國 立 交 通 大 學

電信工程學系碩士班

碩 士 論 文

老人中文語音辨識之初步研究

A Preliminary Study on Elder Mandarin Speech Recognition

研 究 生：楊世帆

指導教授：王逸如 博士

中華民國九十六年九月

老人中文語音辨識之初步研究

A Preliminary Study on Elder Mandarin Speech Recognition

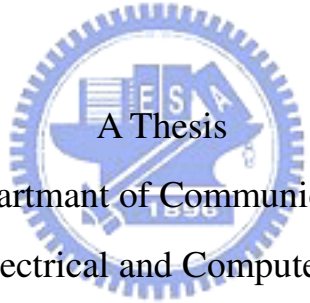
研 究 生：楊世帆　　Student：Shin-Fan Yang

指導教授：王逸如 博士　　Advisor：Dr. Yih-Ru Wang

國 立 交 通 大 學

電 信 工 程 學 系

碩 士 論 文



A Thesis

Submitted to Department of Communication Engineering

College of Electrical and Computer Engineering

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in Communication Engineering

September 2007

Hsinchu, Taiwan, Republic of China

中華民國九十六年九月

老人中文語音辨識之初步研究

研 究 生：楊世帆　　指導教授：王逸如 博士

國立交通大學電信工程學系

中文摘要

在本論文中，從收集的老人語料建立起一個老人中文語音辨識系統，而這個老人中文語音辨識系統的 syllable 辨識率達 44.72%。然後使用 TCC-300 聲學模型來進行老人語料的調適，選用的調適方法為最大可能性線性迴歸；並且在特徵參數抽取時，使用聲道長度正規化來改善老人聲音低沉的特性，當老人語料的聲音頻率被彎曲至較相似年輕人時，再作最大可能性線性迴歸的調適。而且重複 VTLN 加上 MLLR 的調適方法來改善辨識率。最後也分析老人語音腔調差異對辨識與調適的影響，並發現腔調差異的影響可由調適過程來改善；而經由 VTLN 加上 MLLR 的調適過程，可以得到最終的音節辨識率達 51.47%。

A Preliminary Study on Elder Mandarin Speech Recognition

Student：Shin-Fan Yang　　Advisor：Dr. Yih-Ru Wang

Department of Communication Engineering
National Chiao Tung University

Abstract

In this thesis, to build up an elder Mandarin speech recognizer used the collected elder speech corpus, then that syllable recognition to reach 44.72%; moreover, using Maximal Likelihood Linear Regression to adapt the elder corpus by TCC-300 acoustic model. When extracting speech feature, utilizing Vocal Tract Length Normalization to modify the property of the elder voice is to low. When the speech frequency of the elder corpus is warping to be close to the youth speech frequency, we implement the MLLR adaptation; moreover, to use iteration VTLN+MLLR to improve the recognition. Final, to analyze different elder accent to cause distinct result on adaptation and recognition, then we find the MLLR adaptation can decrease the effect by different accent. The VTLN+MLLR adaptation can improve the syllable recognition to reach 51.47%

誌謝

首先要感謝陳信宏老師和王逸如老師的指導，老師們的教誨是對我在研究和做事領域上最大的推力與拉力，讓我確確實實感受到出去外面工作時所會遭遇到的困難，以及學會如何解決這些困難的方法。再來就是要感謝學長們毫不藏私的領導我走進並熟悉語音這塊領域，博班的性獸、志合、希群、阿德以及 barking 在學術研究上的支援，並且在苦悶的交大生活中添加了些許樂趣，還有上屆學長、同屆同學以及下一屆的學弟，都是我在實驗室這段日子的夥伴，特別感謝簡家勇同學在我人生低潮時的幫助，患難見真情才是真朋友，真慶幸一起度過了這個難關。當然也感謝幫我錄音的老人家，沒有你們便沒有這篇論文的存在，還有 my friends，希望你們能諒解我這段時間常常把手機當作 call 機來使用，甚麼之前就這樣，沒這種事，好嗎。

最後也是最重要的，那便是感謝家人的支持，爸、媽感謝你們對我這段期間仍保持著耐心與信心，老哥感謝你對我精神士氣上的加勉，阿嬤也很感謝你的關心。

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究動機

隨著科技的進步，可以讓我們的生活越來越方便，科技產品誕生的目的，就是要學習如何與人溝通、節省工作的時間，而語言也正是我們人類最原始、最簡單、最自然、也是最方便的溝通工具。因此如何發展良好的語音辨識系統來當作人類與機器溝通的橋樑儼然成為非常重要的研究工作。

由於現今台灣社會的老人家們是屬於最晚接觸科技產品的族群，因此在使用科技產品時，對於像是滑鼠、鍵盤或者是觸碰式面板都無法如同青少年或中年人那般順手，使得科技產品對於老人變得較不人性化，鑑於此種原因，我們將於本論文中提出老人語音辨識器，希望能提供給老人家們更便利的使用介面來享受科技。

目前台灣的老人家們(定義六十歲以上者為老人)，由於所接受的教育有極大的落差，有些人是受過國小教育、也有修到博士的學歷、還有接受日本教育的、當然也有完全沒受過教育的人，源自於教育程度與生長環境的差異，將使得老人語音辨識器，不單單只是處理老人語音特性因年紀所造成的影響，也要對台灣老人的腔調做分類，並分析腔調對辨識率的影響。

1.2 研究方向

由於現今國外對於老人語音辨識器，仍無法將其改善至接近於一般年齡層語音辨識系統的水準，並且在國內也沒有相關的辨識系統，所以我們將依循國外的作法：1.基本老人語音辨識系統、2.調適系統(Maximum Likelihood Linear Regression, MLLR)，並加入其他的作法：1.聲道長度正規化(Vocal Tract Length Normalization, VTLN)、2.依腔調來對語料做分類等等，來使得國語老人語音辨識器能有更好的辨識率。所以一個基本的老人語音辨識/調適系統我們會設計如圖1.1：

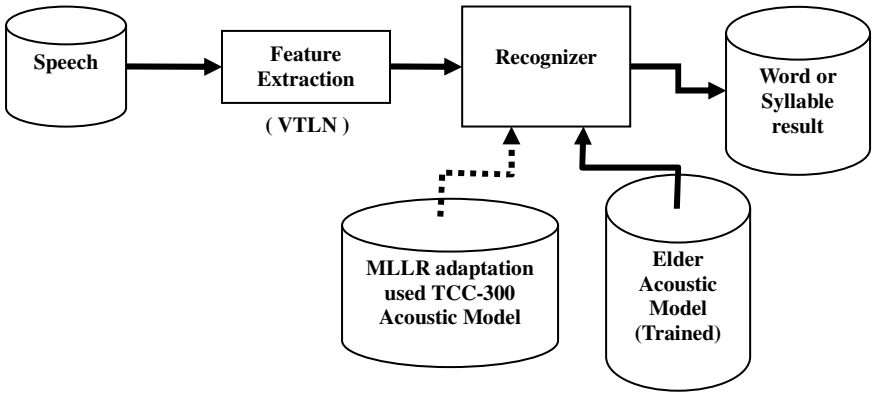


圖1.1 基本的老人語音辨識/調適系統

由圖1.1可看出，一個基本老人語音辨識系統包含了：求取語音參數(Feature Extraction)、老人聲學模型(Elder Acoustic Model)的訓練或者是經由MLLR調適過後產生出新的聲學模型(在本論文中使用既有的TCC-300模型來做調適)，以及進行辨識的方法。圖中實線代表了自行建立的老人語音辨識器，但是根據不同的主題，我們還是要對辨識系統做不同的調整，因此我們將建立一個適合老人的語音辨識器，針對老人語音的特性，如錄音環境、錄音品質的差異、不同的語者腔調等等，探討如何於聲學模型來處理這些問題。在聲學模型上，我們以音節(Syllable)當作辨識的基本單元，而且並不考慮中文的音調變化。

論文中將使用MLLR的調適方法(圖中以虛線表示之)來改善老人語料和一般年齡層(十五至五十歲的語者所構成)辨識系統的匹配程度，並引進VTLN來調整老人語料的頻率，使其和一般年齡層辨識系統較為相似，最後再將VTLN與MLLR一起作到老人語料的調適系統之中，希望以此能有效改善老人語音和一般年齡層語音的差異。

1.3 章節概要

我們將整篇論文以幾個章節區分如下：

第一章：緒論 ─ 說明研究動機、研究方向及章節概要。

第二章：老人語音的特性 ─ 說明老人會隨著年紀的增長，進而改變其發音系統的構造，並改變其語音特性。

第三章：老人語音資料庫 ─ 介紹老人語料庫之起源及其使用的音檔和轉寫格式，

並且有較為詳細的轉寫內容之統計。

第四章：基本老人語音辨識系統 — 說明如何建立一個老人語音的基本辨識系統。

第五章：調適系統與聲道長度正規化 — 採用MLLR來調適老人語料，並由VTLN來改善老人聲道因年紀增長所產生的變化。

第六章：結論與未來發展

第二章 老人語音的特性

本章將簡單地介紹語者發音的模型，並探討老人在年齡的增長對發音器官所產生的變化，進而使老人語音的特性會和一般十五歲到五十歲年齡層的語者有所差異，並在最後我們將逐一列出老人語音特性和一般年齡層語者的不同，依此來當作老人語音辨識器辨識效能較差的起源，並希望能從中排除影響效能的癥結。

2.1 語者發音的原理[1]

語言是由各種語音波形所呈現，語者的語音則是由人體發聲組織，依語者語言內涵，牽動各發音組織器官相關的肌肉，並經由呼吸氣流的帶動，經口腔的管狀構形的變化所形成的共振峰，再通過嘴唇輻射出來，語音在空氣中形成縱向波動的空氣壓力變化形成向外輻射的音壓，當傳至聽者的聲壓波，經聽者中耳、內耳、耳蝸放大，觸動神經，神經訊號傳至大腦，再由大腦轉譯成語言內涵所傳達的意思。前面所述由語音之產生、傳輸、收聽的過程是謂語音鏈路(The Speech Chain)[2]如圖 2.1 所示。

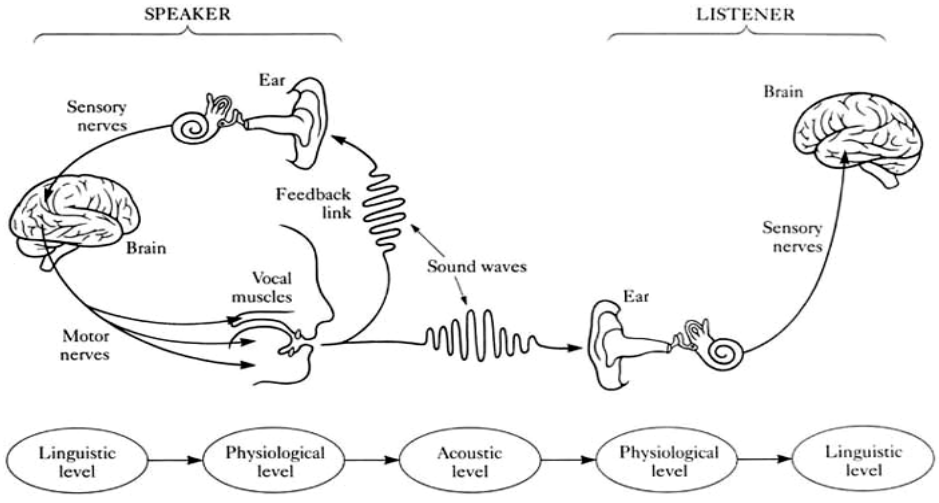


圖 2.1 語音鏈路示意圖 [2]

由語音鏈路中可知，語者的發音過程是必須透過語言學的層級、生理學的層級和聲學的層級來串接構成，在本篇論文中，我們將只討論老人生理構造上的變化對聲學層級的影響。

2.1.1 發音器官簡介[3]

發音器官如圖 2.1 所示，包含了肺部(Lungs)、氣管(Trachea)、聲帶(Vocal folds)(亦即為聲門(Glottis))、喉頭(Larynx)、咽喉(Pharynx)、喉腔、舌頭、軟顎、硬顎、齒齦、齒、唇、口腔、下顎、鼻腔等等。當中由口腔、鼻腔和喉腔構成聲道(Vocal tract)，其具有共振峰(Formant)的濾波器功能，進而將聲門頻譜或雜散氣流作修飾，再經嘴唇幅射出來，而形成各種語音。

