國 立 交 通 大 學

電信工程研究所

碩士論文

自動中文音素錯誤偵測器
Automatic Mandarin Phone Error Detector

研 究 生：賴智誠

指導教授：王逸如 博士

中華民國一百年九月

自動中文音素錯誤偵測器

Automatic Mandarin Phone Error Detector

研 究 生：賴智誠　　Student：Chih-Chen Lai

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

September 2011

Hsinchu, Taiwan, Republic of China

中華民國一百年九月

自動中文音素錯誤偵測器

研 究 生：賴智誠　　指導教授：王逸如 博士

國立交通大學電信工程學系碩士班

中文摘要

傳統上音素錯誤偵測器以事後機率作為信心度指數，跟門檻值比較之後可決定一個音素的正確性，而本論文提出兩種比基本系統好的方法，都是以多層感知器為基礎的錯誤偵測器，其中最主要的概念是引入多維的事後機率向量，而兩種多層感知器偵錯系統的差異在於訓練資料的部分，期許利用多層感知器網路的學習特性，抽取有用的訓練語料，以訓練出效能較佳的多層感知器網路，最後利用外國人的中文發音語料來測試三個系統的偵錯效能。

Automatic Mandarin Phone Error Detector

Student：Chih-Chen Lai　　Advisor：Dr. Yih-Ru Wang

Department of Communication Engineering

National Chiao Tung University

Abstract

Traditionally, phone error detector uses a posterior probability as confidence measure, the correctness of a phone can be decided by comparing it to its corresponding threshold. In this thesis, two systems better than baseline are proposed, both are phone error detectors based on MLP network. The main concept of MLP-based system is the introduction of multiple-dimension a posterior probability. Besides, the difference between the two proposed MLP system is their training data. Hope that we could improve the performance of the MLP network by utilizing its learning property and taking useful data as training data. At last, we test the error detecting performance of these three phone error detectors by the Mandarin corpus of foreign speakers.

誌謝

能如期完成研究真的是非常感動，為期兩年的碩士生涯不長不短，但要感謝的人真的太多。

首先，感謝陳信宏老師不時在我迷失研究方向時給予明確的建議，並且在家裡有狀況時老師的殷切關懷真的讓學生感動萬分；接著，感謝我的指導老師王逸如老師，在最後的緊要關頭老師真的給我很大的幫助，另外除了研究上的指導，總是告訴我正確的做事方法以及研究之道，兩位老師說過的話學生會銘記在心。

接著是博班學長們，跟超人一樣，一個人 handle 幾乎所有人研究的性獸老大，總是幫我解決問題並聽我 complain 的智合，偶爾會一起去買飲料的阿德，時常幫忙隨手關燈的希群，講話莫名好笑的輝哥，感謝學長們平時給我的建議與鼓勵；感謝 96 級的普烏以及小宋學長時常回來實驗室探望大家，以及感謝小廣學長百忙之中幫忙投遞研替的履歷，感謝 97 級的宥余、承燁、小卡、PUMA、皓翔、蔡依玲學姊、雲舒學姊，感謝學長姐們在我們這屆剛進來的時候給予研究上的建議以及生活上的照顧；再來，感謝一同奮鬥的好同學們，減肥後身輕如燕輕易摸框的大胖、五月天瘋狂 fans 文良、重訓王豆腐、雜技王 and 一哥銘傑、鬼切小瞎、修課王的叛逃啟全、消失的佳緯、常看星海直播的冠繹、時常一起聊天的進竹，有你們的陪伴讓我得以安心完成研究；感謝超有趣的下一屆學弟妹們，打球研究把咩都一把罩的 KIWI、好像每天都很累的睿詮、超會吃蝦和螃蟹的俊翰、又潮又是卷哥的子軒、一起共事過效率超高的企鵝、有 hiphop 魂的昂星，小老闆愛將的雅婷，感謝你們時常提供有的沒的娛樂，真的超有舒壓的效果，在過一年就換你們了，加油啊！感謝從大學就混在一起的好兄弟姊妹們，土匪、大胖、大頭良、啃雞、台中人、宗炮、小豬、金易、大葡、000、林舜華、皮皮、阿派，能認識你們是我的榮幸。

最後我要感謝至始至終都陪在我身邊的老爸老媽老哥以及女朋友星萍，你們的體貼與支持是我完成研究的最大動力。

目錄

圖目錄

表目錄

第一章 緒論

1.1 研究背景與動機

在學習第二外語的人數激增以及年齡層下降的現今世代，有別於傳統的真人教學，電腦輔助語言學習(Computer Assisted Language Learning, CALL)系統提供學習者一個無壓力且有彈性的學習環境；然而，在“說”的方面，以目前的技術來說，CALL 系統所能提供的學習成效仍遠不及真人教學，但在發音品質評估的部分，電腦擁有人所不能達到的一致性(consistency)，因此還是有探討其效能的價值。

發音評估(Pronunciation Assessment)和錯誤偵測(Error Detection)是 CALL 系統在語音方面的兩個主要研究方向，前者致力於找出接近人類語言學家的評分方法，常見的是測試語句“標準答案已知”的做法，一般引用傳統自動語音辨識(Automatic Speech Recognition, ASR)技術，使用隱藏式馬可夫模型(Hidden Markov Model, HMM)對測試語句強制對齊(forced align)以得到音素(phone)邊界，再求取語句中各音素的 likelihood 值，接著轉換成合理的個體分數或整體平均分數；後者常會使用發音評估所獲得的評分方法，另外加上門檻值(threshold)的制定，即可決定一個音素的正確與否，達到錯誤偵測的效果。

本論文主要探討音素層級(phone level)的中文發音錯誤偵測，傳統上多以 likelihood-based 的一維音素事後機率為信心度評估(Confidence Measure)指數，因此期許透過類神經網路(Neural Network)中的多層感知器(Multilayer Perceptrons, MLP)，引用多維音素事後機率參數，以達到較佳的錯誤偵測效能。

1.2 文獻回顧

以下列出幾個與錯誤偵測相關的早期文獻：

Ronen 等人提出錯誤發音網路的概念【1】，傳統字典裡一個字能拆解成數個音素，這裡將每個音素又細分成 native 和 non-native 兩類，好像形成一個網路一般，接著將測試語句中的每個字的錯誤發音網路串接起來並使用 Viterbi 演算法得到最佳路徑，計算 native 和 non-native 的量後可得到一個發音錯誤分數，以當作發音品質的評估。

S.M. Witt 等人提出以 likelihood 為基礎的 Goodness of Pronunciation(GOP)【2-3】，該方法結合強制對齊和無限制音素辨識(unconstrained phone recognition)的概念計算出語句中每個音素的分數。

Franco 等人提出兩種方法偵測音素錯誤【4】，第一種利用 native speech 所訓練出來的聲學模型(acoustic model)對語料作強制對齊，再分別計算各音素的事後機率分數，第二種是針對各音素分別用 native speech 和 non-native speech 訓練兩種聲學模型，對每個音段以這兩個聲學模型計算出來的 likelihood 得到一個比例值(ratio)作為分數，兩種方法都把分數跟事先決定好的門檻值做比較以決定錯誤，本論文即是採用第一種方法作為偵測錯誤的方法。

1.3 研究方向

首先建立以一維音素事後機率(A Posterior Probability)為特徵參數的音素錯誤偵測器(Phone Error Detector)，以此當作 Baseline 系統；接著建立以多維音素事後機率向量為參考特徵的 MLP-based 音素錯誤偵測器，並以此系統跟 Baseline 作比較，最後拿外國人的中文發音語料來測試兩種方法的錯誤偵測效能。

1.4 章節概要

本論文的內容共分為六章：

第一章 緒論：介紹本論文之研究背景動機與方向。

第二章 聲學模型簡介：介紹本論文所使用的前級聲學辨認器。

第三章 以音素事後機率為基礎之音素錯誤偵測器：介紹基本音素發音錯誤偵測器系統。

第四章 基於多層感知器之音素發音錯誤偵測器：介紹以 MLP 為後級架構的音素偵測器。

第五章 主觀測試實驗：介紹對由標音員標記的外國語料進行錯誤偵測的結果。

第六章 結論與未來展望：介紹本論文的結論以及未來可改進的方向。

第二章 聲學模型簡介

HMM 模型是傳統自動語音辨識技術中最重要的利器，另外不管是發音評估還是錯誤偵測的研究，多是以 HMM 模型的辨認結果為基礎，再發展出各式各樣的評分或是錯誤偵測方式，而本論文的前級也是以 HMM 模型辨識結果為基礎，再另外轉換成音素事後機率參數，因此本章將介紹論文中使用的 HMM 模型。2.1 節簡單介紹訓練聲學模型所用的 TCC300 語料庫，2.2 節介紹聲學參數的抽取與設定，2.3 節介紹三連音素(tri-phone)聲學模型。

2.1 TCC300 語料庫

本語料庫是由台灣大學、成功大學、交通大學各自擁有之語料庫集合而成【5】，各校錄製之目的是為語音辨認研究，屬於麥克風朗讀語音。其中台大語料庫主要包含詞及短句，文章經過仔細設計，考慮了音節及其相連出現機率，由100人錄製而成；成大及交大語料庫主要包含長文語料，文章由中研院提供之500萬詞詞類標示語料庫中選取，每篇文章包含數百字，再切割成3至4段，每段包含至多231字，由200 人朗讀錄製，每人所讀文章皆不相同，統計資料如表2-1所示：

表 2-1：TCC300 語料庫資訊統計表

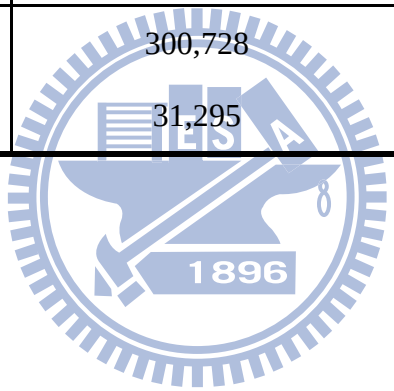
學校名稱	文章屬性	語者總數		總音節數		總檔案數	
台灣大學	短文 (平衡句)	男	50	男	27,541	男	3,425
		女	50	女	24,677	女	3,084
		總計	100	總計	52,218	總計	6,509
交通大學	長文	男	50	男	75,059	男	622
		女	50	女	73,555	女	616
		總計	100	總計	148,614	總計	1,238

		男	50	男	63,127	男	588
成功大學	長文	女	50	女	68,749	女	582
		總計	100	總計	131,876	總計	1,170

整個語料庫又切成兩部分，分別是用來訓練聲學模型的訓練語料以及測試用的測試語料，測試語料的資料量大約是訓練語料的十分之一，詳細的音節音素統計資料如表 2-2 所示：

表 2-2：TCC300 訓練語料和測試語料音節音素統計資訊

	音節數	音素數
訓練語料	300,728	837,597
測試語料	31,295	87,191



2.2 聲學參數

訓練聲學模型所使用的參數是梅爾頻率倒頻譜係數(Mel-scale Frequency Cepstral Coefficients, MFCC)，以 32 毫秒之漢明窗(hamming window)以及 10 毫秒的音框位移(frame shift)求取 12 維MFCC加上 1 維能量係數，再求取此 13 維MFCC的一階(delta)以及二階(delta-delta)變化量，以去除能量係數後的 38 維MFCC為特徵向量，並且執行倒頻譜平均值正規化(Cepstrum Mean Normalization, CMN)，藉以消除不同語音訊號的通道效應。

2.3 三連音素聲學模型

本實驗所使用的聲學模型為不特定語者(speaker independent)、跨詞三連音素(cross-word tri-phone)隱藏式馬可夫模型，平均每個狀態(state)使用 16 mixture 的混和高斯模型(Gaussian Mixture Model, GMM)，除了短停頓(short pause)模型僅使用 1 個狀態之外，其他模型皆使用 3 個狀態。聲學模型所使用的音素總共有 38 個，詳細標記和發音對照請參考附錄一。

為檢察所訓練聲學模型之效能，在沒使用語言模型的情況下，三連音素聲學模型對 TCC300 測試語料的音節辨認率為 70.19%，有音節連接限制的音素辨認率為 84.17%。