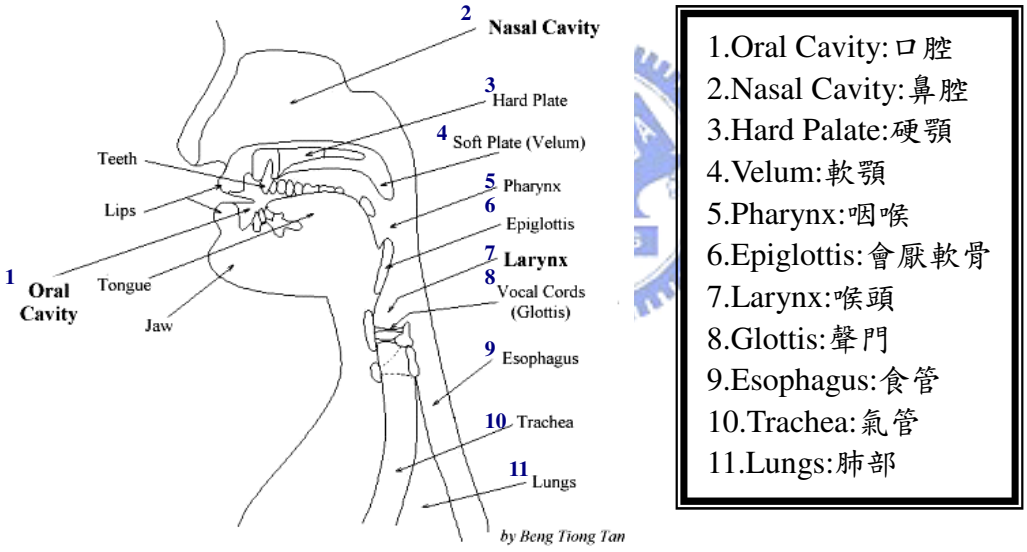


圖 2.2 發音器官

2.1.2 發音過程簡介

當肺部吸入適量空氣後，肺部膨脹儲存空氣；當肺部收縮時，送出壓縮空氣氣流，流經聲帶而成激發音源。聲帶在不說話時，會張開成一口型，讓氣流順暢無阻礙的流過鼻腔或口腔，而產生呼吸氣流聲。

當說話時，聲帶上的肌肉會做某種程度的拉動，使得聲帶予以閉合或微張。當

氣流壓力略大時會把聲帶衝開，整個氣團在聲門聚集，同時因壓差及運動速度的帶動下，往聲道流動，經喉腔、口腔、鼻腔，到口外形成帶聲語音(Voiced)。聲帶內壓力瞬時下降，而使聲帶重新閉合，氣管內氣壓則再次漸升直到再次打開聲帶，以產生振動氣流。上述動作反覆的頻率即為音高(pitch)。簡單來說：

◆聲門震動的快，決定聲音的基本頻率(即音高)。

◆口腔、鼻腔、舌頭的位置、嘴型等，決定聲音的內容(即音色)。

◆肺部壓縮空氣的力量大小，決定音量大小。

而我們已知老人的聲門會因年齡的增長變得較長，使得聲門震動較慢，所以老人的聲音大多數會較年輕人低沉，尤其是男性會特別明顯。在下一節中，我們將會提到國外與老人聲道相關的研究結果。

2.2 老人聲道變化對共振峰的影響

根據國外對老人聲道的研究，An Xue 等人指出[4]，老人的聲道會明顯地較年輕人來的短，假若是採用聲道節管模型[1]以母音/a/的等管路來模擬共振峰$(F_1, F_2, F_3, ..., F_n)$的話，根據下式可知，老人語音的共振峰頻率將會隨著年齡的增長而上升。

$$F_n = \frac{c}{4l_1}(2n-1) \qquad (2.1)$$

但是根據 An Xue 等人的研究結果顯示老人語音共振峰並不會如此規律地隨聲道變短而作改變，其研究由表 2.1 可以看出 First Formant 下降，而 Second Formant 會上升、Third Formant 會下降。不過，根據 Linville 與 Fisher [5]指出 First Formant 及 Second Formant 皆會隨年齡增長而下降；而 Rastatter 與 Jacques [6]有著不同的看法，也就是說，First Formant 會因測量的音節不同而有不一致的變化，其與前兩者相同的是母音/a/的 F1 會下降，而 F2 會上升。對於如此不一致的結論，可以由下表看出一些端倪。

表 2.1 老人聲道長度及共振峰與年輕人的比較 [4]

	Younger subjects	Older subjects
Vocal tract length mean (in cm)	18.15	17.58
SD	1.27	1.26
Range	16.00–19.50	15.50–19.50
Pharyngeal volume mean (in ml)	27.12	31.06
SD	5.56	8.35
Range	17.80–36.10	22.20–53.30
Formant 1 mean (in Hz)	897.83	668.34
SD	136.76	323.10
Range	585.94–1086.43	227.05–1109.62
Formant 2 mean (in Hz)	1631.84	1760.76
SD	590.88	1090.81
Range	1219.48–3260.50	726.32–4226.07
Formant 3 mean (in Hz)	3191.28	2974.28
SD	715.03	1427.18
Range	2486.57–5133.06	1350.10–6640.63

由上表的統計值可看出老人聲道的確會隨著年齡的增長而變短，針對母音/a/而言，Formant 1 也會變得較低，而 Formant 2 與 Formant 3 的統計資料中老人的標準差(SD)與範圍(Range)都比年輕人來的寬廣，而平均值卻相去無幾，這顯示了老人語音特性中 F2、F3 是無法明確去做計算的，這造成了先前的一些相關研究產生矛盾的現象。接下來我們將歸納一些老人語音的特性於下一節中，並以此做為我們老人語音辨識系統研究的起源。

2.3 老人語音特性的歸納

由前兩節已經稍微介紹了一些老人語音的特性，於此我們將歸納這些特性，並將部份可以由一些方法予以處理的特性條列起來，而其他尚未有明確地研究分析出來的特性，例如：Second and Third Formant 具有極大的變異量，將不在後面的章節處理。

一般來說，大家對老人語音的特性直覺為比較大聲、說話速度較慢、聲音較低沉，不過說話較大聲多半是耳朵的功能衰退所造成的相對反應，而說話速度慢有時是感官神經變得較遲緩了、有時是大腦反應變遲鈍了，對於聲音較低沉這個現象，

我們也抱持著多數人具有此種現象，但卻不盡然，也有維持語音頻率的人存在，畢竟這些語音現象的實驗並沒辦法採取同一個人依年齡變化分別作測試。以下是已被語言學家證實的老人語音現象：

◆聲門變長，老人語音較低沉，尤其是男性

◆聲道長度變短，First Formant 下降，其他 Formant 變異量極大

◆說話速率(speech rate)較緩慢

當我們於下一章介紹完老人語音語料庫之後，會將所有老人語音特性與錄音之後產生的問題總結起來，並依這些特性、問題去選擇適當的做法來使得老人語音辨識有較佳的效能。

第三章　老人語音資料庫

本章首先介紹研究中所使用的老人語音資料庫，並說明本語料庫的文字轉寫的由來、錄音規格、錄音者年紀的分布、錄音環境的差異、錄音內容的字數以及錄音語料的處理過程等相關細節。

3.1 老人語料庫之簡介

老人語料庫是使用 IBM-T43 筆記型電腦於 2006-2007 期間所錄製的，其語料錄音者是由台北市與新竹市市民所構成，並限定錄音者的年紀必須在六十歲以上，當中包含了十九名女性和三十五名男性，總共有五十四名錄音者。此錄音的語料屬於朗讀式語音(read speech)，而我們取用的文字轉寫(transcription)源自於 MAT-2000 及 MAT-2500 [6]中四個音節(syllable)以上的句子，使得每位語者皆具有六十句，並將其文字轉寫修正為每句中只包含十個音節左右或十個音節以內，希望以此來減少因年紀增長導致記憶力下降的因素，以提高老人錄音的品質，因此每位語者所錄的句數將變成六十句至八十多句不等。

3.1.1 音檔格式說明

我們是採用 Ergotech 的 ET-E110 型麥克風與 SoundMAX Integrated Digital Audio 的音效卡來錄製老人的語音，以 16kHz 的取樣率和 16bits 的準確度來錄製。

File Format : Wav
Sampling Rate : 16 kHz
Resolution : 16 bits
Channel : Mono

圖 3.1 音檔格式之設定

3.1.2 錄音者的年紀分布

由於第二章之中，我們對於老人語音特性的描述，因此我們將限定錄音者的年紀必須要在六十歲以上，在此我們將列出本語料庫中男性語者和女性語者的年紀分布，當中老人語料庫語者的平均歲數為 69.56 歲。

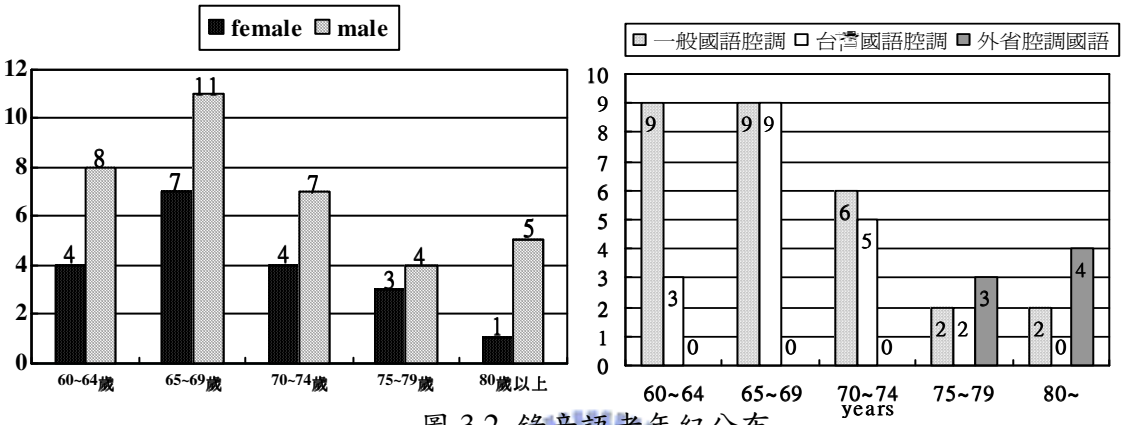


圖 3.2 錄音語者年紀分布

3.1.3 音檔內容的統計

我們於此節中，將把老人語料庫的音節數、句數作一個統計，並顯示於表 3.1 之中。雖然我們的錄音屬於 read speech，但由於語者的習慣，導致某些音檔會把錄音者的一些發語詞、咳嗽聲、吸氣聲等口語現象也一併收錄進來，或者一些不確定音，也就是無法以漢語拼音來表示其文字轉寫的音節。因此我們必須在原先的 HMM 模型中加入一個 filler 的語音音節模型，來應付這些多餘的音節。並將每一個音檔內容中，一些既定文字轉寫之外的發音音節，標示其漢語拼音的文字轉寫，這些多出來的音節多半屬於自發性語音(spontaneous speech)範疇中語者口語現象的詞語修補(repair)現象，由於我們在此是 read speech 的辨識，且語料蒐集不易，故不拿掉出現此現象的句子，儘管在加入語言模型後會影響到辨識結果。且由於錄音環境包含室內及室外，以致於某些音檔會有背景人聲雜音(background speaker noise)，例如：錄音者以外的人們交談聲、無法控制的嬰兒聲音；或者一些較嚴重的背景雜音(noise)，例如：尖銳的鳥叫聲、車子行經所產生的噪音、他人行走時拖鞋拍打地面的聲音、甚至是手機的鈴聲，我們為了降低這些現象對於整個語音辨識系統的影響，所以我

們將發生這些現象的音檔做了如表 3.1 的統計，並標記於這些音檔的文字轉寫中。

表 3.1 語料庫中的音節數、句數、filler、background speaker noise、noise 的統計

	音節數 (Syllable)	句數	filler 出現的音節數	背景人聲雜音出現的句數	背景雜音(noise)出現的句數
女性語者	10741	1423	42	38	40
男性語者	19766	2478	39	26	126
全部語者	30507	3901	81	64	166

3.2 訓練語料與測試語料

因為老人語音中，語者間的語音特性與一般年齡層辨識系統相比，其差異較大，這是由於現在台灣的老人家們有著不同的腔調所造成，例如；台灣國語、一般的國語腔調以及外省腔調的國語(其分布詳見附錄一)，在接下來的章節會對腔調的差異來做分析。由於我們將建立老人語音的聲學模型(acoustic model)來進行語音辨識，我們把老人語料庫中的語者抽取八句長句子來作為測試語料，剩下的其他句子拿來當作HMM模型的訓練語料，採取Multi-speaker的語音辨識，是為了因應語料不足時，若採用語者來分訓練及測試語料，會造成辨識結果較缺乏客觀性，並且使用這種speaker dependent的語料分法，能針對語者腔調的差異來進行實驗。下表為訓練語料與測試語料具有的音節數、句數和filler出現的音節數。

表 3.2 訓練語料中的音節數、句數、filler 的統計

	音節數 (Syllable)	句數	Filler 出現的音節數
女性語者	8841	1270	29
男性語者	16145	2199	33
全部語者	24986	3469	62

表 3.3 測試語料中的音節數、句數、filler 的統計

	音節數 (Syllable)	句數	Filler 出現的音節數
女性語者	1901	152	13
男性語者	3619	280	7
全部語者	5520	432	20

3.3 調適語料與被調適的 HMM 模型

由於老人語音資料的收集有一定難度，所以在論文中老人語音辨識模型將無法如同其他語音辨識模型那麼地強健，也就是說HMM模型可能會因資料量不足，而無法有效地去描述某些音節的特徵。在此我們便採用第五章即將介紹的MLLR來調適語音辨識系統，將使用TCC-300所訓練出來的語音模型作為即將被調適的HMM模型，其詳細的各項設定如附錄二所示。並且取用老人語音辨識系統中的訓練語料來當作調適語料，測試語料仍維持老人語音辨識系統的測試語料，而調適時的訓練與測試語料都將配合TCC-300的acoustic model，把transcription中的filler音節給刪除掉，因為我們使用的TCC-300 acoustic model具有一個描述silence、short pause以及與呼吸聲相關的聲學模型，而filler與這模型相當符合，只要確定具有filler的語句其切割位置是正確的。下圖為調適過程的結構[8]：