第三章 以音素事後機率為基礎之音素發音錯誤偵測器

傳統HMM-based的自動語音辨識的準則(criterion)為Maximum Likelihood，但likelihood是絕對值，沒有相對的概念，因此一般信心度評估常用的量測值是事後機率而非likelihood，例如GOP即是一種以音素事後機率為概念的信心度評估指標，本實驗也採用音素事後機率作為判斷發音品質的依據，得到特定音素事後機率參數後，針對各音素制定合適的門檻值(threshold)，以建立基本音素錯誤偵測器。3.1 節將介紹音素事後機率的抽取流程，3.2 節會定義音素錯誤偵測系統中的效能評估方式，3.3 節將介紹基本音素發音錯誤偵測系統門檻值的設定方式，3.4 節介紹基本系統的測試效能分析。

3.1 音素事後機率計算

假設每個音素的事前機率都一樣，則音素事後機率可表示如下：

$$P(q_i \mid y_t) = \frac{p(y_t \mid q_i)P(q_i)}{\sum_{j=1}^{38} p(y_t \mid q_j)P(q_j)} = \frac{p(y_t \mid q_i)}{\sum_{j=1}^{38} p(y_t \mid q_j)} \tag{3.1}$$

其中 $P(q_i \mid y_t)$ 項是音素 q_i 在音段 y_t (一個音素的長度)的事後機率，表示已知音段 y_t，而這音段是由音素 q_i 產生的機率；$p(y_t \mid q_i)$ 項是音素 q_i 在音段 y_t 的 likelihood，表示已知音素 q_i 的情況下，會產生 y_t 這個觀察向量的機率，所以從(3.1)式可以明顯得知，要得到某個音素在一個音段的事後機率，必須先得到所有音素在該音段的 likelihood 值才行。

因為沒有 TCC300 語料的正確切割位置，為了得到每個音段對所有音素的 likelihood 值，根據強制對齊所得到的音素邊界，我們對所有音段作 Top-N 辨識，

因為聲學模型包含單音素(Mono-phone)、雙連音素(Bi-phone)以及三連音素三種音素結構，所以前 N 名辨識裡表示同一個音素的三種音素結構皆有可能出現，然而雙連音素和三連音素都是單音素在考慮連音效果時所產生的單位，我們取三種音素結構中的最大 likelihood 值當作該音素的 likelihood 分數，至於沒出現在前 N 名辨識結果裡的音素，一律給予該音段辨識結果中最小的 likelihood 值扣掉一個常數當作底限值(ground value)，以這樣的方法取得訓練語料中每個音素在所有音段的 likelihood 值，再利用(3.1)式計算出對應的音素事後機率，這些事後機率就是音素錯誤偵測器的訓練特徵參數。

3.2 音素發音錯誤偵測系統之效能評估方式

3.2.1 錯誤拒絕和錯誤接受

一般來說音素發音錯誤偵測器所偵測出的錯誤可以分成兩類，其一是錯誤拒絕(False Rejection, FR)，表示發音正確卻被偵測出錯誤，另一種是錯誤接受或是錯誤警報(False Acceptance or False Alarm, FA)，表示發音錯誤卻被判斷成正確。根據學習者的心理，“當發音正確的時候被偵測出發音錯誤”所帶來的負面影響要大於“當發音錯誤卻被說發音正確” 【6】，所以一般偵錯系統往往將操作點(operation point)定在錯誤拒絕較少的地方，這麼做的同時不可避免的會使錯誤接受增加。

附帶一提的是，在本論文中訓練音素錯誤偵測器時，因為使用的訓練語料是 Native Speech，所以在訓練的過程中發音正確的語料就是指“目標音素”的語料，而發音錯誤的語料就是“非目標音素”的語料，舉例來說，現在要訓練音素 /a/ 的音素發音錯誤偵測器，那正確的語料就是文本標音為 /a/ 的音段，錯誤語料就是其他 37 個音素音段。

3.2.2 錯誤率

錯誤率是錯誤數量對所有測試資料量取平均後的量測值，根據錯誤的種類而有錯誤拒絕率(False Rejection Rate, FRR)和錯誤接受率或錯誤警報率(False Acceptance Rate or False Alarm Rate, FAR)。

$$FRR = \frac{FR}{N_1} = \frac{FR}{FR + CA} \tag{3.2}$$

$$FAR = \frac{FA}{N_0} = \frac{FA}{FA + CR} \tag{3.3}$$

其中 N_1 表示發音正確(目標音素)的數量；N_0 表示發音錯誤(非目標音素)的數量；FA 表示錯誤接受的音素數量；FR 表示錯誤拒絕的音素數量；CA 表示正確接受(Correct Acceptance)的音素數量；CR 表示正確拒絕(Correct Rejection)的音素數量。

下圖 3-1 是錯誤接受和錯誤拒絕在門檻值決定之後的示意圖，一般小於門檻值的音素事後機率會被拒絕、當作錯誤，而大於門檻值的音素事後機率會被接受、當作正確，但是當被拒絕的資料是來自正確的語料而被接受的語料是來自錯誤語料時，兩種錯誤就發生了。

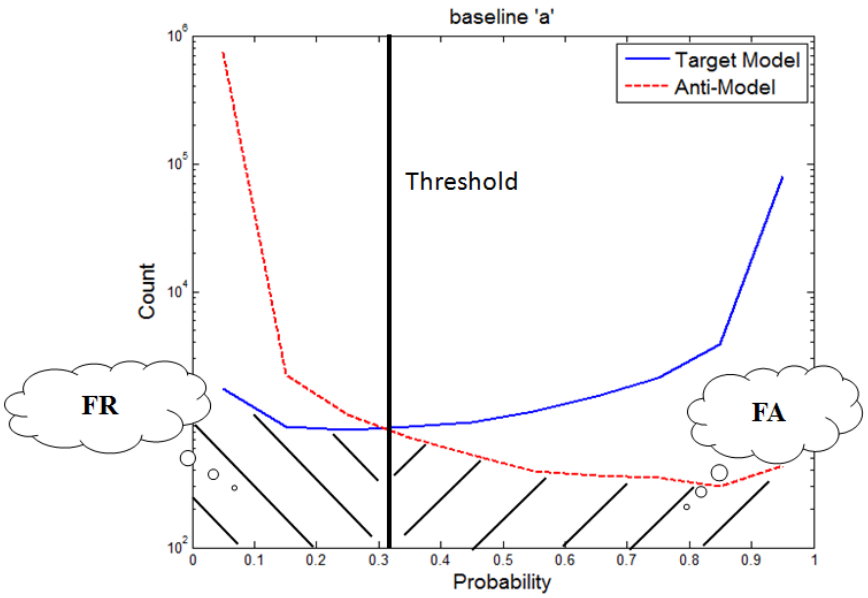


圖 3-1：FR 和 FA 示意圖

3.3 基本音素發音錯誤偵測系統

在得到所有訓練語料的音素事後機率以及定義好錯誤率之後，接下來要決定音素相關(phone dependent)的門檻值，也就是不同音素使用不同的門檻值作為發音品質的判斷，門檻值的值域和事後機率的範圍一樣介於 0 和 1 之間，每個門檻值可以決定一組(FAR，FRR)，一般將門檻值訂在等錯誤率(Equal Error Rate)的位置，但由於在 3.2.1 節提到錯誤拒絕帶來的負面影響較大，所以設計系統將門檻值訂在錯誤拒絕相對少於錯誤接受的地方，因此這裡引用決策代價函數(Decision Cost Function, DCF)來限制操作點錯誤的比例：

$$\begin{aligned} DCF &= C_{FR}P_{Target}P_{FR} + C_{FA}\left(1-P_{Target}\right)P_{FA} \\ &= C_{FR}\times\frac{N_1}{N}\times\frac{FR}{N_1} + C_{FA}\times\frac{N_0}{N}\times\frac{FA}{N_0} \\ &= C_{FR}\times\frac{FR}{N} + C_{FA}\times\frac{FA}{N} \\ &\propto C_{FR}\times FR + C_{FA}\times FA \end{aligned} \tag{3.4}$$

其中 C_{FR} 和 C_{FA} 分別表示錯誤拒絕和錯誤接受的懲罰係數，因為較不希望錯誤拒絕發生，所以設定 $C_{FR} > C_{FA}$，在本論文中將其設定為 $C_{FR}=2$ 和 $C_{FA}=1$；N_1 是發音正確的個數、N_0 是發音錯誤的個數、N 是所有的資料量；P_{Target} 是目標音素出現的機率，也就是該音素語料量佔全部語料量的比例；P_{FR} 和 P_{FA} 是錯誤拒絕率和錯誤接受率；FR 是錯誤拒絕的數量、FA 是錯誤接受的數量。

由(3.4)式推導到最後，因為不除以總資料量並不影響 DCF 的相對大小，所以能寫成最後一個正比式，本論文直接採用當作錯誤總量 DCF_N：

$$\begin{aligned} DCF_N &\equiv C_{FR}\times FR + C_{FA}\times FA \\ &= 2\times FR + 1\times FA \end{aligned} \tag{3.5}$$

圖 3-2 描述了音素 /a/ 的門檻值訓練情形，實線表示 DCF_N 值和門檻值的對應曲線，另外兩條虛線分別表示門檻值和 FA 以及 2 倍 FR 的對應關係，三角形

的位置是最小 DCF_N 值的位置，表示如果選定門檻值為 0.1906 的話，照(3.5)式算出來的 DCF_N 會是 9,399。

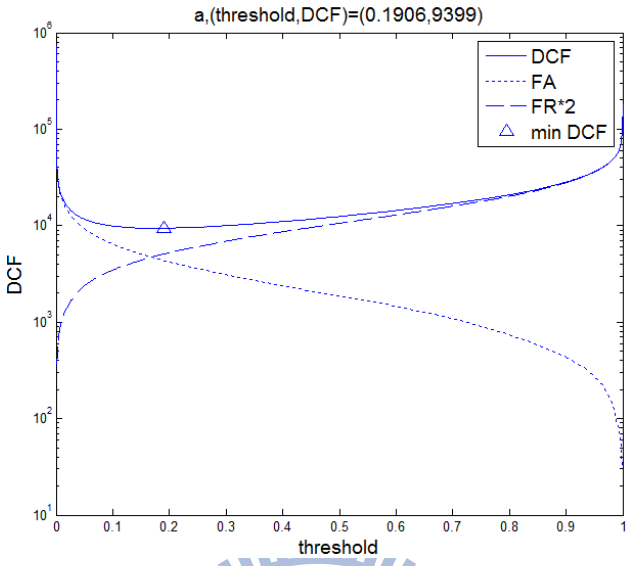


圖 3-2：音素 /a/ 的 threshold-DCF 曲線

3.4 基本音素發音錯誤偵測系統測試結果

為了討論方便，之後的論文裡將本章所探討的基本音素發音錯誤偵測系統簡稱作 baseline 系統，此系統使用 TCC300 訓練語料針對 38 個音素個別設定最小 DCF_N 之門檻值，criterion 如 3.3 節所描述，詳細的訓練語料音素資料量請參考附錄二，接下來介紹 baseline 系統的測試效能與分析。

3.4.1 內部測試

內部測試(inside test)使用的測試語料為用來訓練 baseline 系統的 TCC300 訓練語料，表 3-1 是測試結果，包含門檻值、相對的 DCF 值、FA 個數、FR 個數、FAR 錯誤率以及 FRR 錯誤率等資訊，而表 3-2 是每個錯誤偵測器的前三名錯誤表，其中包含錯誤拒絕和錯誤接受，表中的括號內數字表示該音素錯誤的數量。

由表中可以觀察到一些現象，首先就錯誤拒絕率 FRR 來看，超過 20%的音素有 /FNULL2/ 、 /z(ㄗ)/ 、 /c(ㄘ)/ 、 /s(ㄙ)/ 四個，而除了 /FNULL2/ 以外的三個音素恰巧是中文常見的混淆音對的其中之一，而 /FNULL2/ 剛好也是 /ㄗ/ 、 /ㄘ/ 、 /ㄙ/三個音的韻母，猜測照成這個現象的原因是 TCC300 語料庫語料並非由字正腔圓的發音員錄製而成，而是由捲舌音不夠明顯的台灣大學生錄製而成，使用這樣的語料訓練出來的聲學模型所抽取出來的聲學參數就會有一定程度的混淆效果，所以會導致這裡的單方面混淆現象；接著就 DCF 值來看，超過 9,000 的音素有 /a/ 、 /en/ 、 /e/ 、 /ng/ 、 /sh/ 、 /zh/ 、 /z/ 七個，這裡可以跟表 3-2 的前三名錯誤表做對照，這七個音來自四個混淆對，/a/ 和 /e/ 、 /ng/ 和 /en/ 、 /zh/ 和 /z/ 以及 /sh/ 和 /s/ ，明顯可以發現，成混淆對的音所測出來的 DCF 值都不小，這些音素在語言學上都是發音容易混淆的音。