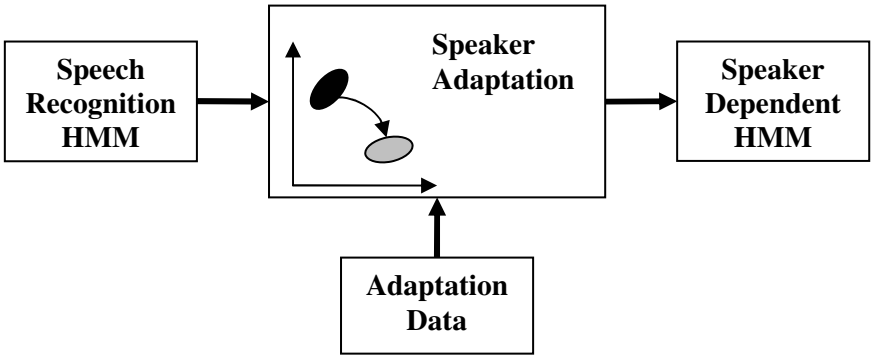


圖 3.3 基本調適過程的結構圖

根據上圖所示，我們使用 MLLR 的語者調適過程將已訓練好的 TCC-300 之 HMM 模型，透過老人調適語料的特性將 TCC-300 之 HMM 模型轉換成符合老人語者的語者相關 HMM 模型，詳細調適內容將於第五章仔細介紹。

3.4 老人語料的特性與問題歸納

經由第二章與本章的整理之後，將預先把老人語料會在將來實行辨識及調適時需要注意的一些語音特性及錄音引起的問題作個歸納，並以此考量這些歸納後的事項需要由何種方法來解決或降低其影響，以下為我們所歸納的事項及對策：

1.由於錄音環境的差異(室內與室外)造成了錄音品質良、劣差異極大，也產生了各種的雜訊，這些錄音引起的問題與老人的語音特性毫無關係，所以我們採取的對策便是在建立老人語音辨識器與調適老人語料前，先剔除這些語句，甚至是語者。

2.因為錄音時，老人語者未必會唸出完全對應於transcription的發音，所以預先聽過所有語者的音檔，並確認其transcription是否與發音一致，若出現不一致的現象，便更正transcription為相對應的漢語拼音。

3.少數的語言學現象，例如：咳嗽聲、發語詞或是一些無法由漢語拼音構成的音節，都將採用filler model來替代之，不過被調適的TCC-300聲學模型，並不具有filler的音節，但具有一個包含silence、short pause以及與呼吸聲相關的聲學模型，所以在調適時把transcription中的filler給予剔除，並確定具有filler的語句中其切割位置是正確的，讓filler存在的音節使用這個與呼吸聲相關的模型所調適，如此一來便能使用調適去把TCC-300 HMM model轉換至較符合老人語料的新HMM model，所以僅有在建立老人語音辨識系統時採用filler model。

4.聲門與聲道因年齡的增長所產生的變化，總的來講，就是聲音會變低沉，不過Second Formant、Third Formant…等共振峰與年齡的相關性，並沒有相關的研究確實指出其合理的改變行為，所以我們於調適老人語料時，將採用VTLN來warping老人語音的頻率，並視每一位老人語者的頻率變化為線性的變化(詳見第五章內容)。

5.依腔調(一般腔調國語、台灣國語腔調與外省腔調國語)來劃分老人語者，並比較腔調的差異對於辨識與調適兩種方法的影響。

第四章 基本老人語音辨識系統

本論文的基本語音辨識系統架構是建構在 HTK(HMM Tool Kit)v3.4 [9]上。HTK 是由英國劍橋大學針對語音辨識，所開發出來的工具；這套工具主要是基於 HMM 用來實行語音辨識；在使用 HMM 實行大詞彙語音辨識上，已被證實能夠獲得不錯的效果；而 HMM 最大的特點是它利用訊號的統計特性去描述訊號。故本論文中將採用 HTK 來建立老人語音的辨識系統，並且使用 MLLR 的方法來調適老人語料。

4.1 老人語音的基本辨識系統

老人語音辨識系統的基本架構[10]，它的方塊圖見圖 4.1，主要包含三個部份：(1)特徵參數擷取；(2)聲學模型訓練；(3)辨識比對。

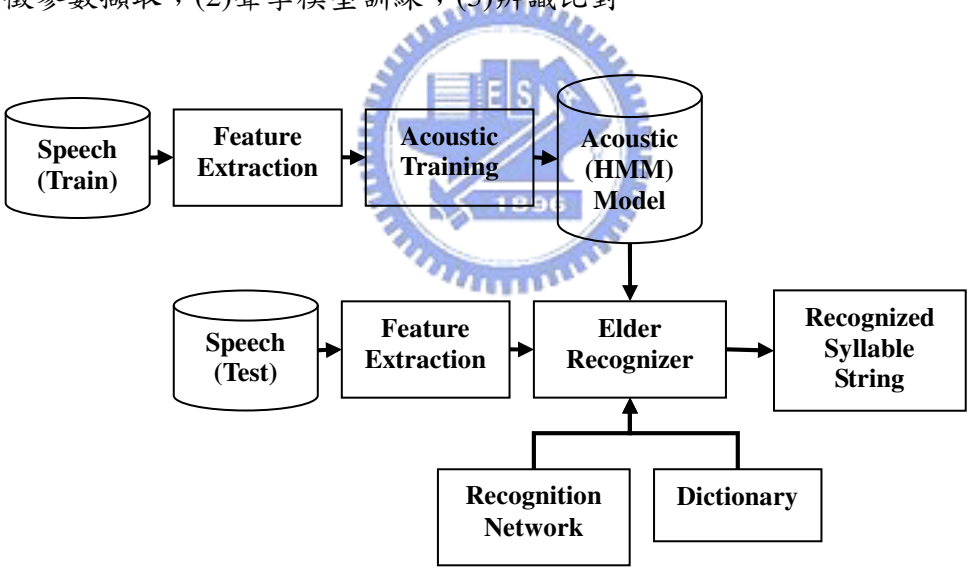


圖4.1老人語音辨識系統方塊圖

各個方塊之功能說明如下：

(1)Feature extraction：對音訊做處理，抽取出能表示此語音的特徵訊息，以作為語音訓練及辨識的參數。

(2)Training：利用特徵參數與隱藏式馬可夫模型，訓練出聲學模型。

(3)Dictionary：此檔案記錄了聲學模型與基本音節的對應關係。在辨識時可經由查詢此檔案的動作，來將辨識後的聲學模型符號，轉成基本音節的拼音符號。

(4)Recognition network：作為辨識時所依據的搜尋網路，先前的辨識網路並無加任何限制。

接著，將簡述各方塊之系統參數。

4.1.1 系統環境參數設定

1.HTK特徵向量求取原理說明

我們首先做一個假設－聲音訊號在幾毫秒內是stationary。然後，再將聲音訊號做一連串的處理，處理順序如下[11]：

- ◆將聲音訊號切成若干的區塊，而每個區塊大小32ms間隔10 ms，相互重疊。
- ◆對聲音訊號做預強調的動作(即高頻放大)，用來補償從口腔所引起的發散衰減現象。然後，再對每塊區塊取Hamming Windows作smoothing動作。
- ◆接下來就是MFCCs(Mel-Frequency Cepstral Coefficients)參數求取流程，如圖4.2所示。

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

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

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

$$\begin{cases} \Delta: \text{delta coefficients} \\ \Delta^2: \text{acceleration coefficients} \\ dt: \text{the first differential coefficient} \\ dt^2: \text{the second differential coefficient} \\ c_t: \text{MFCC coefficient} \\ \Theta: \text{delta window size} \end{cases} \tag{4.1}$$

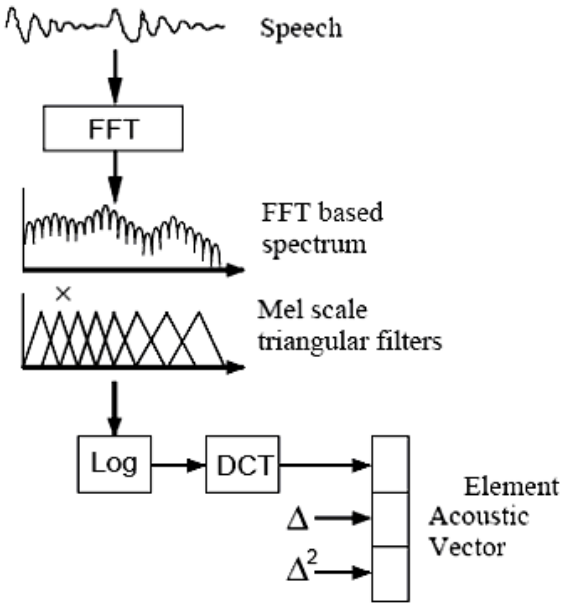


圖 4.2 抽取參數流程圖 [9]

2.特徵向量求取環境設定

◆預強調： $1-0.97z^{-1}$

◆視窗類別：漢明視窗(Hamming Window)

◆音框長度：32 ms

◆音框偏移量：10 ms

◆Filter Band 頻率範圍：0~8kHz

◆通道數：24

◆38維之參數向量

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

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

因此，針對語音訊號進行參數求取之後，得到 38 維的語音參數向量，而在第五章使用的 VTLN 即是於此處對老人語者語音做頻率彎曲的動作。

4.1.2 聲學模型及其訓練與測試

於此我們採用的是 left-to-right HMM，雖然口腔聲道會隨時間而變，但因為語音訊號具備短時間的穩定特性，因此假設在同一音框(Frame)中，口腔狀態是相同的。此外，代表音框與各狀態的相似程度的狀態觀測機率(State Observation Probability)，使用混合高斯模型(Mixture Gaussian Model)來表示[12]。

另外，訓練模型時重估模型參數則利用 Baum-Welch 重估演算法，並重複估測聲學模型至穩定為止；至於辨識工作的進行則是使用 Viterbi 演算，將輸入的訓練語音針對一個模型計算其由該模型產生的機率，並找出最佳狀態序列[13]。

並在訓練模型時使用 HTK v3.4 中的 proportional state 的做法，去達到訓練聲學模型時能夠去對照其統計資料，產生出較適合我們老人語料資料量較少的模型，並重複訓練聲學模型七次，使得聲學模型能達到穩定，而在辨識時求出 optimal penalty，使得辨識率是最佳的。

4.2 建立多語者(multi-speaker)聲學模型與實驗

在本論文中，將建立一個 multi-speaker 的老人語音 HMM 模型，而 HMM 初始模型將採用 flat start 的方式來建立，這種方法首先假設一段語音中的切割位置平均分布，並先用訓練語料訓練出一個初始的模型，將它提供給所有的次音節

(Sub-syllable，在此即為中文的聲母(initial)和韻母(final))使用，雖然此種作法在一段語音較長的情況下容易發生切割位置錯誤的情形，並且需要花較多的時間訓練出一個正確的模型，但在先前所提到的老人語音資料庫，因過長的句子都修改為十個多音節或以下，且整個資料庫字數約三萬多個音節，所以使用 flat start 仍是合理的選擇。

而我們採用的辨識網路(word net)並沒有加任何的限制，即為任何音節皆可接任何音節，示意圖如下所示。

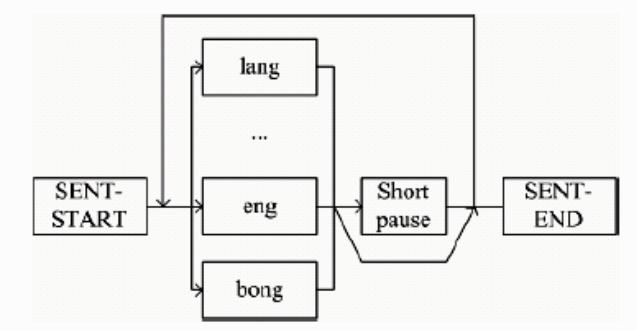


圖 4.3 辨識網路 [9]

到目前為止，老人語音語料中的國語 411 音、filler、silence 都準備妥當，便可以進行初始模型訓練。另外我們還會建立一個 sp (short pause)的 HMM 模型，這是代表音節與音節之間的短暫靜音，sp 只有一個狀態，此狀態允許跳躍(skip)，並且與 silence 的中間狀態合併(tying)，如下圖所示。

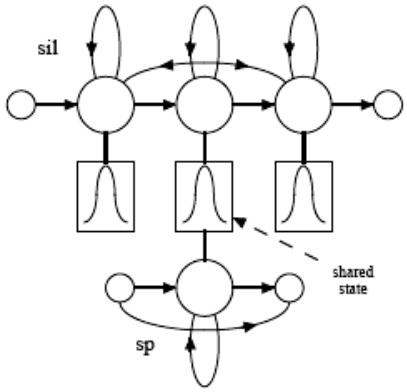


圖 4.4 silence 與 sp 做 tying 示意圖 [9]