表 3-1：Baseline 系統內部測試效能

Phone	DCF	FA	FR	FAR(%)	FRR(%)	threshold
FNULL1	5,280	3,106	1,087	0.38	6.29	0.264
FNULL2	3,839	1,621	1,109	0.19	**22.96**	0.572
a	**9,399**	4,335	2,532	0.58	2.68	0.191
b	3,944	1,728	1,108	0.21	7.90	0.391
ch	4,457	1,909	1,274	0.23	13.15	0.337
c	4,297	1,535	1,381	0.18	**28.30**	0.427
d	6,249	2,875	1,687	0.35	6.25	0.301
en	**22,687**	12,889	4,899	1.64	9.07	0.360
er	1,050	284	383	0.03	17.53	0.475
e	**15,325**	6,941	4,192	0.91	5.49	0.288
eh	6,589	2,675	1,957	0.33	7.96	0.502
f	2,970	1,208	881	0.15	9.93	0.400
g	1,885	615	635	0.07	4.50	0.360
h	1,960	740	610	0.09	4.79	0.380
j	4,703	2,507	1,098	0.31	5.17	0.245
k	987	289	349	0.03	6.16	0.517
l	4,562	1,982	1,290	0.24	8.80	0.473
m	1,031	385	323	0.05	3.72	0.408
ng	**26,334**	11,578	7,378	1.47	14.62	0.306
n	2,312	856	728	0.10	11.41	0.623
o	5,743	2,287	1,728	0.28	6.73	0.345
p	1,499	623	438	0.07	10.75	0.438
q	4,220	1,754	1,233	0.21	11.68	0.393
r	3,469	1,285	1,092	0.15	13.89	0.367
s	6,181	2,243	1,969	0.27	**34.34**	0.518
sh	**10,412**	5,592	2,410	0.69	10.60	0.213
t	3,189	1,419	885	0.17	7.78	0.301
wu1	2,904	1,502	701	0.19	1.49	0.315
wu2	2,224	932	646	0.11	3.02	0.633
wu3	6,612	3,650	1,481	0.45	4.99	0.352
x	5,460	2,826	1,317	0.34	8.08	0.336
yi1	5,798	3,092	1,353	0.40	2.23	0.176
yi2	4,732	1,788	1,472	0.22	5.66	0.594
yi3	7,237	3,409	1,914	0.42	6.56	0.351
yu1	3,271	1,189	1,041	0.14	10.71	0.453

yu2	1,784	648	568	0.08	7.06	0.688
zh	**10,585**	5,855	2,365	0.72	12.55	0.204
z	**9,550**	3,690	2,930	0.45	**27.51**	0.409

表 3-2：Baseline 系統前 3 名錯誤表

Phone	No.1	No.2	No.3
FNULL1	FNULL2(1,644)	FNULL1(1,087)	e(312)
FNULL2	FNULL1(1,552)	FNULL2(1,109)	e(21)
a	a(2,532)	e(2,126)	o(626)
b	b(1,108)	f(650)	d(614)
ch	ch(1,274)	c(1,075)	zh(246)
c	c(1,381)	ch(1,003)	z(130)
d	d(1,687)	b(701)	zh(383)
en	ng(11,442)	en(4,899)	yi3(534)
er	er(383)	e(166)	a(37)
e	e(4,192)	a(2,807)	eh(987)
eh	eh(1,957)	a(1,270)	yi3(626)
f	f(881)	b(547)	d(110)
g	g(635)	d(232)	k(138)
h	h(610)	g(126)	t(123)
j	q(1,152)	j(1,098)	zh(703)
k	k(349)	g(135)	h(62)
l	l(1,290)	r(589)	d(509)
m	m(323)	n(110)	l(65)
ng	en(9,366)	ng(7,378)	wu3(835)
n	n(728)	l(249)	en(154)
o	o(1,728)	e(586)	a(560)
p	p(438)	t(319)	b(116)
q	q(1,233)	j(1,133)	x(393)
r	r(1,092)	l(377)	d(311)
s	s(1,969)	sh(1,901)	z(98)
sh	sh(2,410)	s(2,273)	x(1,892)
t	t(885)	p(284)	d(229)
wu1	wu1(701)	o(428)	a(219)
wu2	wu2(646)	o(252)	wu3(155)
wu3	ng(1,913)	wu3(1,481)	o(633)
x	sh(1,679)	x(1,317)	q(636)

yi1	yi1(1,353)	yu1(1,208)	yi2(420)
yi2	yi2(1,472)	yi1(743)	e(362)
yi3	yi3(1,914)	en(1,197)	e(662)
yu1	yu1(1,041)	yi1(917)	wu1(80)
yu2	yu2(568)	yi2(238)	yu1(100)
zh	z(3,654)	zh(2,365)	j(854)
z	zh(3,064)	z(2,930)	j(142)

3.4.2 外部測試

外部測試(outside test)使用的測試語料是 TCC300 語料庫的測試語料部分，詳細內容請參考附錄二，測試的方法是將訓練出來的門檻值套用到由測試語料所得到的音素事後機率參數，藉以計算 DCF、FR、FA、FRR 以及 FAR 等五個效能評估指數。

這裡的測試結果跟內部測試的部分有相當高的正相關性，從 3.4.2 節討論的 FRR 和 DCF 值來觀察，在這裡可以看到一致的現象。

表 3-3：Baseline 系統外部測試效能

Phone	DCF	FA	FR	FAR(%)	FRR(%)
FNULL1	879	487	196	0.57	11.07
FNULL2	725	279	223	0.32	**41.84**
a	**1,708**	848	430	1.10	4.37
b	718	238	240	0.28	16.21
ch	754	348	203	0.40	20.65
c	737	227	255	0.26	**49.61**
d	984	424	280	0.50	10.30
en	**3,248**	1,706	771	2.09	13.60
er	254	96	79	0.11	30.74
e	**2,378**	1,026	676	1.29	8.65
eh	1,067	383	342	0.45	13.31
f	365	177	94	0.20	11.75
g	370	98	136	0.11	9.23
h	360	122	119	0.14	8.32

j	698	412	143	0.49	6.22
k	176	54	61	0.06	11.40
l	970	440	265	0.51	16.33
m	195	81	57	0.09	6.58
ng	**3,762**	1,716	1023	2.09	19.88
n	572	156	208	0.18	31.14
o	973	385	294	0.46	10.67
p	323	83	120	0.10	27.15
q	762	278	242	0.32	20.72
r	681	269	206	0.31	26.14
s	923	287	318	0.33	**52.65**
sh	**1,427**	743	342	0.87	15.33
t	591	263	164	0.31	13.86
wu1	509	243	133	0.30	2.75
wu2	375	145	115	0.17	5.38
wu3	1,028	548	240	0.65	7.78
x	754	364	195	0.43	11.77
yi1	973	567	203	0.70	3.18
yi2	821	225	298	0.27	10.74
yi3	1,119	477	321	0.57	10.29
yu1	619	209	205	0.24	18.96
yu2	300	136	82	0.16	9.57
zh	**1,534**	892	321	1.05	16.39
z	**1,394**	484	455	0.56	**39.67**

第四章 基於多層感知器之音素發音錯誤偵測器

在 baseline 系統中只使用到目標音素的“一維”音素事後機率資訊，相信引入更多維的資訊將會使錯誤偵測的效能有所提升，而類神經網路中的多層感知器是好用的多輸入多輸出(MIMO)經驗建模(empirical modeling)工具，其優點是擁有適應性學習(adaptive learning)的能力，另外，擁有足夠隱藏層節點(hidden nodes)的雙層倒傳遞網路(Back-Propagation Network)，已經被證實是通用的逼近器(universal approximators)【7】，這也是本實驗的 MLP 網路僅使用兩層的原因。4.1 節簡單介紹 MLP 網路的架構以及訓練準則【8】，4.2 節介紹 MLP-based 的音素發音錯誤偵測器之測試效能分析，4.3 節介紹改良的 MLP-based 偵測器之訓練理念，以及其效能分析。

4.1 多層感知器網路簡介

4.1.1 網路架構

神經元(Neuron)是多層感知器網路中最基本的訊息處理單元，一個完整的神經元模型包含三個組成元素：突觸(synapse)、加法器(adder)以及活化函數(activation function)，其結構如圖 4-1 所示，而多層感知器的多層結構使網路能處理較複雜的非線性可分割的分類問題，本論文所使用的 MLP 網路架構如圖 4-2 所示。

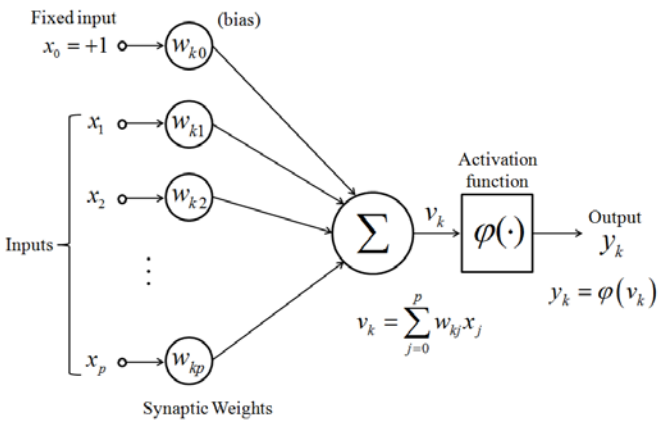


圖 4-1：神經元模型

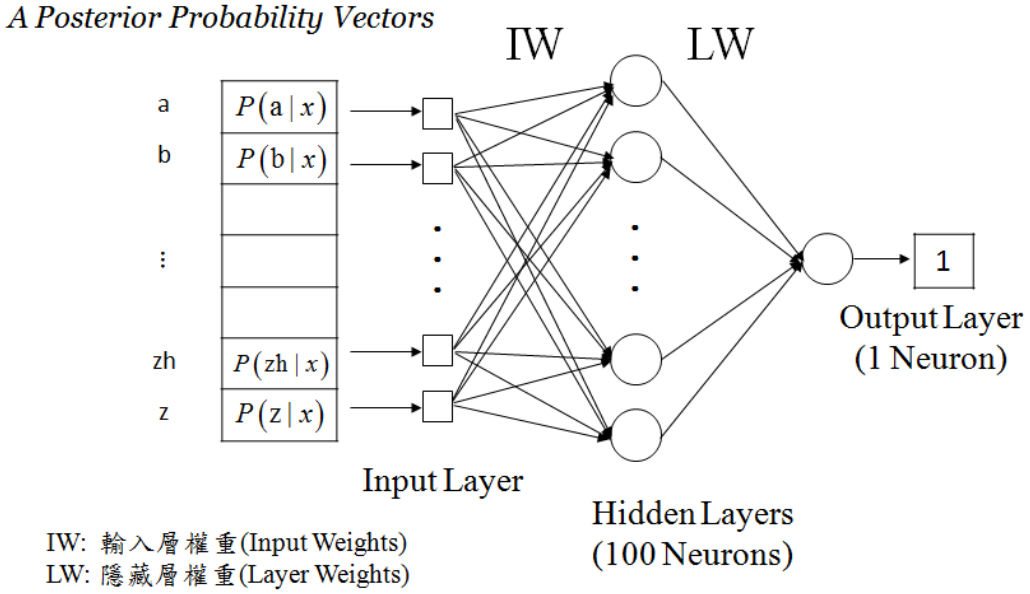


圖 4-2：本論文的 MLP 架構示意圖

如上圖所示，本實驗採用雙層 MLP 網路(一層隱藏層加上一層輸出層)，隱藏 0 層的神經元數量為 100，輸出層只使用 1 個神經元，輸入為 38 維音素事後機率向量，每個向量代表訓練語料中某個音段所計算出的 38 個音素事後機率，而機率向量根據來源分成“目標模型”和“非目標模型”兩群，輸出為 1 或 0，1 對應目標模型的機率向量，0 則對應非目標模型的機率向量。

隱藏層以及輸出層都有神經元，其中使用的活化函數都是雙曲線正切函數(hyperbolic tangent function)，是值域介於-1 和 1 之間的函數，如圖 4-3 所示。

$$\varphi(x) = \tanh(x) = \frac{\sinh(x)}{\cosh(x)} \tag{4.1}$$

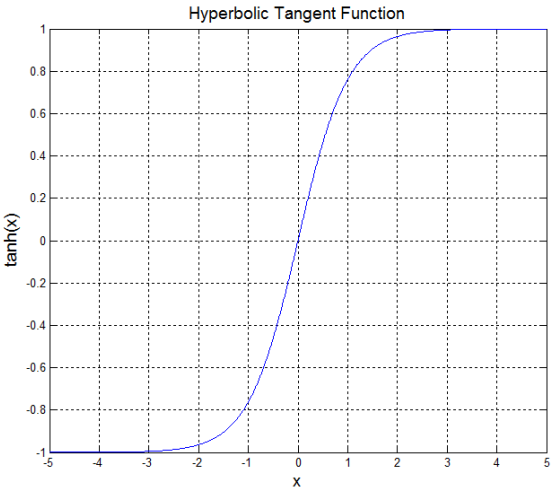


圖 4-3：MLP 網路使用的活化函數

4.1.2 多層感知器網路之訓練

MLP 的訓練是為了讓目標輸出與實際網路輸出之間的誤差變小，一般表示成下式：

$$J = \frac{1}{2} E\left[e^2\right] \tag{4.2}$$

其中 J 為方均誤差(Square Mean Error)，e 是目標輸出與實際網路輸出的差值，$E[\cdot]$ 表示期望值，整個多層感知器的訓練 criterion 就是為了讓 J 能達到最小。

通常 MLP 網路都會使用倒傳遞演算法(Back-Propagation Algorithm)來調整權重值(weights)和偏權值(bias)，調整的方式引用最陡坡降法(Steepest Descend)的概念，將權重往“讓平均誤差變小最快的方向”作修正，而之所以會稱作倒傳遞演算法是因為在調整權重的過程分成兩個階段，一個是前向傳遞(forward pass)，在這個階段權重保持固定，將輸入向量從輸入端輸入，一層一層計算中間值直到得到最後的輸出層結果，然後計算出誤差，第二階段是反向傳遞(backward pass)，將前一步驟得到的誤差反向回傳計算權重改變量，一層一層更新權重值，直到整個

網路的權重值全部更新完畢。

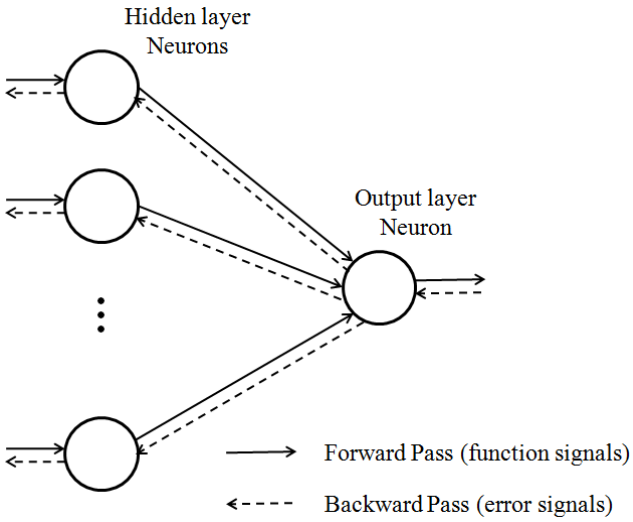


圖 4-4：二階段倒傳遞演算法

關於使用最陡坡降法作為更新權重的部分，如果學習速率太小的話，權重的改變在誤差曲面(Error Surface)上的反映會是平緩的往最佳值改變，但需要長時間才能收斂；反之如果學習速率太大，學習速率會因而變快，但在誤差曲面上會有震盪無法穩定的現象，改善這兩種情況的常見方法就是引入動量項(Momentum term)， 除了改善前面提到的兩個問題之外，還能一定程度的避免權重收斂到誤差曲面的區域最小值(local minimum)。

4.2 使用多層感知器之音素發音錯誤偵測器

MLP-based 的音素發音錯誤偵測器和 baseline 系統使用一樣的訓練語料，資料量如附錄二所示，採用 4.1.1 節的網路架構，分別完成 38 個 MLP 網路的訓練後即可進行內部測試與外部測試，並分析其偵錯的效能。

4.2.1 內部測試與比較

為了討論方便，在接下來的論文將這節訓練出來的發音錯誤偵測器稱作“MLP-1”，並且將此偵測器就內部測試的部分和 baseline 系統作比較，表 4-1 列出了每個音素發音錯誤偵測器的效能，圖 4-5 為 MLP-1 和 baseline 的 DCF 差距圖(baseline 的 DCF 減去 MLP-1 的 DCF)，表 4-2 為兩系統的 DCF 改變量相對於 baseline 系統的百分比，表 4-3 為 MLP-1 系統整體變好的量與百分比分析。

結果顯示， baseline 系統擁有的偵測器特性，這裡的系統也有相當的一致性，全部 38 個音素中，有 37 個音素 MLP-1 偵測器比 Baseline 好，僅有 1 個音素(FNULL2, index = 2)較差，但也只略差 0.2%而已，明顯各項指標在 MLP-based 的偵測器上有全面性的改進，詳細的音素索引對照表請參考附錄三。

表 4-1：MLP-1 內部測試系統效能

Phone	DCF	FA	FR	FAR(%)	FRR(%)	threshold
FNULL1	4,536	2,842	847	0.35	4.90	0.339
FNULL2	3,846	1,572	1137	0.19	23.54	0.328
a	8,603	4,037	2283	0.54	2.42	0.474
b	3,894	1,736	1079	0.21	7.69	0.290
ch	4,328	2,210	1059	0.27	10.93	0.278
c	4,273	1,699	1287	0.20	26.38	0.306
d	5,866	2,690	1588	0.33	5.88	0.356
en	22,212	13,328	4442	1.70	8.23	0.293
er	1,022	228	397	0.03	18.17	0.357

e	13,941	6,197	3872	0.81	5.07	0.340
eh	5,728	2,056	1836	0.25	7.47	0.387
f	2,893	1,313	790	0.16	8.91	0.211
g	1,797	733	532	0.09	3.77	0.169
h	1,910	796	557	0.10	4.37	0.221
j	4,627	2,329	1149	0.29	5.41	0.359
k	972	346	313	0.04	5.52	0.270
l	4,374	1,766	1304	0.21	8.90	0.304
m	1,004	314	345	0.04	3.98	0.322
ng	25,793	12,695	6549	1.61	12.98	0.288
n	2,240	1,044	598	0.13	9.38	0.227
o	5,597	2,415	1591	0.30	6.20	0.377
p	1,479	647	416	0.08	10.21	0.173
q	4,074	1,526	1274	0.18	12.07	0.448
r	3,346	1,484	931	0.18	11.84	0.260
s	6,123	2,155	1984	0.26	34.60	0.381
sh	9,656	5,886	1885	0.72	8.29	0.319
t	3,112	1,334	889	0.16	7.81	0.292
wu1	2,695	1,255	720	0.16	1.53	0.306
wu2	2,184	830	677	0.10	3.17	0.449
wu3	6,344	3,136	1604	0.39	5.40	0.319
x	5,300	2,622	1339	0.32	8.21	0.306
yi1	5,388	2,872	1258	0.37	2.07	0.371
yi2	4,039	1,571	1234	0.19	4.75	0.343
yi3	6,813	3,115	1849	0.39	6.34	0.332
yu1	3,104	1,174	965	0.14	9.93	0.182
yu2	1,653	689	482	0.08	5.99	0.235
zh	10,320	5,702	2309	0.70	12.25	0.327
z	9,400	3,776	2812	0.46	26.40	0.264

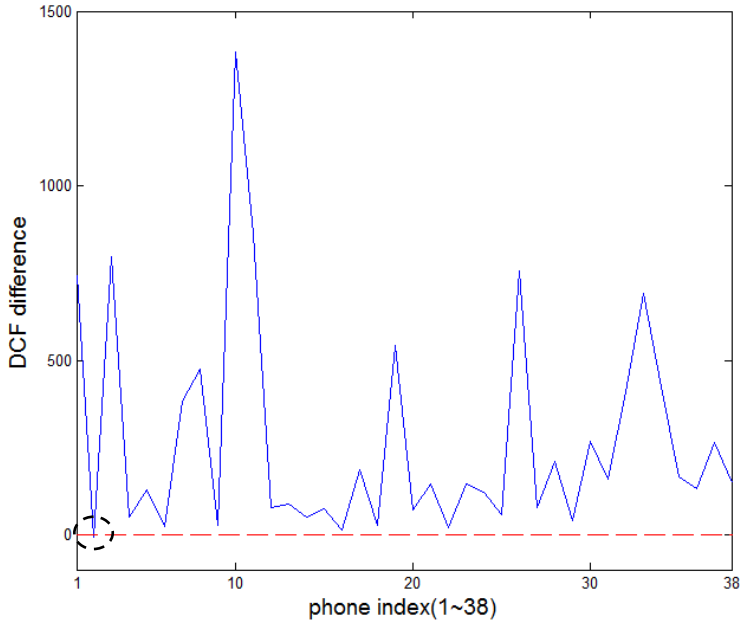


圖 4-5：Baseline 系統和 MLP-1 系統的內部測試 DCF 值差距圖

表 4-2：MLP-1 系統和 Baseline 系統的內部測試 DCF 值改變百分比

FNULL1	14.1%	n	3.1%
FNULL2	**-0.2%**	o	2.5%
a	8.5%	p	1.3%
b	1.3%	q	3.5%
ch	2.9%	r	3.5%
c	0.6%	s	0.9%
d	6.1%	sh	7.3%
en	2.1%	t	2.4%
er	2.7%	wu1	7.2%
e	9.0%	wu2	1.8%
eh	13.1%	wu3	4.1%
f	2.6%	x	2.9%
g	4.7%	yi1	7.1%
h	2.6%	yi2	14.6%
j	1.6%	yi3	5.9%
k	1.5%	yu1	5.1%
l	4.1%	yu2	7.3%
m	2.6%	zh	2.5%
ng	2.1%	z	1.6%

表 4-3：MLP-1 系統內部測試整體效能改進分析

	整體變好的量	整體變好的百分比
DCF	10,195	4.54%
FA	2,719	2.62%
FR	3,738	6.18%

4.2.2 外部測試與比較

外部測試使用 TCC300 語料庫的測試語料部分，表 4-4 列出所有音素的外部測試效能，圖 4-6 是 Baseline 系統和 MLP-1 系統的 DCF 差距圖，表 4-5 為 DCF 改變量百分比，表 4-6 為整體效能改進分析表。

表 4-4：MLP-1 外部測試系統效能

Phone	DCF	FA	FR	FAR(%)	FRR(%)
FNULL1	728	426	151	0.50	8.53
FNULL2	725	269	228	0.31	42.78
a	1,625	823	401	1.06	4.08
b	701	235	233	0.27	15.73
ch	704	380	162	0.44	16.48
c	736	250	243	0.29	47.28
d	929	391	269	0.46	9.89
en	3,200	1,840	680	2.26	12.00
er	238	78	80	0.09	31.13
e	2,298	922	688	1.16	8.81
eh	996	320	338	0.38	13.15
f	372	204	84	0.24	10.50
g	357	119	119	0.14	8.08
h	373	141	116	0.16	8.11
j	680	380	150	0.45	6.52
k	180	72	54	0.08	10.09
l	946	402	272	0.47	16.76
m	189	71	59	0.08	6.81
ng	3,738	1,858	940	2.26	18.27
n	548	192	178	0.22	26.65

o	944	422	261	0.50	9.47
p	327	85	121	0.10	27.38
q	763	239	262	0.28	22.43
r	668	314	177	0.36	22.46
s	915	275	320	0.32	52.98
sh	1,338	768	285	0.90	12.77
t	585	249	168	0.29	14.20
wu1	513	215	149	0.26	3.08
wu2	365	135	115	0.16	5.38
wu3	1,004	474	265	0.56	8.59
x	746	348	199	0.41	12.01
yi1	928	534	197	0.66	3.08
yi2	695	173	261	0.20	9.41
yi3	1,065	477	294	0.57	9.42
yu1	582	202	190	0.23	17.58
yu2	334	146	94	0.17	10.97
zh	1,470	858	306	1.01	15.63
z	1,397	499	449	0.58	39.15