而其中 filler 是老人語料中存在許多自然語音中經常出現的口語現象，這些現象是不易個別處理的，例如說話前的清喉嚨聲、笑聲等語言學現象和無法辨識其漢語拼音的音節，因為在語料中存在的資料量並不足夠訓練出個別的聲學模型，所以我們採用的方法是建立一個共同的特殊聲音模型對這些現象進行處理，並將這些現象統稱為「filler」，而稱呼用來取代各種特殊聲音的語音模型為「filler model」，並把 filler 加入於字典(dictionary)當中，且訓練 filler 為三個狀態的 HMM 模型。下表顯示出老人語音辨識系統 HMM 模型的參數設定：

表 4.1 老人語音辨識系統 HMM 參數設定

模型類型	個數	狀態數	Mixture/狀態
聲母	100(RCD)	3	1~16
韻母	40	5	1~16
filler	1	3	1~16
silence	1	3	32
short pause	1	1	32

『實驗一』以 TCC-300 聲學模型(詳細的設定見附錄二)對老人語音語料作測試

在此我們先作老人語音測試語料對一般年齡層的 TCC-300 聲學模型作辨識，下表為 syllable 辨識率與 penalty 的關係：

表 4.2 以 TCC-300 聲學模型對老人語音語料作測試

penalty	%Del	%Sub	%Ins	%Corr	%Acc
-50	2.3	59.18	17.77	38.51	20.74
-60	2.59	59.04	16.67	38.37	21.70
-90	2.95	59.13	15.45	37.92	22.46
-120	3.08	59.26	15.11	37.66	22.55

表中的縮寫分別為Corr = correct percentage，Acc = accuracy figure，Del = deletion error，Sub = substitution error，Ins = insertion error，N = total syllable(這次實驗的N = 5520)。而辨識率的計算方式如下：

$$Acc(\%) = \frac{N - (Del + Sub + Ins)}{N} \tag{4.2}$$

$$Corr(\%) = \frac{N - (Del + Sub)}{N} \tag{4.3}$$

由此次實驗可知，老人語音語料與一般年齡的語料有很大差異性，當中有相當多的因素，我們必須將這些因素一一分析，首先要把錄音品質、各種雜訊這類與老人語音特性無關的外在因素予以降低，接著我們來看下一個實驗。

『實驗二』建立基本的老人語音辨識系統

依前述的作法，我們首先建立一個老人語音辨識系統，且觀察訓練好的 HMM 模型檢查其統計資料，發現有五個 sub-syllables (分別為兩個韻母：eh、yo 及三個聲母：c_o、s_o、n_o)沒有出現於訓練和測試語料當中，所以這五個訓練出來的模型是取用 global mean 時的參數，故我們將其在 HMM 模型中刪除掉，且在 dictionary、phone table(修改過後的 phone list 詳見附錄三)與 grammar 中做修改，進而產生新的 word net，並計算總辨識率、每位語者的辨識細節(亦即為 deletion、insertion error 的比例)以及每個句子的辨識細節，並希望能從中發現於第三章所標記的錄音品質、背景人聲雜訊、背景雜訊等因素對辨識細節的影響，下表為總辨識率：

表 4.3 基本老人語音辨識系統之辨識率

penalty	%Del	%Sub	%Ins	%Corr	%Acc
-50	4.55	49.33	4.6	46.12	41.52

於本次實驗中，我們使用 HTK 的工具求得 optimal penalty，而在實驗一中也使用了相同的作法，但卻無法求得 optimal penalty。上表如預期其辨識效能比實驗一好很多，但我們由每位語者的辨識細節以及每個句子的辨識細節可比較出錄音品質、各種雜訊對老人語音辨識系統的影響，所以將每一個辨識較差的句子(句中包含了太多的 deletion 或 insertion error，而不是以辨識率高低來區別)與先前標記好的音檔細

節作比對，並分別去聽音檔的內容；因為其中有三個語者出現了大量的 deletion 或 insertion error，比對之後發現這三個語者具有極差的錄音品質(m002、m003 是由於環境周圍有太多的人聲，導致這兩位語者的每一個音檔都具備了極嚴重的背景人聲雜訊，f014 是由於語者的聲音太小且背景有著不小的雜音，因此這個語者的音檔幾乎是全為雜訊的)，所以我們也對訓練語料做相同的比對，也有相同的現象，為了降低這三個語者因錄音產生的問題對整個辨識系統的影響，故我們將這三個語者從語料庫中去除，下表為這三個語者的 deletion、insertion error(%)：

表 4.4 去除的語者之 deletion、insertion error(%)

speaker	Deletion %	Insertion %
M002	**22**	1
M003	**21**	1.6
F014	0	**42**

對照所有標記的問題音檔，進而去除掉一些被極嚴重的偶發性地背景雜訊或極嚴重的背景人聲雜訊給影響到產生多數 deletion、insertion error 的句子，但不須把具有腔調或咬字不清、較慢等具老人語音特性的句子去除，儘管此句的辨識率為零。因此，除了前三個語者，又分別在測試語料與訓練語料去除了 13 與 58 句，儘管如此也只能降低錄音時的因素對辨識系統的影響，無法完全排除之，但受限於錄音語料已經夠少了，而且找尋老人錄音者有一定難度，所以也只好針對特別嚴重的句子做去除，以下為去除之後所剩餘的語料統計：

表 4.5 修改後訓練、測試語料中的音節數、句數、filler 的統計

	音節數 (Syllable)	句數	感歎詞(filler)出現的音節數
訓練語料	23442	3253	53
測試語料	5038	395	20

『實驗三』以 TCC-300 聲學模型對去除問題語句後的老人語音語料作測試

在去除掉那些因錄音時所造成辨識極差的句子後，再一次測試老人語料對

TCC-300 聲學模型的辨識率：

表 4.6 TCC-300 對去除問題語句後的老人語音語料作辨識

	%Del	%Sub	**%Ins**	%Corr	**%Acc**
實驗一	2.95	59.13	**15.45**	37.92	**22.46**
本實驗	4.4	56.57	**6.9**	39.02	**32.12**

這次實驗與(實驗一)未去除問題語句的實驗採用相同的 penalty 為-90，在去除問題語句之後，在實驗一中極端嚴重的 insertion error 有著大幅度地下降，也因此大幅度地提升了%Acc。這次實驗也可以有效的平衡了 deletion 與 insertion error。

『實驗四』去除問題語句後的老人語音辨識系統

在去除掉那些因錄音時所將會被雜訊污染老人語料庫的句子後，再一次建立老人語音的基本辨識系統，其中 penalty 採用與(實驗二)未去除問題語句的老人語音辨識系統相同的-50，並列出其辨識率的變化如下表：

表 4.7 去除問題語句後的老人語音辨識系統之辨識率

penalty	%Del	%Sub	%Ins	%Corr	**%Acc**
實驗二	4.55	49.33	4.6	46.12	**41.52**
本實驗	4.4	47.72	3.2	47.88	**44.72**

由表中可看出，和實驗二相比，%Acc 分別上升了約 3 個百分比，deletion 和 insertion error 也明顯地降低了約 1.5 個百分比。

4.3 說話速度(speaking rate)對老人語音辨識系統之影響

相對於一般年齡層語者在 read speech 的說話速度(約為每秒五個 syllables 左右)，而老人語料庫的整體 speaking rate 為每秒 2.658 個

syllables，可見老人的說話速度明顯較一般年齡層語者慢，所以必須探討 speaking rate 對於老人語音辨識率的影響。下圖為每位老人語者 speaking rate 的分布：

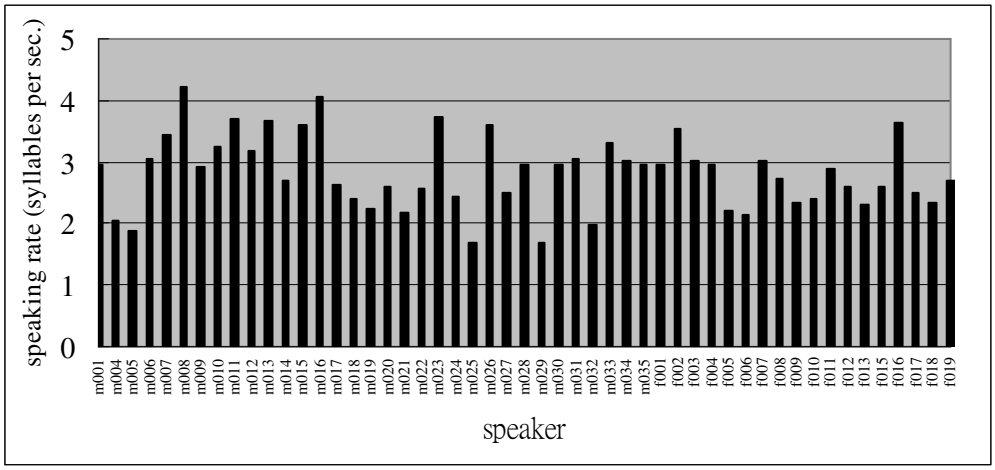


圖 4.5 老人語者 speaking rate 分布圖

又以腔調及年齡來統計 speaking rate 的變化，並對照這些變化與老人語音辨識器之結果，如圖 4.6、圖 4.7 所示：

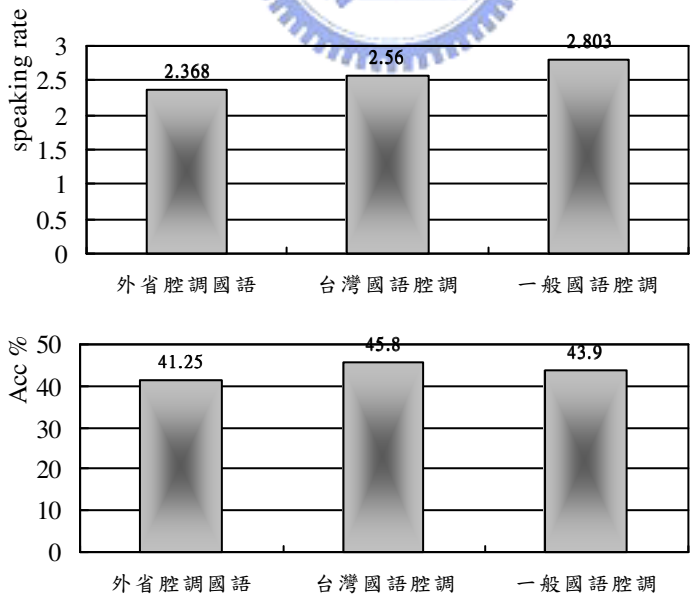


圖 4.6 腔調分類對 speaking rate 與老人語音辨識率的關係

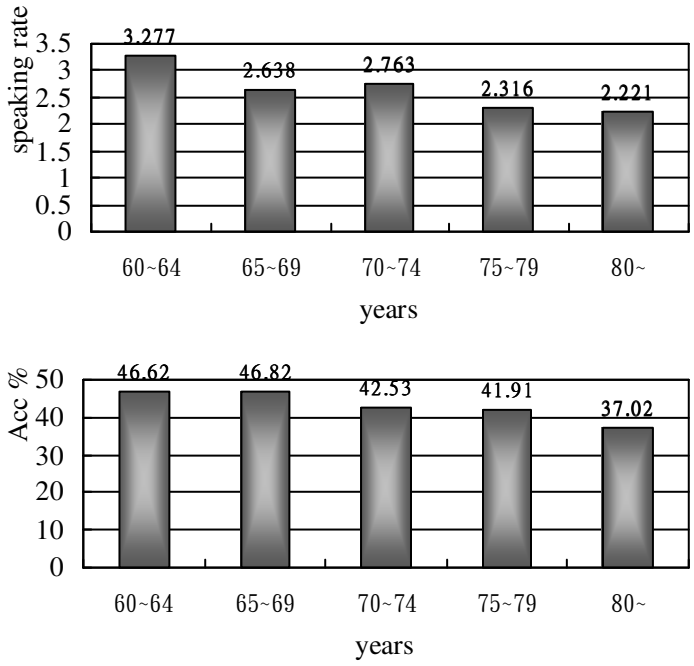


圖 4.7 年齡分類對 speaking rate 與老人語音辨識率的關係

圖 4.6 中顯示了台灣國語腔調的 speaking rate 較一般國語腔調慢，但是台灣國語腔調的辨識率卻較佳；而圖 4.7 具有 speaking rate 隨著年齡而下降的趨勢，對應其辨識率也有此種趨勢，但是於 65~69 歲的區段，會因此區段的台灣國語人數較多，導致辨識率會跳脫此一趨勢。以上對腔調的分析，會在下一章中做個結論。

第五章 調適系統與聲道長度正規化

本章將使用 MLLR 的方法來進行老人語料的調適，並且利用 VTLN 來把老人語料的頻率 warping 至符合 TCC-300 acoustic model 的一般年齡層音高，再將 VTLN 與 MLLR 一起對老人語料來進行調適，最後由這些作法來觀察老人語音的腔調與辨識率的關係。

5.1 最大可能性線性回歸(MLLR)簡介[14]

MLLR 應用基礎主要在迴歸類別(regression classes)的觀念。每一迴歸類別可由一群高斯成份(mixture component)集合所組成，在此類別當中每一高斯分布，可利用相同的轉換矩陣來進行調適，而在每一迴歸類別當中高斯分布成員數目則不一定，最多每一個高斯成分可單獨成為一個類別，至少則可把所有模型參數當做是一個類別，而做這種分類的好處是，當調適語料在有限的情形下，有的高斯成份在調適時並沒有觀測到任何的調適語料，也可以使用相同轉換矩陣來進行調適。使用轉換矩陣 $\mathbf{W}_s$ 進行語者調適可以寫成 Eq.(5.1)式：

$$\hat{\boldsymbol{\mu}}_s = \mathbf{W}_s \cdot \boldsymbol{\xi}_s \qquad (5.1)$$

其中 $\mathbf{W}_s$ 為對高斯成份 s，維度為（$n \times n+1$）的轉換矩陣，n 是音框特徵向量的維度，$\hat{\boldsymbol{\mu}}_s$ 是對高斯成份 s 調適後的平均值向量，而 $\boldsymbol{\xi}_s$ 稱之為高斯成份之擴大平均值向量(extended mean vector)，定義成 Eq.(5.2)式：

$$\boldsymbol{\xi}_s = [\omega, \mu_{s1}, ..., \mu_{sn}]' = [\omega : \boldsymbol{\mu}_s]' \qquad (5.2)$$

$\boldsymbol{\mu}_s$ 為高斯成份 s 原來之平均值向量，而 ω 是一項位移量(offset term)，若調適語者的錄音環境與初始模型錄音環境不同時，可以加入的一項參數，通常設為 1。圖 5.1 為將高斯成份分成兩個迴歸類別，並使用 MLLR 對平均值向量於聲學特徵向量空間中做轉換的例子，圖 5.2 為 ω 值與 MLLR 轉換的關係，圖中實線代表平均值向量，虛線代表轉換的動作：

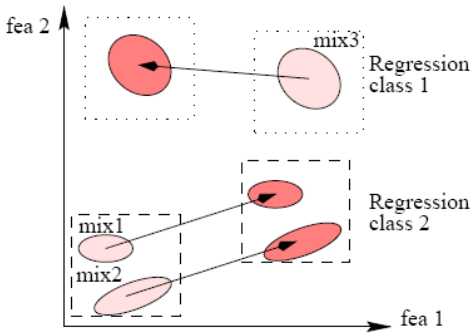


圖 5.1 迴歸類別依 MLLR 法作轉換的圖例 [8]

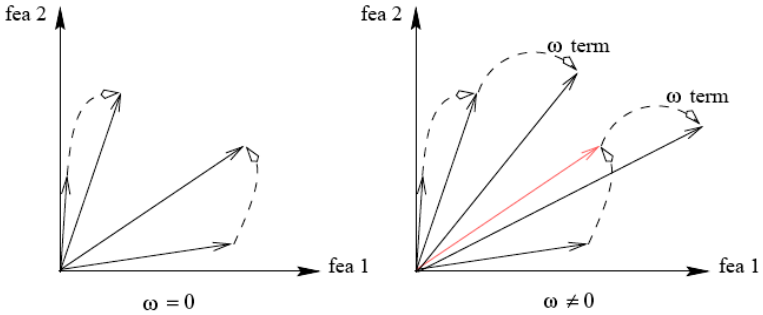


圖 5.2 ω值與 MLLR 轉換的關係 [8]

在本論文中，不仔細地推導 MLLR 轉換矩陣的估計，僅僅是做 MLLR 的一些基本觀念的介紹。應用最大可能性(Maximum Likelihood, ML)估測，也就是必須找到一轉換矩陣 $\hat{\mathbf{W}}_s$，當透過 $\hat{\mathbf{W}}_s$ 轉換之後模型參數 $\hat{\boldsymbol{\lambda}}$，使得調適語料能產生最大機率值：

$$\hat{\mathbf{W}}_s = \max_{\mathbf{W}_s} P\left(\mathbf{O}|\hat{\boldsymbol{\lambda}}\right) \tag{5.3}$$

其中，$\mathbf{O} = \{o_1, o_2, ..., o_T\}$ 為觀測到調適語料的特徵向量，$\hat{\boldsymbol{\lambda}}$ 則定義為調適後模型的參數。求解 Eq.(5.3)式的過程相當於在訓練隱藏式馬可夫模型時重估模型參數時所用到的 Baum-Welch 演算法。在這裡，定義一輔助方程式：

$$Q\left(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}\right) = \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} P\left(\mathbf{O}, \boldsymbol{\theta}|\boldsymbol{\lambda}\right) \cdot \log\left(P\left(\mathbf{O}, \boldsymbol{\theta}\middle|\hat{\boldsymbol{\lambda}}\right)\right) \tag{5.4}$$

其中，$\boldsymbol{\theta}$ 為一種可能狀態序列，$\boldsymbol{\Theta}$ 則是所有可能狀態序列集合，每種狀態序列長度為 T，為所觀測到調適語料的總音框數。Huang 等人[15]在 1990 年證明有這樣的性質，假使能找到參數 $\hat{\boldsymbol{\lambda}}$，使得輔助方程式 $Q(\cdot)$ 有最大值，亦即是：

$$Q\left(\lambda,\hat{\lambda}\right)\geq Q(\lambda,\lambda)\Rightarrow P\left(\mathbf{O}\middle|\hat{\lambda}\right)\geq P\left(\mathbf{O}\middle|\lambda\right) \qquad (5.5)$$

則會有最大 $P\left(\mathbf{O}\middle|\hat{\lambda}\right)$ 值，所以我們的目的就是讓輔助方程式有最大值。以上便是 MLLR 的起手式，詳盡的推導或變化在此不做描述。

5.2 老人語料進行 MLLR 調適之實驗

經由上一節對 MLLR 的原理推導之後，在此將作老人語音語料對 TCC-300 的聲學模型作調適的動作，當中所使用的調適語料為修改過後的基本老人辨識系統的訓練語料(請參照 4.2 節實驗四)，而測試語料為修改過後的基本老人辨識系統的測試語料，但是 TCC-300 的聲學模型中並沒有 filler 這個 HMM 模型，但具有一個包含 silence、short pause 以及與呼吸聲相關的聲學模型，所以在調適時把 transcription 中的 filler 給予剔除，並確定具有 filler 的語句中其切割位置是正確的，讓 filler 存在的音節使用這個與呼吸聲相關的模型所調適，如此一來便能使用調適去把 TCC-300 HMM model 轉換至較符合老人語料的新 HMM model。於此節中，將作 MLLR 的調適，並且將依語者腔調的不同，分別對 TCC-300 模型作直接辨識與 MLLR 的調適，並且從中觀察腔調對辨識率的影響。

『實驗五』以 MLLR 對老人語料做調適

在本實驗中，我們將採用 HTK 所支援的工具來進行 MLLR 的調適過程，如前述可知，被調適的模型為 TCC-300 的聲學模型，所以在老人語料先行刪除 TCC-300 中沒有的 filler 這個音節的文字轉寫，其中抽取語音參數與訓練的設定維持與 TCC-300 的設定相同，並依照上一節的原理，在 MLLR 中的 linear transforms 加入 bias，使用的 transformation kind 為 MLLRMEAN，產生出 mean vector transform 的轉換矩陣，並嘗試使用另一種作法：在 MLLR mean transform 中加上 diagonal variance transform 的轉換矩陣，以上兩種不同的轉換方式依 regression tree 而分出不同數量的 regression classes，並進行 MLLR 調適所得的 syllable 辨識率(其中為了使實驗具有一致性，實驗中採用 diagonal variance transform 求出來的 optimal penalty -70，作為 mean

vector transform 實驗中的 penalty)：

表 5.1 老人語料以 MLLR 調適的結果

MLLR with mean vector transform					
Regression Classes	**%Del**	**%Sub**	**%Ins**	**%Corr**	**%Acc**
32	3.55	52.17	6.24	44.28	38.04
64	3.69	50.12	5.92	46.19	40.28
128	3.61	48.92	5.68	47.47	41.79
256	3.67	46.89	5.68	49.44	43.76
MLLR with mean vector plus diagonal variance transform					
32	5.06	49.34	2.67	45.6	42.93
64	5.06	47.81	2.55	47.09	44.54
128	5.06	45.91	2.45	49.02	46.57
256	5.16	43.5	2.45	51.34	48.88

一般來說，MLLR 的作法採用 mean vector transform 就可以改進不少辨識率，從這次的實驗與實驗三相比，可明顯地觀察出對老人語料作 MLLR 的調適有一定的效果；而在增加 diagonal variance transform 之後，多數的研究顯示都只比採用 mean vector transform 好上一些些，但由此次實驗顯示對於我們的老人語料庫作調適採用 diagonal variance transform 能有效地提升調適的效能。

而 regression classes 分的越多，將可依 TCC-300 聲學模型的統計資料，去把更多相異統計特性的 syllable HMM model 中之 mixture 給分到不同 classes 裡，將使得更多組的 regression classes 都能依照其統計特性分別使用 MLLR 來轉換，所以這個實驗也顯示了我們的老人語料庫採用較多組的 regression classes 能獲得較佳的辨識率。

本次實驗的結果是對整體老人語音語料的，我們更想分析老人語料會不會因腔調的差異，而對辨識及調適系統產生不一樣結果，但受限於語料庫的內容仍不夠以

腔調分類來建立個別的辨識系統(如下圖所示)，只能以實驗四的基本老人語音辨識系統計算語者的個別辨識率，再去統計依腔調分類中的辨識率。

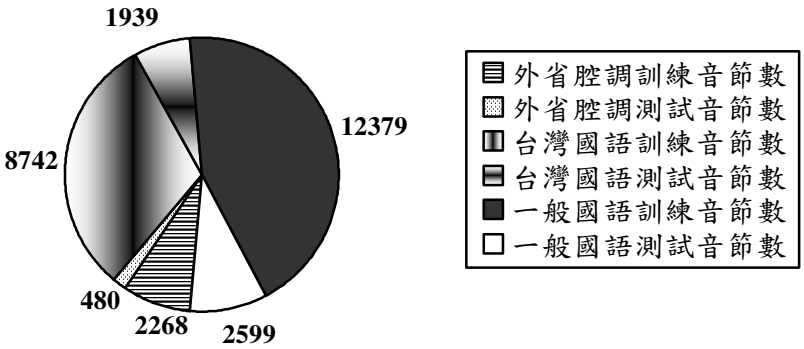


圖 5.3 依腔調分類後語料的統計

5.3 老人語音腔調的差異

首先我們必須如實驗三去對一般國語腔調、台灣國語腔調和外省國語腔調做個別的辨識，並以此為基準去比較上述的基本老人語音辨識系統依腔調分類的辨識率，以及基本老人語音調適系統依腔調分類的辨識率。

『實驗六』由 TCC-300 聲學模型直接對老人腔調分類的語料作辨識

在本次實驗中我們需要將實驗三的結果拿來作比較，並設定 penalty 為一致的 -90，下表即為老人語音依腔調分類後，對 TCC-300 聲學模型分別作辨識的結果：

表 5.2：TCC-300 聲學模型直接對老人腔調分類的語料作辨識之結果

腔調分類	%Del	%Sub	%Ins	%Corr	%Acc
實驗三	4.4	56.57	6.91	39.02	32.12
一般國語	5.03	53.88	5.72	41.10	35.38
台灣國語	3.74	57.38	7.38	38.88	31.51
外省國語	3.75	67.92	11.46	28.33	16.88

由這次實驗中，可清楚地發現整體辨識率被外省腔調的國語給大幅拉低，台灣國語也比一般的老人國語腔調對一般年齡的語音辨識器來得更不匹配。

『實驗七』以 MLLR 對老人腔調分類的語料作調適

實驗六顯示了腔調對於辨識有著顯著地影響，現在來觀察作 MLLR 調適能否有效改善腔調的問題，我們先將腔調依上個實驗分類，再個別對這三種腔調的語料進行對 TCC-300 的聲學模型作 MLLR 的調適過程，此次實驗的 optimal penalty 為-70，由於老人語料又被分成三種資料，所以 MLLR 中 regression classes 設定為 64，其個別的結果如下表所示：

表 5.3 MLLR 對老人腔調分類的語料作調適的結果

腔調分類	%Del	%Sub	%Ins	%Corr	%Acc
實驗五	5.06	45.91	2.45	49.02	46.57
一般國語	5.39	44.59	3.77	50.02	46.25
台灣國語	4.13	47.4	4.95	48.48	43.53
外省國語	2.92	50	3.96	47.08	43.12

在這個實驗與(實驗五)以 MLLR 對老人語料進行調適中 regression classes = 128(採用 128 的原因是由於圖 5.3 中整個老人語料被腔調分類約略被一般國語與台灣國語給分成兩類)的實驗相比，並且與(實驗六)由 TCC-300 直接對老人腔調分類的語料作辨識的結果比對(如圖 5.4 所示)，我們發現採用 MLLR 的調適過程可以有效的降低因腔調的差異所產生的影響。