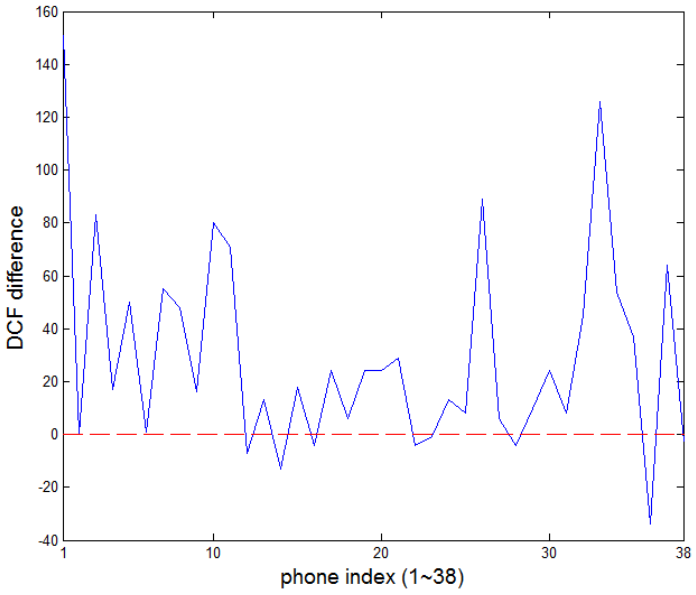


圖 4-6：Baseline 系統和 MLP-1 系統的外部測試 DCF 值差距圖

表 4-5：MLP-1 系統和 Baseline 系統的外部測試 DCF 值改變百分比

FNULL1	17.2%	n	4.2%
FNULL2	0.0%	o	3.0%
a	4.9%	p	**-1.2%**
b	2.4%	q	**-0.1%**
ch	6.6%	r	1.9%
c	0.1%	s	0.9%
d	5.6%	sh	6.2%
en	1.5%	t	1.0%
er	6.3%	wu1	**-0.8%**
e	3.4%	wu2	2.7%
eh	6.7%	wu3	2.3%
f	**-1.9%**	x	1.1%
g	3.5%	yi1	4.6%
h	**-3.6%**	yi2	15.3%
j	2.6%	yi3	4.8%
k	**-2.3%**	yu1	6.0%
l	2.5%	yu2	**-11.3%**
m	3.1%	zh	4.2%
ng	0.6%	z	**-0.2%**

表 4-6：MLP-1 系統外部測試整體效能改進分析

	Overall 變好的量	Overall 變好的百分比
DCF	1,106	3.07%
FA	270	1.70%
FR	418	4.16%

4.3 MLP 音素發音錯誤偵測器之改良

在前面的訓練中，已經取得效能不錯的 MLP-1 偵測器，但前面的訓練是所有的語料做訓練，事實上，對訓練語料資料做可以得到更好的效能，其中考慮的幾個概念如下：1) 少選取不可靠的訓練語料，例如對目標模型來說，音素事後機率值很低的資料就是不可靠的資料；2) 多取容易混淆的語料，一般認為 MLP 訓練完後，如果測試的資料輸出值落在 0.5 附近就算是容易混淆的資料，如此做的原因是初步訓練出來的 MLP 已經對很可靠的資料學習得很好，再重新訓練的話對這些資料幫助可能不大，但容易混淆的資料很容易受到權重值調整的影響，所以這種資料要多取，讓網路對這些資料調適得好一點；3)原本 MLP 的訓練 criterion 是最小均方誤差(Minimum Mean Square Error, MMSE)，而我們所希望的音素發音錯誤偵測器的 criterion 是最小 DCF(MDCF)，或許能透過操作資料取樣使得達到 MMSE 的同時也達到 MDCF 的 criterion；4)MLP 網路在學習的過程中有偏向資料量多的資料的特性，所以對每個音素錯誤偵測器來說，音素間的數量關係可由 baseline 系統的錯誤情況求得，多錯的資料就多取一點，讓 MLP 網路對這些資料好好調適。接下來將細節介紹考慮了這先因素後的取樣方法，以及改良過後的音素發音錯誤偵測器之效能分析。

4.3.1 資料取樣

資料的取樣都是以亂數產生器產生亂數值，再將該值拿去比對希望達到的取樣機率分佈，以這樣的概念達到隨機卻又能保持資料機率分佈的取樣方式。

作法上如同前面所述的概念，結合 1)和 2)的概念，我們分別將目標音素和非目標音素的取樣參考機率密度函數(Probability Density Function, PDF)訂作圖 4-7 以及圖 4-8 的形狀，從 PDF 圖可以看出最不可靠的部分完全沒取，以及容易混淆的部分取樣機率很高，如此決定好亂數取一個音素資料時的取樣機率分佈。

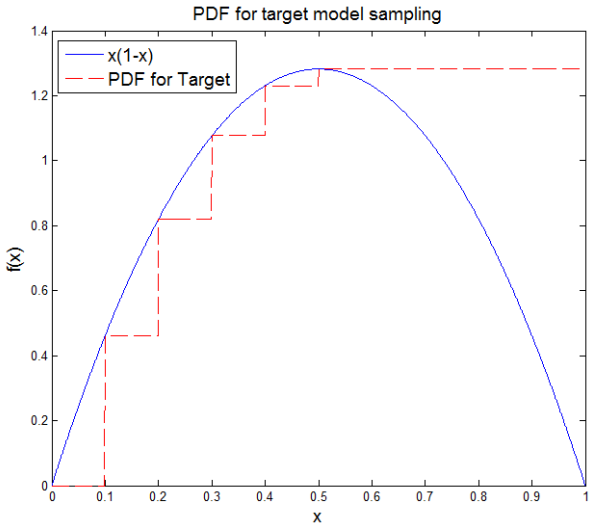


圖 4-7：目標音素的取樣參考 PDF

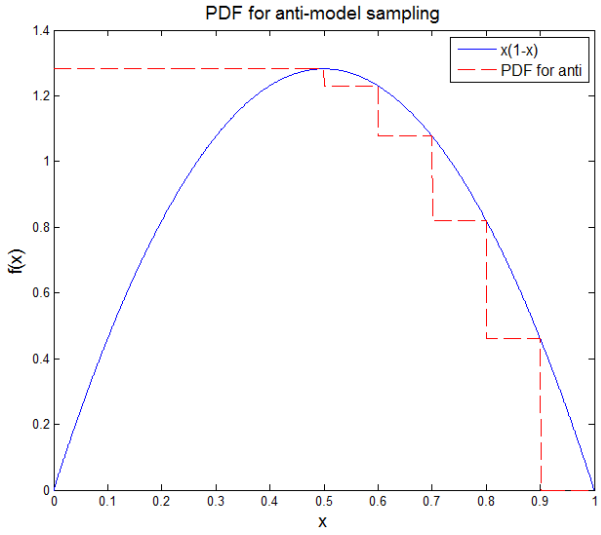


圖 4-8：非目標音素的取樣參考 PDF

接著要考慮的是，音素間的取樣數量關係，這部分結合 3)、4)的概念，所以根據baseline的方法，將門檻值定在等錯誤量(Equal Error Count)的地方，把這時各音素的錯誤個數當作取樣參考值，對於沒有錯或是錯誤量過少的音素我們給其一個底限值，避免某些音素取樣量不足的情況發生；另外，考慮前面所提及的DCF函數，這裡依照錯誤拒絕和錯誤接受的懲罰係數，將目標音素也就是對應於“對”的資料，在決定完各音素的取樣量之後再多取 1 倍的資料量，期許做這樣的使系統能達到最小DCF的訓練criterion。例如圖 4-9 即是以音素 /ch/ 為範例，參考 baseline 系統錯誤狀況所求出來的 PDF 以及累積分部函數(Cumulative

Distribution Function, CDF)如圖所示，由其中可看出一個音素的錯誤情況，包含錯誤拒絕和錯誤接受，虛線的部份表示音素ch(ㄔ)的取樣參考PDF，可以觀察到較高的幾個PDF值表示這個音素的baseline偵測器較容易偵測失敗的音素，從這個例子來看，錯誤接受較多的部分有index為 6 的 /c(ㄘ)/ 、37 的 /zh(ㄓ)/ 、26 的 /sh(ㄕ)/ 、23 的 /q(ㄑ)/ …等，對於這些音素我們給予較高的被取樣機率，期許MLP偵測器能對這些音素的偵測有所改善。

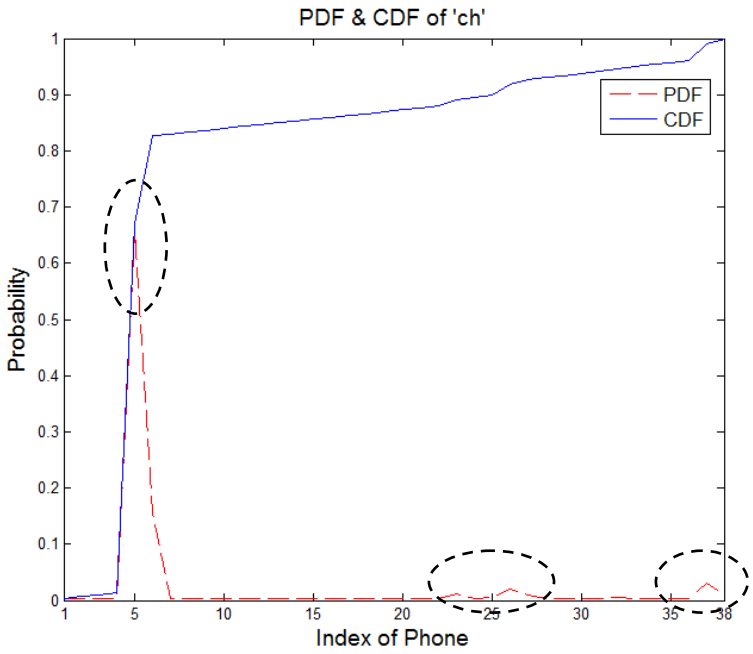


圖 4-9：音素 /ch/ 的取樣參考 PDF 和 CDF

在決定了音素的挑選機率和資料區間的挑選機率後就可用亂數產生的方式完成隨機取樣的動作，對每個音素偵測器都隨機挑選 50,000 筆資料當作重新訓練 MLP-1 網路的訓練語料，在隨機挑選的過程中，可能會碰到某個音素的某個區間語料被取完或者是某個音素全部的資料都被取完的情況，解決的方法是，對於目標音素來說，資料可重複取樣，如此一來就能讓取樣的資料維持 PDF 分佈，並且不會有前面所述的情況發生；而對於非目標音素來說，根據實驗測試，讓資料不可重複取樣的結果比讓資料可重複取樣的結果好，所以採用不可重複取樣的方式，一旦碰到某個區間的資料取完，就重新產生亂數，去取還有剩餘資料的部

分，而碰到某個非目標音素資料被取完，一樣重新產生亂數，去取還有剩餘資料的音素語料，圖 4-10 即是以音素 /ch/ 為例，對取樣完且由 MLP 網路輸出的目標模型訓練語料和非目標模型訓練語料畫 10 點的 Histogram 圖，可以發現目標模型的部分因為可重複取樣而維持和 PDF 差不多的形狀；非目標模型的部分則因為不能重複取樣，受限於原始資料分佈。

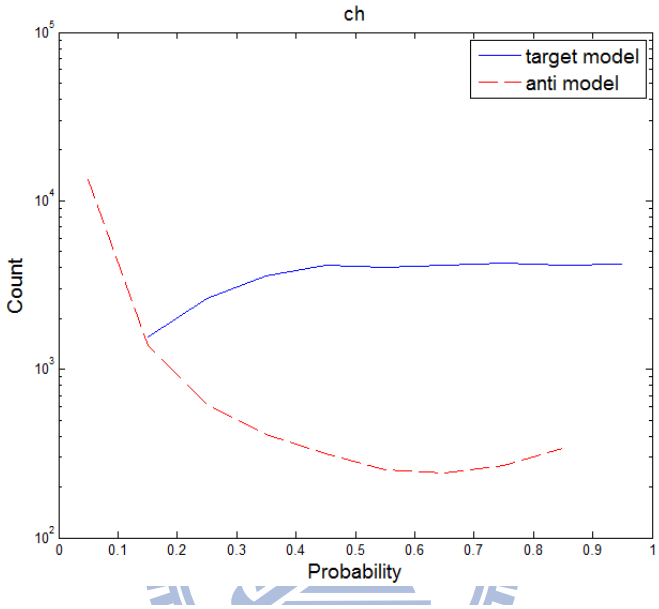


圖 4-10：音素 /ch/ 的取樣資料 Histogram 圖

4.3.2 內部測試與比較

經過前面的取樣後，重新對 MLP-1 做重新訓練，得到改良後的音素錯誤偵測器系統，之後將這個系統稱作“MLP-2”以方便討論。表 4-7 列出內部測試的系統效能，其中顯示跟 baseline 和 MLP-1 系統一致的 FRR 和 DCF 特性，另外可以注意到 MLP-2 訓練完的門檻值普遍大於 0.9，這是因為重新訓練網路的資料量，目標模型所佔的比重跟 MLP-1 訓練用的資料量比起來要大許多，使得目標音素效能變好而非目標音素效能變差，門檻值因而偏向讓錯誤接受減少的方向，儘管如此，系統在這樣門檻值的情況下仍然有不錯的效能，表 4-8 和圖 4-11 顯示 MLP-2 和 MLP-1 在 DCF 上有優有劣，而表 4-9 顯示雖然整體在 DCF 和 FA 量上

MLP-1 略勝一籌，但在 FR 上 MLP-2 有較好的表現。

表 4-7：MLP-2 內部測試系統效能

Phone	DCF	FA	FR	FAR(%)	FRR(%)	threshold
FNULL1	4,501	2,763	869	0.34	5.03	0.973
FNULL2	3,842	1,610	1,116	0.19	23.11	0.976
a	8,700	4,094	2,303	0.55	2.44	0.979
b	3,878	1,720	1,079	0.21	7.69	0.969
ch	4,332	2,066	1,133	0.25	11.69	0.971
c	4,267	1,551	1,358	0.19	27.83	0.959
d	5,885	2,903	1,491	0.36	5.53	0.966
en	22,322	13,582	4,370	1.73	8.09	0.578
er	1,033	255	389	0.03	17.80	0.996
e	13,910	6,258	3,826	0.82	5.01	0.783
eh	5,728	2,234	1,747	0.27	7.10	0.963
f	2,891	1,181	855	0.14	9.64	0.958
g	1,776	724	526	0.09	3.73	0.967
h	1,925	877	524	0.11	4.11	0.929
j	4,622	2,200	1,211	0.27	5.70	0.982
k	972	336	318	0.04	5.61	0.995
l	4,366	1,798	1,284	0.22	8.76	0.987
m	1,017	323	347	0.04	4.00	0.998
ng	25,924	12,890	6,517	1.64	12.92	0.704
n	2,243	1,045	599	0.13	9.39	0.975
o	5,589	2,151	1,719	0.26	6.70	0.978
p	1,476	632	422	0.08	10.36	0.963
q	4,063	1,693	1,185	0.20	11.23	0.976
r	3,340	1,510	915	0.18	11.64	0.964
s	6,106	2,206	1,950	0.27	34.01	0.899
sh	9,657	5,861	1,898	0.72	8.35	0.853
t	3,095	1,349	873	0.16	7.67	0.971
wu1	2,683	1,123	780	0.14	1.66	0.995
wu2	2,194	952	621	0.12	2.90	0.989
wu3	6,377	3,311	1,533	0.41	5.16	0.976
x	5,309	2,683	1,313	0.33	8.05	0.961
yi1	5,452	2,910	1,271	0.37	2.09	0.977
yi2	4,098	1,646	1,226	0.20	4.72	0.977
yi3	6,824	3,160	1,832	0.39	6.28	0.969

yu1	3,108	1,164	972	0.14	10.00	0.956
yu2	1,652	658	497	0.08	6.18	0.977
zh	10,339	5,197	2,571	0.63	13.64	0.855
z	9,398	3,430	2,984	0.41	28.02	0.933

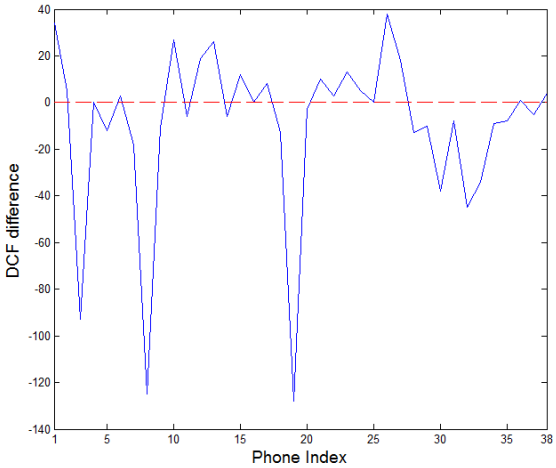


圖 4-11：MLP-1 系統和 MLP-2 系統的內部測試 DCF 值差距圖

表 4-8：MLP-1 系統和 MLP-2 系統的內部測試 DCF 值改變百分比

FNULL1	0.7%	n	-0.1%
FNULL2	0.1%	o	0.2%
a	-1.1%	p	0.2%
b	0.0%	q	0.3%
ch	-0.3%	r	0.1%
c	0.1%	s	0.0%
d	-0.3%	sh	0.4%
en	-0.6%	t	0.6%
er	-1.1%	wu1	-0.5%
e	0.2%	wu2	-0.5%
eh	-0.1%	wu3	-0.6%
f	0.7%	x	-0.2%
g	1.4%	yi1	-0.8%
h	-0.3%	yi2	-0.8%
j	0.7%	yi3	-0.1%
k	0.1%	yu1	-0.3%
l	-1.1%	yu2	0.1%
m	0.0%	zh	0.0%
ng	-0.3%	z	0.0%

表 4-9：MLP-2 相對 MLP-1 的內部測試效能改變資訊

	Overall 改變量	Overall 改變百分比
DCF	-359	-0.17%
FA	-931	-0.92%
FR	286	0.50%

4.3.3 外部測試與比較

外部測試的部分結果如表 4-10 和 4-11 所示，同樣與內部測試有相當高的一致性。

表 4-10：MLP-2 外部測試系統效能

Phone	DCF	FA	FR	FAR(%)	FRR(%)
FNULL1	716	420	148	0.49	8.36
FNULL2	726	278	224	0.32	42.03
a	1,645	839	403	1.08	4.10
b	700	236	232	0.28	15.67
ch	723	361	181	0.42	18.41
c	731	231	250	0.27	48.64
d	947	431	258	0.51	9.49
en	3,227	1,863	682	2.29	12.03
er	244	84	80	0.10	31.13
e	2,306	940	683	1.18	8.74
eh	990	342	324	0.40	12.61
f	359	181	89	0.21	11.13
g	346	110	118	0.13	8.01
h	368	154	107	0.18	7.48
j	679	355	162	0.42	7.04
k	178	70	54	0.08	10.09
l	956	416	270	0.49	16.64
m	190	70	60	0.08	6.93
ng	3,782	1,864	959	2.27	18.64
n	552	196	178	0.23	26.65
o	961	381	290	0.45	10.52

p	320	84	118	0.10	26.70
q	770	274	248	0.32	21.23
r	667	319	174	0.37	22.08
s	917	289	314	0.33	51.99
sh	1,330	760	285	0.89	12.77
t	585	251	167	0.29	14.12
wu1	541	205	168	0.25	3.47
wu2	356	148	104	0.17	4.87
wu3	989	489	250	0.58	8.11
x	753	355	199	0.42	12.01
yi1	930	536	197	0.66	3.08
yi2	698	186	256	0.22	9.23
yi3	1,057	475	291	0.57	9.32
yu1	585	201	192	0.23	17.76
yu2	328	140	94	0.16	10.97
zh	1,497	799	349	0.94	17.82
z	1,404	460	472	0.53	41.15

表 4-11：MLP-2 相對 MLP-1 的外部測試效能改變資訊

	Overall 改變量	Overall 改變百分比
DCF	-133	-0.38%
FA	-147	-0.94%
FR	7	0.07%

第五章 主觀測試實驗

前面章節介紹的音素錯誤偵測器使用的測試語料為 native speech，並且屬於客觀測試(objective test)，而本章主旨為使用音素錯誤偵測器分析外國人錄製的語料，也就是 non-native speech 語料，同時以 8 位錯誤標記員標記的錯誤為參考，進行主觀測試(subjective test)分析。5.1 節介紹使用的外國人中文發音語料，5.2 節介紹抽取音素事後機率參數之前要先進行語者調適，5.3 節為標記員的錯誤標記說明，以及一致性上的分析，5.4 節為主觀測試實驗結果與分析。

5.1 外國人中文語料介紹

這是由 9 個在交大的外國交換學生錄製而成的中文語料，屬於麥克風朗讀語音，共計 6 男 3 女，分別來自 5 個不同國家，其中以來自烏克蘭的人最多(2 男 3 女)，接觸中文的時間從 1 年到 3 年不等，錄製的語料內容每個人皆相同，為 20 個一組的中文二字詞共 5 組，總計 100 個二字詞，二字詞在設計時考慮“每組”涵蓋所有音調的組合以及所有音素，但不考慮音素量的平衡性，詳細的音素資料量請參考附錄四，表 5-1 為其中一組範例：

表 5-1：外國人中文二字詞語料範例

		第二個字				
		一聲	二聲	三聲	四聲	清聲
第	一聲	誇張	猜拳	縮小	喝醉	他的
一	二聲	昨天	籃球	傳染	螃蟹	橘子
個	三聲	耳機	倒楣	水餃	冷氣	你們
字	四聲	犯規	棒球	特產	笑話	騙子

5.2 音素事後機率參數抽取

在進行分析之前必須先求得音素事後機率參數，而參數抽取的前級是 HMM 聲學辨識器，對語料強制對齊之後再進行音段辨認，最後將辨認出來的音素 likelihood 轉成事後機率向量，所以對齊出來的切割位置以及辨識結果的 likelihood 對音素錯誤偵測器的效能影響甚鉅，尤其現在是以 native speech 訓練出來的聲學模型去對 non-native 語料進行對齊以及辨識，語者的特性差異會使效能低落，所以傳統上使用語者調適(Speaker Adaptation) 【9】以獲得較好的辨識結果或切割位置，本實驗亦採取同樣的作法，接下來將介紹語者調適的作法以及調適後在辨識率和切割位置上的改進狀況，在彌補語者的特性差異後，即可完成如同前面章節所述的音素事後機率參數抽取。

5.2.1 語者調適

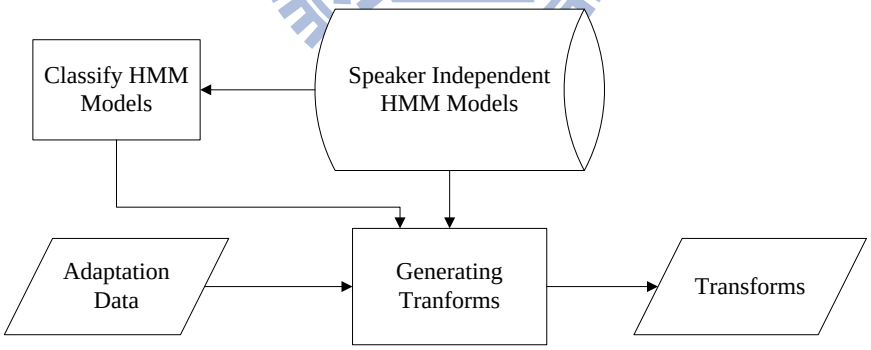


圖 5-1：語者調適方塊圖

如上圖所示，首先準備好調適用的語料，一般來說，對於每個語者使用約莫 30 秒的調適語料即可看出一定程度的效能提升【10】，所以本實驗針對每個語者挑選 30 秒左右的“正確”語料作為調適資料，大約佔總語料的六分之一，而這裡所謂的正確是指標音員沒有標記錯誤的語料，語料錯誤標記的部分稍後會再加以介紹，準備好調適用的語料後，接下來將 speaker independent 的 HMM 模型做分