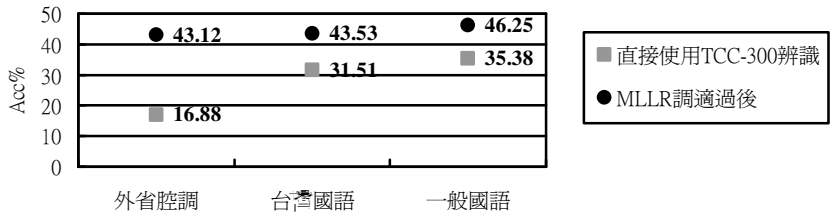


圖 5.4 腔調分類經 MLLR 調適過後的效果

『實驗八』由基本老人語音辨識器分別計算依腔調分類後的辨識率

由於老人語料庫的資料量不足以依腔調分類來各自建立辨識系統，所以我們選用實驗四中 penalty 為 -40 的辨識結果為基準，並且求出每一位語者的個別辨識率，並依腔調分類將這些語者辨識細節作統計，並計算出辨識率(Acc%)，下表為此次實驗的結果：

表 5.4 由基本老人語音辨識器分別計算依腔調分類後的辨識率

腔調分類	(本實驗) 老人語音辨識器 之辨識率	(實驗六) 由 TCC-300 聲學模 型辨識之結果	(實驗七) 由 MLLR 調適 之辨識率
一般國語	43.90	**35.38**	46.25
台灣國語	**45.80**	31.51	43.53
外省國語	41.25	16.88	43.12

這次實驗的結果與實驗六對照後產生出了一個有趣的現象，實驗六的辨識結果以一般腔調國語最佳，而台灣國語腔調次之，這代表著一般腔調國語為 TCC-300 聲學模型中之主要腔調；而老人語料和老人語音辨識器則出現了台灣國語腔調最為匹配，雖然台灣國語腔調的語料量少於一般腔調國語，但老人語音辨識器的腔調仍偏向於台灣國語腔調。以上只是針對我們所建立的老人語料庫，當往後的語料庫建立得較龐大時(特別是將來錄音者在台灣普及教育的環境下，腔調差異也會跟著下降)，老人語音辨識系統極有可能變成一般國語為主要腔調，也能對腔調做更進一步的分析。

5.4 聲道長度正規化

接下來我們將做聲道長度正規化(Vocal Tract Length Normalization, VTLN)的動作，這個動作能把老人語音語料的頻率彎曲(frequency warping)成較接近一般年齡的語音，再來對 TCC-300 聲學模型作辨識則會有較好的辨識率，下圖所示為 frequency warping 的轉換圖：

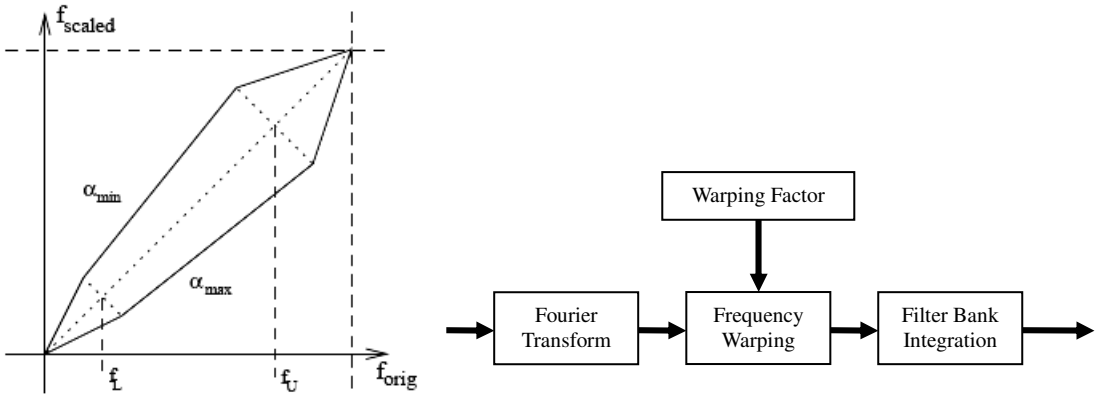


圖 5.5 Frequency Warping 示意圖與方塊圖

由圖 5.5 右圖所示，是 frequency warping 位於整個辨識系統的位置，可參考第四章中抽取語音參數的過程。圖 5.4 左圖描述的是 frequency warping 的示意圖，frequency warping 是把老人語料的頻率採用 piecewise warping 的方式轉換到與 TCC-300 acoustic model 較相似，圖中的虛線即為沒有改變頻率的 warping factor(亦即為 $\alpha = 1$)，而 α_{max} 與 α_{min} 分別為降低與調高老人語料頻率的 warping factor；在使用 VTLN 時，需要去設定預定轉換的頻率範圍 lower and upper boundary frequency($f_L \sim f_U$)，我們參考國外研究的範圍[16]，並由實驗去找最佳的 $f_L \sim f_U$ ，而這裡採取的 f_L 為 100Hz、 f_U 為 6000Hz。

在進行VTLN前，必須先把每一位老人語者語料對TCC-300 acoustic model的最佳 α 值求出來，首先將老人訓練語料對TCC-300模型作forced alignment，進而統計各個 α 值的likelihood值，再依合理的 α 值範圍(0.8~1.2)去找尋各個語者likelihood值最高的 α 值，這個likelihood值並不包含silence與short pause的分數，圖5.5為男、女語者 α 值與人數的統計。

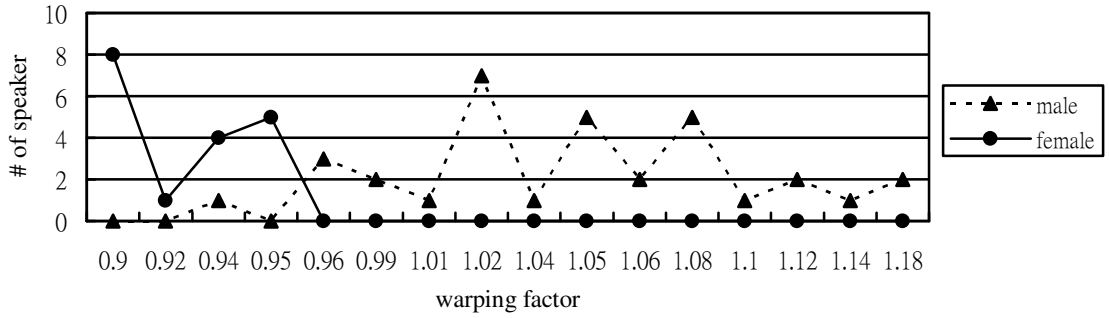


圖5.6 男、女語者α值與人數的統計

上圖中α值準確度為0.01，顯而易見地女性語者α值都落於0.95以下(與TCC-300模型相比其聲音頻率較高需要降低其頻率)，男性語者多半落在α值1以上(聲音較低沉需要調高其頻率)；一般的VTLN實驗時，圖5.6應當呈現男性語者為一個類似平均α值大於1一些的高斯分佈，女性語者為一個類似平均α值小於1一些的高斯分佈，而我們這裡並沒有如此趨勢，有兩種可能：1.語者數不足以產生這種高斯分佈，2.如同第三章對老人語音特性的分析，老人語音的頻率變化呈現極難預估的變動，在此仍希望待語料量充足後再來分析。

『實驗九』語者做過VTLN後對TCC-300 acoustic model作辨識

接下來將把求得的α值對每一位語者的測試語料重新做參數抽取，並對TCC-300 acoustic model作辨識，以下為其結果：

表5.5 語者VTLN對TCC-300 acoustic model作辨識

	%Del	%Sub	%Ins	%Corr	%Acc
實驗三	4.41	56.57	6.91	39.02	32.12
本實驗	4.18	55.04	7.13	40.77	33.64

由上表可知，在做過VTLN後辨識率提升約略1.5個百分比，這是對於整體的辨識率，而我們對於每一位語者的α值、likelihood的變化以及辨識率的增量之間相關性更感興趣，以下將把這三者的關係顯示於下圖之中：

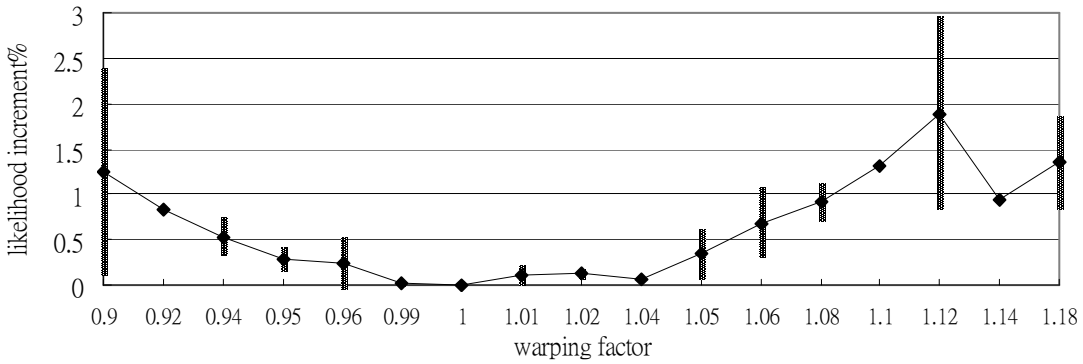


圖5.7 warping factor與likelihood increment的關係

圖中likelihood increment%定義如下式：

$$\text{likelihood increment\%} = \frac{(\text{likelihood when } \alpha \text{ is best}) - (\text{likelihood when } \alpha = 1)}{\text{likelihood when } \alpha = 1} \%$$

(5.6)

黑點代表相同warping factor語者其likelihood increment的平均值，而黑點上具有的直條代表著其likelihood increment一倍標準差的範圍。合理來說，上圖中的平均值分佈應當呈現以$\alpha=1$為底點的倒V字形，也就是$\alpha=1$時likelihood increment = 0，而當α值離1越來越遠時，其likelihood increment也將跟著上升，其代表的意義為當α值離1越來越遠時，語者的頻率被warping得較多，與原先的頻率相差愈多。在圖5.7中，也稍微呈現著以$\alpha=1$為底點的倒V字形，相信只要語者再增加後會有更明顯的形狀。

以下將比較warping factor與(Acc% increment)%的關係，而(Acc% increment)%定義如下式：

$$\text{(Acc\% increment)\%} = \frac{(\text{Acc\% by TCC-300 model with VTLN}) - (\text{Acc\% by TCC-300 model without VTLN})}{\text{Acc\% by TCC-300 model without VTLN}} \%$$

(5.7)

合理來說，Acc% increment與likelihood increment應為正相關的關係，但儘管如此，由國外相關的研究顯示[16]，最高的likelihood值求出的α值不一定能保證Acc% increment會是正的，也就是說，不保證每一位語者在做了VTLN後會提升辨識率，但是整體的辨識率會提升。而下圖為warping factor與(Acc% increment)%的關係圖，

其中黑色方形代表男性語者、白色三角形代表女性語者，我們也希望下圖會呈現以$\alpha=1$為底點的倒V字形，可惜的是成效並沒有那麼明顯，假若忽略負的Acc% increment，這個圖形將會較接近我們的預期。

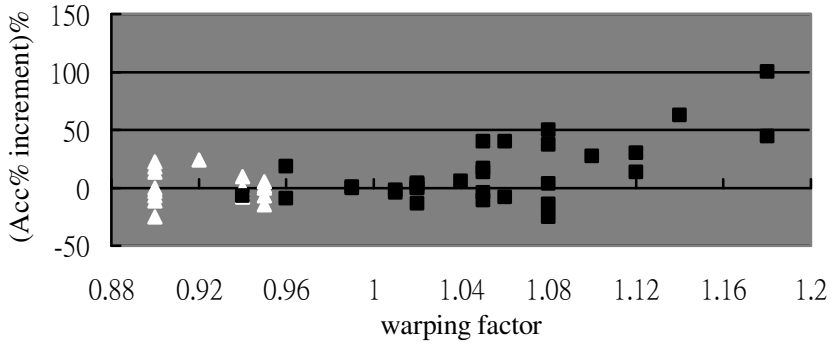


圖5.8 warping factor與(Acc% increment)%的關係

這裡顯示了VTLN對每一位語者的資訊，接下來將把VTLN與MLLR一起進行老人語音語料的調適。

『實驗十』加入VTLN後的MLLR調適

以下先採用VTLN對老人調適與測試語料(亦即為實驗五的語料)做frequency warping得動作，再進行MLLR的調適過程，並以此觀察VTLN與MLLR是否有相輔相成的效果。以下為加入VTLN後的MLLR調適的結果：

表5.6 加入VTLN後的MLLR調適

Regression Classes	%Del	%Sub	%Ins	%Corr	%Acc
32	5.08	47.1	2.65	47.83	45.18
64	5.12	45.64	2.51	49.24	46.73
128	4.96	43.98	2.51	51.06	48.55
256	5.38	40.91	2.23	53.71	51.47

這次實驗的辨識率Acc%可以和(實驗五)以MLLR對老人語料做調適後的辨識率Acc%做比較，而(實驗三)以TCC-300對老人語料作辨識的結果與(實驗九)語者做過VTLN後對TCC-300作辨識的結果做比較，將上述兩者放在一起便可以觀察出我們感興趣的資訊，也就是VTLN與MLLR是否有相輔相成的效果。下圖為增加VTLN後對辨識與調適的效能增量評估：