類的動作，分類的依據是訓練聲學模型過程中產生的統計檔，以 Gaussian mixture 為單位分成特性相近的 16 群，對每群產生各自的轉換式，另外非語音的部分(長靜音、短停頓)不做分類以及轉換式的產生，因為這部分不會因人而異。這裡所求得的轉換式是準備用來做參數上(MFCC)的轉換，即是在強制對齊和音段辨識前先做特性轉換的前處理，所以本實驗的語者調適是指參數的調適。

5.2.2 調適前後的辨識效能和切割位置比較

辨識外國人語料採用前面章節所述的三連音 HMM 模型，表 5-2 是沒使用語言模型以及未經過語者調適的音節辨識率，其中包含各語者的辨認率，f01 表示第一位女性外國語者，m01 表示第一位男性外國語者，其餘依此類推，而表 5-3 是沒使用語言模型但經過語者調適後的音節辨識率，可以發現在語者調適過後提升了將近 20%的辨認率。

表 5-2：外國語料語者調適前的音節辨認率

```
================== HTK Results Analysis ==================
---------------------------- Speaker Results ------------------------------
spkr: %Corr( %Acc )   [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-----------------------------------------------------------------------------
f01:      33.00( 32.00)   [H=   66, D=   1, S=133, I=   2, N= 200]   13.00 [N=100]
f02:      38.50( 38.00)   [H=   77, D=   0, S=123, I=   1, N= 200]   16.00 [N=100]
f03:      34.00( 33.50)   [H=   68, D=   0, S=132, I=   1, N= 200]   14.00 [N=100]
m01:      28.00( 28.00)   [H=   56, D=   0, S=144, I=   0, N= 200]    9.00 [N=100]
m02:      42.00( 41.00)   [H=   84, D=   0, S=116, I=   2, N= 200]   20.00 [N=100]
m03:      26.50( 25.50)   [H=   53, D=   1, S=146, I=   2, N= 200]    5.00 [N=100]
m04:      27.00( 27.00)   [H=   54, D=   4, S=142, I=   0, N= 200]    5.00 [N=100]
m05:      29.00( 28.50)   [H=   58, D=   5, S=137, I=   1, N= 200]   10.00 [N=100]
m06:      34.00( 32.00)   [H=   68, D=   0, S=132, I=   4, N= 200]   11.00 [N=100]
---------------------------- Overall Results -------------------------------
SENT:    %Correct=11.44 [H=103, S=797, N=900]
WORD:  %Corr=32.44, Acc=31.72 [H=584, D=11, S=1205, I=13, N=1800]
=============================================================
```

表 5-3：外國語料語者調適後的音節辨認率

```
================ HTK Results Analysis =======================
---------------------------- Speaker Results ------------------------------
spkr: %Corr( %Acc )    [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-------------------------------------------------------------------------
f01:       50.00( 50.00) [H= 100, D=   1, S= 99,  I=   0, N= 200]    29.00 [N=100]
f02:       65.50( 65.00) [H= 131, D=   1, S= 68,  I=   1, N= 200]    46.00 [N=100]
f03:       45.50( 45.50) [H=  91, D=   0, S=109, I=   0, N= 200]    26.00 [N=100]
m01:       49.00( 48.50) [H=  98, D=   0, S=102, I=   1, N= 200]    23.00 [N=100]
m02:       58.50( 58.00) [H= 117, D=   0, S= 83,  I=   1, N= 200]    36.00 [N=100]
m03:       44.50( 44.00) [H=  89, D=   0, S=111, I=   1, N= 200]    27.00 [N=100]
m04:       57.00( 57.00) [H= 114, D=   1, S= 85,  I=   0, N= 200]    38.00 [N=100]
m05:       37.50( 37.50) [H=  75, D=   6, S=119, I=   0, N= 200]    15.00 [N=100]
m06:       59.00( 58.00) [H= 118, D=   0, S= 82,  I=   2, N= 200]    39.00 [N=100]
---------------------------- Overall Results ------------------------------
SENT: %Correct=31.00 [H=279, S=621, N=900]
WORD: %Corr=51.83, Acc=51.50 [H=933, D=9, S=858, I=6, N=1800]
=========================================================================
```

接下來觀察調適前後切割位置的差異，圖 5-2 中，從上往下分別是調適前的切割位置、調適後的切割位置、音檔波形圖以及音檔頻譜，可以發現調適後的位置略比調適前精準，其他音檔的切割位置也差不多是這樣的情況，雖然與 /p/ 發音方法相似的音都會有空白過長的現象，但本實驗不對此現象做探討，這樣的音素邊界品質已經足夠拿來測試音素錯誤偵測器的效能。

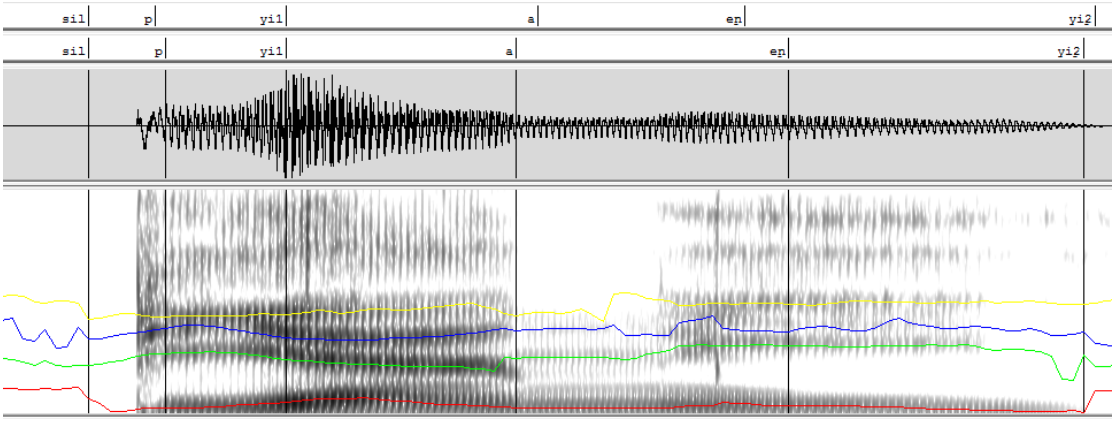


圖 5-2：外國語料調適前後的切割狀況比較

5.3 主觀錯誤標記

在錯誤分析之前必須要有測試語料的答案，而這裡的答案跟前面 native speech 語料的答案不同，前面說的答案即是文本(transcription)，但外國人語料不能盡信文本所標示，需要另請專人聆聽音檔後對有問題的發音做標記，本實驗請了 8 個 native speaker 來對所有音檔做錯誤標記，即是對總共 900 個共計 28 分鐘 14 秒的二字詞語料，忽略韻律上的錯誤，單就聲學上的錯誤進行錯誤標記，表 5-4 是各標音員間的交叉相關性(cross correlation)分析表，彼此間以相關係數(correlation coefficients)為相關性的量測值，標記時以“1”表示認為有錯的音，以“0”表示沒錯的音，表 5-5 為各標音員對總計 5,184 個音素的語料所標記的音素錯誤數量，由絕對數值和相關係數來看，大家的標音標準有一定程度的不一致性，這是主觀上所無法控制的。

$$R_{i,j} = \frac{cov(x_i, x_j)}{\sigma_{x_i}\sigma_{x_j}} = \frac{E\left[(x_i - \mu_{x_i})(x_j - \mu_{x_j})\right]}{\sigma_{x_i}\sigma_{x_j}} \tag{5.1}$$

其中 x_i 為第 i 個標音員標記的對錯序列向量，$cov(x_i, x_j)$ 為任兩標音員間的共變異數(covariance)，μ_{x_i} 、σ_{x_i} 分別為第 i 個標音員標記的向量平均值以及標準差(standard deviation)，$E[\cdot]$ 為期望值符號。

表 5-4：各標音員間的相關係數

	1	2	3	4	5	6	7	8
1	1.00	0.23	0.31	0.31	0.24	0.28	0.26	0.39
2		1.00	0.28	0.28	0.34	0.34	0.29	0.29
3			1.00	0.30	0.22	0.24	0.19	0.37
4				1.00	0.24	0.38	0.28	0.33
5					1.00	0.29	0.29	0.36
6						1.00	0.34	0.33
7							1.00	0.33
8								1.00

表 5-5：各標音員標記音素錯誤數量

標音員編號	1	2	3	4	5	6	7	8
標記錯誤音素數量(/5,184)	251	57	206	100	63	52	58	140

5.4 實驗結果分析

準備好測試語料參數以及主觀測試用的答案後，即可利用前面訓練的音素錯誤偵測器來進行主觀上的兩個實驗分析，首先，我們將探討根據標音員共同標記的錯誤數量，是否越多人標錯的音，其信心指數會分佈在較低的位置，而使用 Baseline 和 MLP 的音素偵測器在效能上的差異如何；再者，根據標音員的標音結果，在不同的制定錯誤標準下，探討整體的錯誤率以及 DCF 值等效能，是否 MLP 系統擁有較佳的偵錯能力。

5.4.1 主觀實驗一

本節以標音員標記的錯誤狀況來做音素錯誤偵測器的分析，表 5-6 為標音員共同對一個音素標記錯誤的音素數量表，將總計 5,184 個音素依照被標記錯誤的數量分成 8 群，其中有 4,692 個音素大家一致認為沒有問題，約佔 90.5%，而一個音素最多同時由 7 個人標記錯誤，總計有 12 個音素。

表 5-6：音素一致錯誤數量表

對同一個音素的標記錯誤數量	0	1	2	3	4	5	6	7	8
Phone 個數	4,692	299	79	53	29	9	11	12	0

首先從 baseline 系統的觀點，也就是對每個音素觀察其 1 維事後機率值，將以上這 8 群資料畫成 10 點 Histogram 圖，如圖 5-3 所示，將其換成百分比表格如表 5-5 所示，其中 0~5 六群有如預期的分佈狀況，錯誤標記越少的資料群聚在事後機率較高的位置，而 6~7 兩群雖然是大家一致認為錯誤的資料，卻還是有資料分佈在信心度較高的地方，這是較奇怪的部分。

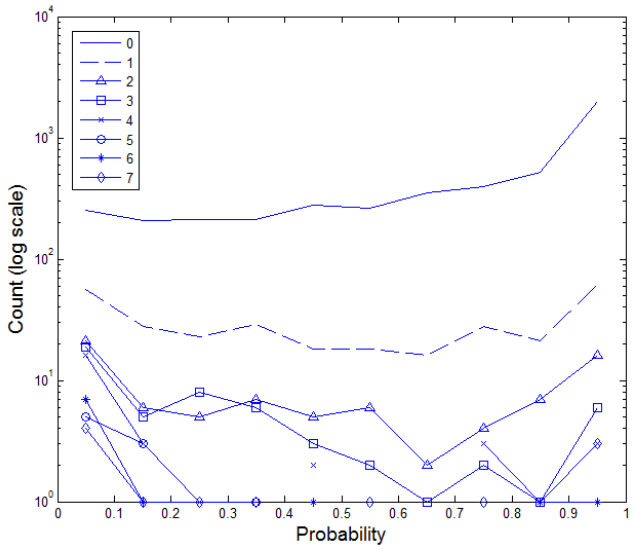


圖 5-3：主觀標記錯誤資料 1 維音素事後機率 Histogram 圖

表 5-7：主觀標記錯誤資料 1 維音素事後機率的 Histogram 百分比表

	0.0 ~0.1	0.1 ~0.2	0.2 ~0.3	0.3 ~0.4	0.4 ~0.5	0.5 ~0.6	0.6 ~0.7	0.7 ~0.8	0.8 ~0.9	0.9 ~1.0
0	5.4%	4.5%	4.5%	4.5%	5.9%	5.6%	7.4%	8.5%	11.0%	42.6%
1	18.7%	9.4%	7.7%	9.7%	6.0%	6.0%	5.4%	9.4%	7.0%	20.7%
2	26.6%	7.6%	6.3%	8.9%	6.3%	7.6%	2.5%	5.1%	8.9%	20.3%
3	35.8%	9.4%	15.1%	11.3%	5.7%	3.8%	1.9%	3.8%	1.9%	11.3%
4	55.2%	10.3%	3.4%	0.0%	6.9%	0.0%	0.0%	10.3%	3.4%	10.3%
5	55.6%	33.3%	0.0%	11.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
6	63.6%	9.1%	0.0%	0.0%	9.1%	0.0%	0.0%	0.0%	9.1%	9.1%
7	33.3%	8.3%	8.3%	8.3%	0.0%	8.3%	0.0%	8.3%	0.0%	25.0%