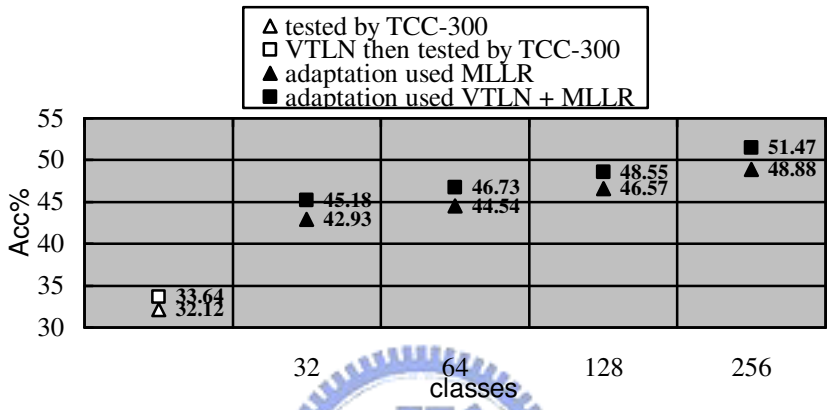


圖5.9 加入VTLN後對辨識與調適的效能增量

依據上圖的資訊可知，單獨對老人語料做frequency warping的動作，能提升對TCC-300模型作辨識時辨識效能Acc%約1.5個百分比；而當採用MLLR對老人語料做調適時，先於抽取語料時進行frequency warping再作調適，約可提升辨識效能Acc%約2~2.5個百分比，如果圖5.9辨識率之間的差距為等距的話，那麼代表著VTLN對MLLR的調適是完全沒有幫助的，經由這個實驗可以確定VTLN的做法對老人語音的辨識與調適都有小幅度的改進。

接下來將使用這個實驗產生出來的聲學模型(regression classes = 256)再對老人語者作一次VTLN的動作，並希望求出來的α值能更逼近1，而使用這個α值來對老人調適語料與測試語料進行抽取參數的動作，進而對這個新的聲學模型重複MLLR調適的動作。雖然每個語者α值的likelihood值都有下降，但可惜的是重新求出的α值僅有七個語者的α值有變化，大多數都與第一次求α值時相同(第一次與第二次做VTLN時，α值與其對應之likelihood值的變動，詳見附錄四)，以此也可預期重複MLLR調適的動作只會有微量的辨識率提升。在先前提到的男、女語者α值應該呈

現接近1的高斯分布，由這兩次VTLN的動作可以比較α值的平均值與標準差的變化，下表統計了男、女語者α值對新(VTLN+MLLR調適過後所產生)、舊(TCC-300)聲學模型產生的變化：

表5.7 二次VTLN過程中warping factor之統計性質

	男性語者		女性語者	
	平均值	標準差	平均值	標準差
對TCC-300聲學模型求α值	**1.0482**	**0.0589**	0.9239	0.0230
對VTLN+MLLR調適過後所產生的聲學模型求α值	**1.0382**	**0.0485**	0.9222	0.0213

上表可看出女性語者的α值沒有因新的聲學模型產生多少改變，但是男性語者在經過一次VTLN+MLLR的調適之後，老人男性語料與新聲學模型的聲音頻率高低較TCC-300聲學模型來得更相似，也就是說α值更接近1，而且α值的標準差較先前更集中了。並且觀察α值與其對應之likelihood值的變動，如圖5.10所示：

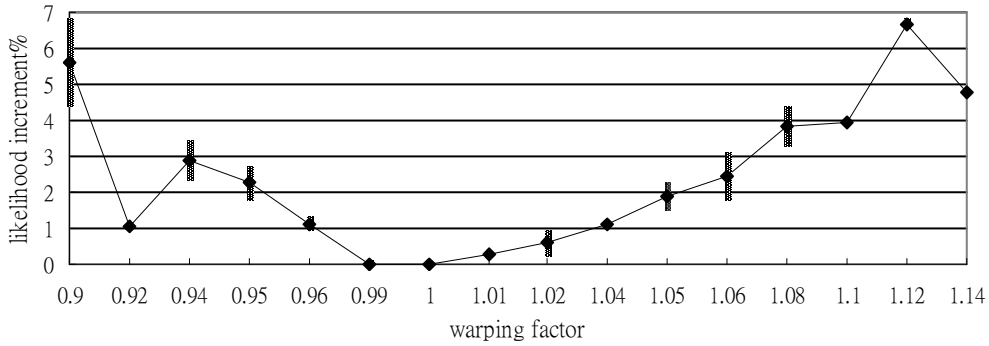


圖5.10 新聲學模型的warping factor與likelihood increment的關係

對照圖5.7的Y軸，可清楚發現經VTLN新聲學模型的likelihood increment較第一次做VTLN時增加了約三倍，對於被調適過聲學模型，上圖之呈現較符合男、女語者聲道長度正規化，而不是針對老人語音聲道長度的改變。

『實驗十一』重複VTLN+MLLR調適

重新對第一次VTLN+MLLR調適所產生的新聲學模型求出其α值(2nd warping factor)，並以此α值對調適語料、測試語料做frequency warping，而抽取出warping過後的語音特徵參數，最後再進行一次MLLR對新聲學模型進行調適，當中的過程可由下兩圖描述之：

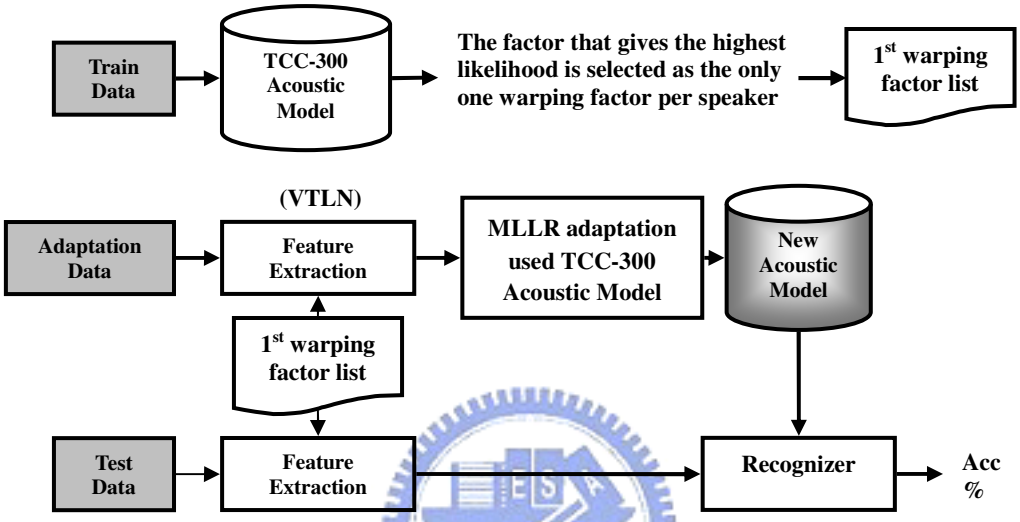


圖5.11 第一次VTLN+MLLR調適流程方塊圖

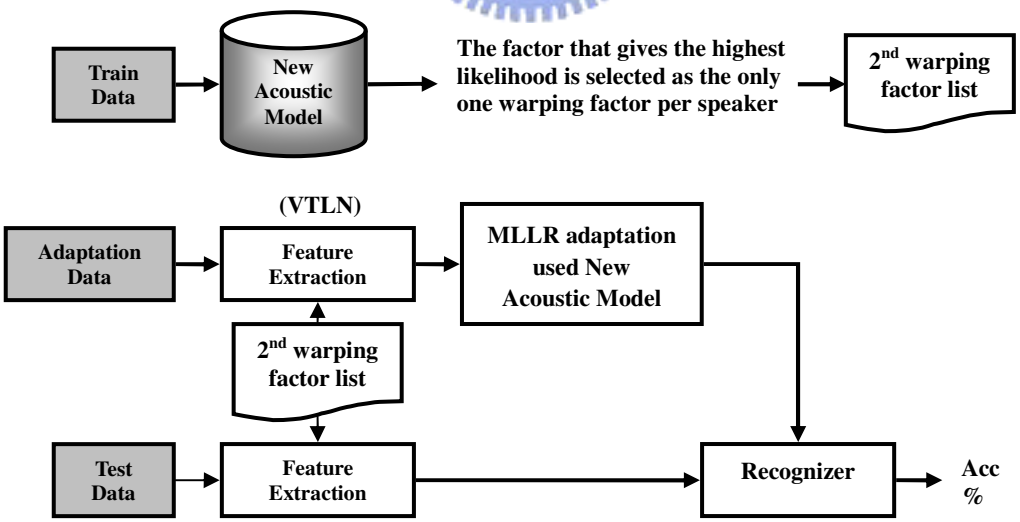


圖5.12 第二次VTLN+MLLR調適流程方塊圖

由前面的敘述可知這次實驗的流程(圖5.12)，而2nd與1stwarping factor list之間只

有七位語者改變了α值，這代表著在進行第一、第二次VTLN時，並不會有太大的改變，預計辨識率的提升也有限，而老人語料經由二次VTLN+MLLR調適之後，可以得到以下的辨識結果：

表5.8 二次VTLN+MLLR調適之辨識結果比較

	classes	%Del	%Sub	%Ins	%Corr	%Acc
1st VTLN+MLLR adaptation	256	5.38	40.91	2.23	53.71	51.47
2nd VTLN+MLLR adaptation	256	5.18	41.65	2.17	53.17	51.00

這次的結果顯示出了對VTLN+MLLR做iteration的動作不會有進一步地好處，這是由於第一次VTLN+MLLR所產生的新聲學模型是把TCC-300模型作轉換的動作，使得模型中有332個mixture在此過程中會遺失，也代表著老人語料庫仍不夠大，足以去實行有效的2nd VTLN+MLLR調適。這個結果顯示了對相同的老人語料只能做一次MLLR及VTLN。

第六章 結論與未來展望

6.1 結論

在論文中使用TCC-300聲學模型作為老人語音辨識系統比較的基準；並依老人語音腔調的不同來分析其對老人語音辨識系統的差異；然後利用MLLR來對TCC-300聲學模型進行老人語料的調適，最後依據老人語音特性：聲帶變長、聲道變短而導致老人聲音較青、壯年人來的低沉，所以採用VTLN方法來對老人語料進行frequency warping的動作，將老人聲音的頻率調整至與一般年齡層較相似，並結合MLLR來對TCC-300聲學模型進行老人語料的調適；最後把VTLN+MLLR產生出來的新聲學模型再一次地求取老人語料對此模型更適當frequency warping factor，再作一次VTLN+MLLR的調適，經由一連串的實驗與分析，歸納出以下的結論：

1. 對於我們的老人語料庫，直接對TCC-300聲學模型作辨識與自行建立的老人語音辨識系統作辨識，兩者的辨識率(Acc%)分別為33.12%與44.72%，以此可見老人語音特性與一般年齡層有著不小的差異，若對照TCC-300聲學模型採用TCC-300語料之辨識結果(Acc% = 72.4%)相比，更可佐證此一論點。
2. 使用MLLR來對TCC-300聲學模型進行老人語料的調適時，regression classes越多，則調適的效果越佳，而採用diagonal variance transform(當mixtures分成256 classes時，Acc% = 48.88%)遠比mean vector transform(43.76%)更有效。
3. MLLR調適老人語料可有效地降低因腔調不同所產生的辨識結果差異。
4. VTLN與MLLR調適將會產生相輔相成的作用，而由第一次VTLN與MLLR調適所產生的新聲學模型進行再一次的調適，辨識率並不會有些微提升，使得最佳的調適為1st VTLN+MLLR，其syllable辨識率達到51.47%。

6.2 未來展望

由於目前具有的老人語音資料庫只有三萬多個音節數，導致分析出來的結果之可靠性不足，希望未來在擴充語料量之後，可以對上述的結果進行更詳盡的分析；對於老人腔調的差異，也能在足夠的語料量下各自建立其辨識系統進行更仔細的探

討。

論文中採用的VTLN方法為piecewise warping，由於目前老人聲道長度對語音特性改變之相關研究，仍無法明確指出老人語音特性具有何種變化規則，而VTLN有另一種方法為bilinear warping，未來可以試著由bilinear warping來得到更符合老人語音特性的warping curve。

參考文獻

【1】吳光明，”呼吸氣流對發音特徵與模型影響之研究”，台灣科技大學，中華民國九十四年七月

【2】P.B. Denes and E.N. Pinson,"The Speech Chain", 1993, W.H. Freeman New York, p.5

【3】http--www.telecom.tuc.gr-~ntsourak-tutorial_acoustic.htm

【4】An Xue, Jack Jiang, Emily Lin and Peter B. Mueller,"Age-related changes in human vocal tract configurations and the effects on speakers' vowel formant frequencies: a pilot study",Ohio University, Athens, OH, Northwestern University Medical School, Chicago, IL, Kent State University, Kent OH, USA,Log Phon Vocol 1998; 24: 132–137

【5】Linville SE, Fisher HB. "Acoustic characteristics of women’s voices with advancing age." J Gerontol 1985; 40:324–30.

【6】Rastatter MP, Jacques RD. "Formant frequency structure of the aging male and female vocal tract." Folia Phoniatr 1990; 42: 312–9.

【7】Association for Computational Linguistics and Chinese Language Processing Institute of Information Science, Academia Sinica,"MAT-2000" and "MAT-2500"

【8】Heidi Christensen, Ove Andersen, Borge Lindberg, "Speaker Adaptation of Hidden Markov Models using Maximum Likelihood Linear Regression" Aalborg University, 1996

【9】Steve Young, Gunnar Evermann, Mark Gales, etc. "The HTK Book (for HTK Version 3.4)", Cambridge University Engineering Department, 2001-2006

【10】梁振豐，"台語語音辨識及智慧型口語對話汽車導航系統"，國立交通大學，中華民國九十五年八月

【11】張隆勳，"國語廣播新聞語音基本系統之建立"，國立交通大學，中華民國九十四年六月

【12】Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, "Spoken Language Processing, A guide to Theory,Algorithm, and System Development," Prentice-Hall, Inc.

【13】王小川，"語音訊號處理"，全華科技圖書，中華民國九十三年

【14】陳克巽，"非監督式快速語者調適演算法研究"，國立中央大學，中華民國八十九年六月

【15】X. Huang and K.F. Lee, "On Speaker-Independent, Speaker-Dependent, and Speaker-Adaptive Speech Recognition." IEEE Trans. on Speech and Audio Proc., Vol. 12, pp. 150–157, April 1993

【16】Puming Zhan and Alex Waibel, "Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition", CMU-CS-97-148, May 1997

附錄一 老人語料庫基本資料統計

	年紀	腔調	偶發的背景音	錄音品質	# of filler	# of noise	# of b.s.n.	錄音環境
f001	65	台灣國語	人聲、鳥叫聲	一般	7	15	8	室外
f002	72	一般國語		優良	2	3		室內
f003	63	一般國語	人聲	低劣			1	室外
f004	73	台灣國語	人聲	一般	7	9		室內
f005	77	台灣國語		優良	1			室內
f006	71	台灣國語		優良				室內
f007	76	一般國語		優良	2			室內
f008	66	台灣國語	人聲	一般	2			室內
f009	69	一般國語		優良	1			室內
f010	70	台灣國語	鳥叫聲	一般			1	室外
f011	61	一般國語		優良				室內
f012	65	台灣國語		優良	15			室內
f013	63	一般國語		一般	1			室外
f014	82	外省腔調國語	人聲	低劣		2	1	室內
f015	66	一般國語		優良				室內
f016	62	一般國語		優良				室內
f017	75	台灣國語	鳥叫聲	一般				室外
f018	69	一般國語	鳥叫聲	一般	1			室外
f019	65	一般國語	人聲	一般		11	27	室內

	年紀	腔調	偶發的背景音	錄音品質	# of filler	# of noise	# of b.s.n	錄音環境
m001	72	一般國語	人聲	低劣				室外
m002	76	外省腔調國語	人聲	低劣	1			室外
m003	60	一般國語	人聲	低劣	3		7	室外
m004	82	一般國語		優良				室內
m005	65	台灣國語	鳥叫聲	一般	2	12		室外
m006	85	外省腔調國語		優良		1		室內
m007	68	一般國語		優良		12		室內

	年紀	腔調	偶發的背景音	錄音品質	# of filler	# of noise	# of b.s.n	錄音環境
m008	66	台灣國語		一般				室內
m009	62	台灣國語		優良	4	1		室內
m010	71	一般國語		優良		2	1	室內
m011	66	一般國語		優良			1	室內
m012	63	一般國語	人聲	一般				室內
m013	67	一般國語	人聲	一般		3		室內
m014	68	台灣國語		一般	3	8		室內
m015	63	一般國語		優良				室內
m016	60	一般國語		優良	2			室內
m017	78	外省腔調國語	鳥叫聲	一般	3	3		室外
m018	78	一般國語	鳥叫聲	一般	1	14		室外
m019	66	台灣國語	鳥叫聲	低劣	6	3		室外
m020	71	一般國語		優良				室內
m021	81	外省腔調國語		優良		3		室內
m022	67	台灣國語		優良		5		室內
m023	63	台灣國語		優良	1			室內
m024	67	一般國語		優良	1	3		室內
m025	80	一般國語	車聲、鳥叫聲	低劣	6			室外
m026	60	一般國語	車聲	一般				室外
m027	70	台灣國語	人聲、鳥叫聲	一般				室外
m028	71	台灣國語		優良	6	22	1	室內
m029	76	外省腔調國語	人聲	一般		1	5	室內
m030	73	一般國語	人聲	低劣	5	1		室外
m031	67	一般國語	人聲	一般	6	13	5	室外
m032	65	台灣國語		優良		10	6	室內
m033	81	外省腔調國語		優良				室內
m034	64	台灣國語		優良				室內
m035	74	一般國語	人聲	一般	2	9		室內

註解：錄音品質的區分是根據整體的背景雜訊而定，低劣的錄音品質中可能包含了大量的 noise、background speaker noise(b.s.n.)，但由於這些雜訊過多，所以不將這些雜訊額外統計於倒數第二與第三欄中。

附錄二 TCC-300 聲學模型

Configuration of Feature Extraction

```
# byte order
#BYTEORDER=VAX
NATURALREADORDER=TRUE
NATURALWRITEORDER=TRUE
# Waveform parameters
SOURCEFORMAT=ALIEN
HEADERSIZE=4096
SOURCERATE=625.0
# Coding parameters
TARGETKIND=MFCC_E
TARGETRATE=100000.0
SAVECOMPRESSED=F
SAVEWITHCRC=T
WINDOWSIZE=320000.0
ZMEANSOURCE=T
USEHAMMING=T
PREEMCOEF=0.97
NUMCHANS=24
USEPOWER=F
CEPLIFTER=22
LOFREQ=0
HIFREQ=8000
NUMCEPS=12
ENORMALISE=T
DELTAWINDOW=2
ACCWINDOW=2
ALLOWCXTEXP=F
```

Configuration of Training

```
# byte order
#BYTEORDER=VAX
NATURALREADORDER=TRUE
NATURALWRITEORDER=TRUE
# MFCC parameters
SOURCEFORMAT=HTK
SOURCERATE=100000.0
TARGETKIND=MFCC_E_D_A_N_Z
TARGETRATE=100000.0
DELATWINDOW=2
ACCWINDOW=2
#new variable can replace the varFloor
VARFLOORPERCANTILE = 20
```

註解：以上的設定檔當中的詳細內容，可參考 HTK book version 3.4

附錄三 修改過後的 phone table

音碼	注音	漢語拼音	子音拼音	母音拼音
64	*ㄝ*	*eh*	*INULL_e*	*eh*
102	*ㄘㄡ*	*cou*	*c_o*	*ou*
103	*ㄙㄡ*	*sou*	*s_o*	*ou*
109	*ㄋㄡ*	*nou*	*n_o*	*ou*
408	*ㄧㄛ*	*yo*	*INULL_y*	*yo*
412	**各種語言現象**	**filler**	**filler**	

註解：此附錄在一般常見的 411 音碼表(省略)中增加(黑粗體字)了 filler，並且刪除(灰階斜體字)了具有 c_o, n_o, s_o, eh, yo 這五個 sub-syllable 的 syllable，因此老人語音辨識器所使用的 syllable 為 406 個。

附錄四 老人語者 warping factor 與 likelihood 值的統計資料

	1st VTLN				2nd VTLN			
speaker	warping factor	likelihood (α=1)	likelihood (best α)	likelihood increment	warping factor	likelihood (α=1)	likelihood (best α)	likelihood increment
f001	0.9	-31046.66	-30523.25	523.411	0.9	-29456.03	-27376.2	2079.834
f002	0.9	-30154.06	-29899.16	254.9052	0.9	-28681.6	-27066.63	1614.97
f003	0.9	-31776.64	-31434.58	342.0534	0.9	-28742.39	-27564.77	1177.617
f004	0.92	-26616.6	-26395.48	221.1223	0.92	-23625.99	-23383.04	242.9419
f005	0.95	-32390.39	-32338.21	52.17372	0.95	-28385.98	-27832.12	553.8558
f006	0.94	-32751.5	-32499.23	252.263	0.94	-29532.24	-28603.47	928.7704
f007	**0.95**	-33504.35	-33400.81	103.5351	**0.94**	-30361.2	-29340.02	1021.187
f008	**0.95**	-30206.3	-30130.29	76.01388	**0.94**	-26989.47	-26103.79	885.6781
f009	0.9	-32459.61	-31223.97	1235.639	0.9	-30033.24	-28620.26	1412.983
f010	**0.94**	-30073.81	-29881.23	192.587	**0.95**	-27194.23	-26498.21	696.0112
f011	0.9	-29145.6	-29037.56	108.044	0.9	-27418.07	-26031.04	1387.031
f012	0.94	-29110.73	-29012.34	98.38758	0.94	-26341.85	-25675.24	666.6157
f013	**0.95**	-27838.19	-27716.45	121.7349	**0.94**	-25407.8	-24723.66	684.1453
f015	0.94	-26914.12	-26767.53	146.5967	0.94	-24578.74	-23794.87	783.8749
f016	0.9	-31865.48	-31549.09	316.39	0.9	-30255.18	-27967.82	2287.356
f017	0.9	-28126.4	-28055.74	70.65582	0.9	-26077.99	-24749.91	1328.078
f018	**0.95**	-30211.18	-30131.94	79.23372	**0.94**	-27066.87	-26221.54	845.3296
f019	0.9	-32506.01	-32197.88	308.1324	0.9	-29828.26	-28169.08	1659.174
m001	1.02	-30938.94	-30907.97	30.96549	1.02	-27429.66	-27301.8	127.8671
m004	**1.18**	-32195.57	-31649.98	545.5945	**1.02**	-27333.27	-26996.24	337.0355
m005	1.04	-32466.78	-32445.43	21.34227	1.04	-28970.13	-28611.63	358.5053
m006	1.02	-27738.07	-27689.11	48.95851	1.02	-24790.83	-24691.11	99.72026
m007	0.94	-28339.52	-28241.11	98.41373	0.94	-26142.73	-25653.25	489.4796
m008	1.02	-36161.63	-36120.36	41.26611	1.02	-32744.29	-32617.65	126.6381
m009	1.02	-34482.81	-34429.34	53.46401	1.02	-30354.58	-30156.17	198.4119
m010	0.99	-24541.36	-24538.19	3.169024	0.99	-22122.08	-22120.87	1.211983
m011	1.05	-29728.04	-29694.7	33.33838	1.05	-27303.37	-26828.2	475.1746
m012	1.08	-34253.84	-34001.11	252.7276	1.08	-30828.19	-29697.31	1130.877

	1st VTLN				2nd VTLN			
speaker	warping factor	likelihood (α=1)	likelihood (best α)	likelihood increment	warping factor	likelihood (α=1)	likelihood (best α)	likelihood increment
m013	1.12	-35976.37	-35031.02	945.3471	1.12	-33749.46	-31509.36	2240.104
m014	0.99	-31061.39	-31057.73	3.659202	0.99	-28849.81	-28836.53	13.28258
m015	0.96	-34863.58	-34678.58	185.001	0.96	-31553.79	-31147.57	406.2244
m016	1.08	-30635.69	-30290.5	345.19	1.08	-28798.68	-27439.96	1358.72
m017	1.05	-35736.41	-35489.07	247.3386	1.05	-31339.87	-30646.69	693.1862
m018	0.96	-24672.65	-24640.79	31.86196	0.96	-21748.02	-21531.32	216.7003
m019	1.06	-32293.58	-32153.92	139.6659	1.06	-28683.81	-28112.53	571.2827
m020	1.02	-33346.36	-33316.59	29.77238	1.02	-29230.89	-29134.91	95.98081
m021	1.08	-35538.34	-35166.61	371.7323	1.08	-31060.73	-29904.81	1155.919
m022	1.02	-31516.72	-31483.43	33.28325	1.02	-27830.47	-27708.47	122.0026
m023	1.01	-37135.5	-37121.42	14.08222	1.01	-32209.7	-32128.78	80.91806
m024	**1.01**	-32521.4	-32467.41	53.99497	**1.02**	-28206.64	-27958.37	248.2781
m025	0.96	-27125.52	-27108.39	17.13185	0.96	-23583.32	-23319.21	264.1079
m026	1.05	-28807.6	-28753.31	54.2909	1.05	-26363.7	-25769.62	594.0773
m027	1.05	-26551.16	-26409.58	141.5769	1.05	-23230.69	-22894.77	335.9277
m028	1.08	-28480.22	-28283.07	197.15	1.08	-25696.33	-24821.27	875.0649
m029	1.08	-27930.47	-27664.8	265.6737	1.08	-24482.17	-23577.07	905.1042
m030	1.14	-21878.47	-21673.71	204.7597	1.14	-19787.72	-18837.12	950.6068
m031	1.12	-32763.54	-32383.5	380.0494	1.12	-30393.19	-28343.35	2049.85
m032	1.05	-27954.03	-27900.99	53.03667	1.05	-24348.33	-23922.26	426.0686
m033	**1.18**	-30348.82	-30044.09	304.7296	**1**	-26505.12	-26505.12	0
m034	1.1	-29388.86	-29005.19	383.6699	1.1	-27090.02	-26019.96	1070.066
m035	1.06	-31927.26	-31627.58	299.6789	1.06	-28682.63	-27854.2	828.428

註解：當中的1st VTLN為使用老人訓練語料對TCC-300 acoustic model進行force alignment所計算出來的likelihood與warping factor，而2nd VTLN是對(第一次VTLN+MLLR調適所產生出來的)新聲學模型進行force alignment，進而計算出上列數值。