接下來使用前面訓練好的 MLP-1 系統，將 38 維音素事後機率當作輸入所得到的輸出，畫成 Histogram 再轉成百分比如表 5-8 所示，結果看起來優於表 5-7 的 baseline 結果，表示 MLP-1 系統對主觀錯誤標記資料有較佳的鑑別能力。

表 5-8：以 MLP-1 偵測器測試主觀標記錯誤資料的 Histogram 百分比表

	0.0 ~0.1	0.1 ~0.2	0.2 ~0.3	0.3 ~0.4	0.4 ~0.5	0.5 ~0.6	0.6 ~0.7	0.7 ~0.8	0.8 ~0.9	0.9 ~1.0
0	8.4%	3.5%	3.2%	3.5%	3.8%	4.1%	4.2%	7.2%	10.6%	51.4%
1	27.4%	6.0%	4.3%	5.4%	4.7%	3.3%	7.4%	6.7%	7.4%	27.4%
2	35.4%	6.3%	1.3%	5.1%	2.5%	6.3%	6.3%	2.5%	5.1%	29.1%
3	39.6%	18.9%	3.8%	11.3%	1.9%	1.9%	3.8%	1.9%	3.8%	13.2%
4	58.6%	10.3%	0.0%	0.0%	6.9%	0.0%	3.4%	0.0%	6.9%	13.8%
5	77.8%	11.1%	11.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
6	63.6%	9.1%	0.0%	0.0%	9.1%	0.0%	0.0%	0.0%	0.0%	18.2%
7	41.7%	0.0%	8.3%	0.0%	8.3%	0.0%	8.3%	8.3%	0.0%	25.0%

最後是 MLP-2 系統的結果，如表 5-9 所示，看起來在 0~4 五群效果優於 MLP-1 系統，而 5~7 三群效能較差。

表 5-9：以 MLP-2 偵測器測試主觀標記錯誤資料的 Histogram 百分比表

	0.0 ~0.1	0.1 ~0.2	0.2 ~0.3	0.3 ~0.4	0.4 ~0.5	0.5 ~0.6	0.6 ~0.7	0.7 ~0.8	0.8 ~0.9	0.9 ~1.0
0	3.2%	1.7%	1.6%	1.7%	2.0%	3.2%	4.7%	9.4%	17.5%	55.0%
1	13.7%	5.4%	4.0%	4.7%	3.0%	4.7%	5.4%	13.0%	18.1%	28.1%
2	21.5%	2.5%	3.8%	5.1%	3.8%	6.3%	5.1%	8.9%	12.7%	30.4%
3	20.8%	11.3%	3.8%	3.8%	9.4%	5.7%	13.2%	7.5%	13.2%	11.3%
4	44.8%	10.3%	3.4%	6.9%	3.4%	0.0%	3.4%	6.9%	10.3%	10.3%
5	66.7%	0.0%	0.0%	0.0%	11.1%	0.0%	11.1%	0.0%	11.1%	0.0%
6	54.5%	0.0%	0.0%	18.2%	0.0%	0.0%	0.0%	0.0%	9.1%	18.2%
7	25.0%	8.3%	8.3%	0.0%	0.0%	8.3%	8.3%	8.3%	8.3%	25.0%

5.4.2 主觀實驗二

此部分實驗一樣根據標音員的標音結果作測試，利用標音員的標錯結果進行投票，舉例來說，如果判定錯誤的標準較為嚴格，可以設定 3 個人以上標記為錯的音素即是錯，如果標準較為寬鬆，則可以設定 5 個人以上標記錯才算錯，首先

以 baseline 錯誤偵測器來偵測外國人二字詞語料，分別在三種標準下進行分析：3 個人、4 個人以及 5 個人以上標記錯誤的音素算錯誤，如表 5-10 所示，其中 N1 表示總共正確的個數，N0 表示總共錯誤的個數，其他的定義都跟前面章節一樣，整體來說，不管在怎樣的標準之下，Baseline 系統都有 70%左右的正確偵錯能力。

表 5-10：Baseline 錯誤偵測器對外國語料偵錯的整體結果

Standard	N1	N0	FA	FR	CA	CR	FAR	FRR	CAR	CRR	DCF
3	5,070	114	35	958	4,112	79	30.70%	18.90%	81.10%	69.30%	1,951
4	5,123	61	18	977	4,146	43	29.51%	19.07%	80.93%	70.49%	1,972
5	5,152	32	9	988	4,164	23	28.13%	19.18%	80.82%	71.88%	1,985

接著是 MLP-1 偵測器的部分，如表 5-11 所示，看起來錯誤偵測的部分跟 Baseline 差不多，但 DCF 值下降了，主要是因為錯誤拒絕變少了。

表 5-11：MLP-1 錯誤偵測器對外國語料偵錯的整體結果

Standard	N1	N0	FA	FR	CA	CR	FAR	FRR	CAR	CRR	DCF
3	5,070	114	38	936	4,134	76	33.33%	18.46%	81.54%	66.67%	1,910
4	5,123	61	18	949	4,174	43	29.51%	18.52%	81.48%	70.49%	1,916
5	5,152	32	9	960	4,192	23	28.13%	18.63%	81.37%	71.88%	1,929

最後是 MLP-2 偵測器，在錯誤拒絕的部分略優於 baseline 系統，錯誤偵測的部分跟 baseline 和 MLP-1 都差不多，猜測是因為標音員的標準太寬鬆，以及標音的不一致性使然，整體來說這個測試的部分以 MLP-1 錯誤偵測器效能最好，其他兩個則差不多。

表 5-12：MLP-1 錯誤偵測器對外國語料偵錯的整體結果

Standard	N1	N0	FA	FR	CA	CR	FAR	FRR	CAR	CRR	DCF
3	5,070	114	39	959	4,111	75	34.21%	18.92%	81.08%	65.79%	1,957
4	5,123	61	18	970	4,153	43	29.51%	18.93%	81.07%	70.49%	1,958
5	5,152	32	9	981	4,171	23	28.13%	19.04%	80.96%	71.88%	1,971

第六章 結論與未來展望

6.1 結論

本論文主要探討以音素事後機率為基礎的中文音素發音錯誤偵測器，總共做出三種偵測器，其一是傳統的方法，僅使用一維音素事後機率資訊做為音素的信心度評估指數；接著為了利用多維的資訊，引用了多層感知器網路作為 HMM 辨識器的後級，在訓練的方式上，使用跟聲學模型一樣的訓練語料；最後，期許透過滿足決策代價函數以及 MLP 網路訓練特性等訴求的取樣方式，重新訓練前面訓練出來的 MLP 音素錯誤偵測器得到改良過的第三個音素偵測器。

在測試效能的方面，嘗試了使用訓練語料的內部測試、Native Speech 的外部測試，以及外國人二字詞語料的兩種外部測試共計 4 種，其中前 3 種較能明顯顯示 MLP 偵測器優於傳統的偵測器，而至於最後一種測試則未能明顯看出 MLP 偵測器的優勢。

整體而言，使用 MLP 當作後級的偵測器有一定程度上的效能改進，但仍受限於前級 HMM 辨識器的效能，所以能提升的能力有限，加上 MLP 本身的參數調整也因使用需求而異，並沒有一個通用的調適方法，所以其使用價值仍有待商榷。

6.2 未來展望

本論文僅單純以聲學特徵為偵測對象，並沒考慮中文最重要的因素—聲調在內，加入聲調或是韻律的輔助以及適當的回饋較能對學習者有所幫助，畢竟就算音對了，聲調不正確則完全不能達意，所以未來可以考慮加入聲調的部分以提升整個中文錯誤偵測器的完整性。

參考文獻

【1】 O. Ronen, L. Neumeyer and H. Franco, “Automatic Detection of Mispronunciation for Language Instruction,” Proceedings Eurospeech 97, Rhodes, Greece, 649-652, 1997.

【2】 S. M. Witt, “Use of Speech Recognition in Computer-Assisted Language Learning,” PhD thesis, Department of Engineering, University of Cambridge, 1999.

【3】 S. M. Witt and S. J. Young, “Phone-level Pronunciation Scoring and Assessment for Interactive language learning,” Speech Communication 30, 95-108, 2000.

【4】 H. Franco, L. Neumeyer, M. Ramos and H. Bratt, “Automatic Detection of Phone-Level Mispronunciation for Language Learning,” Proceedings Eurospeech 99, Budapest, Hungary, 851-854, 1999.

【5】 TCC-300 麥克風語音資料庫使用說明

【6】 Eskenazi, M., “An overview of Spoken Language Technology for Education”, Speech Communication, 2009.

【7】 K. Hornik, M. Stinchcombe and H. White (1989). Multilayer feedforward networks are universal approximators. Neural Networks, 2, 359-366.

【8】 S. Haykin, “*Neural Networks*,” Yew York, 1994.

【9】 D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, K. Hirose, "Analysis and utilization of MLLR speaker adaptation technique for learners' pronunciation evaluation, " Proc. INTERSPEECH2009, pp.608-611 2009-9.

【10】 S. Yong, “*The HTK Book (Version 3.4)*,” 2009.

附錄一：中文音素發音對照表

FNULL1	ㄓㄔㄕㄖ韻母	**n**	ㄋ
FNULL2	ㄗㄘㄙ韻母	**o**	ㄛ
a	ㄚ	**p**	ㄆ
b	ㄅ	**q**	ㄑ
ch	ㄔ的子音	**r**	ㄖ的子音
c	ㄘ的子音	**s**	ㄙ的子音
d	ㄉ	**sh**	ㄕ的子音
en	ㄣ的韻尾	**t**	ㄊ
er	ㄦ	**wu1**	ㄨ在介音
e	ㄜ	**wu2**	ㄨ在韻腹
eh	ㄝ	**wu3**	ㄠ的韻尾
f	ㄈ	**x**	ㄒ
g	ㄍ	**yi1**	ㄧ在介音
h	ㄏ	**yi2**	ㄧ在韻腹
j	ㄐ	**yi3**	ㄞ的韻尾
k	ㄎ	**yu1**	ㄩ在介音
l	ㄌ	**yu2**	ㄩ在韻腹
m	ㄇ	**zh**	ㄓ的子音
ng	ㄤ的韻尾	**z**	ㄗ的子音

附錄二：TCC300 語料庫音素資料量

訓練語料			
a	94,410	**g**	14,117
e	76,351	**b**	14,034
yi1	60,804	**h**	12,741
en	54,001	**t**	11,381
ng	50,450	**z**	10,651
wu1	46,954	**q**	10,553
wu3	29,686	**yu1**	9,718
yi3	29,176	**ch**	9,691
d	26,986	**f**	8,870
yi2	25,994	**m**	8,676
o	25,658	**yu2**	8,043
eh	24,589	**r**	7,860
sh	22,744	**n**	6,378
wu2	21,378	**s**	5,734
j	21,239	**k**	5,667
zh	18,844	**c**	4,879
FNULL1	17,285	**FNULL2**	4,830
x	16,306	**p**	4,073
l	14,654	**er**	2,185
Total		837,590	

測試語料			
a	9,829	**b**	1,481
e	7,813	**g**	1,473
yi1	6,389	**h**	1,431
en	5,668	**t**	1,183
ng	5,146	**q**	1,168
wu1	4,836	**z**	1,147
yi3	3,121	**yu1**	1,081
wu3	3,084	**ch**	983
yi2	2,774	**m**	866
o	2,756	**yu2**	857
d	2,719	**f**	800
eh	2,570	**r**	788
j	2,300	**n**	668
sh	2,231	**s**	604
wu2	2,136	**k**	535
zh	1,958	**FNULL2**	533
FNULL1	1,770	**c**	514
x	1,657	**p**	442
l	1,623	**er**	257
Total		87,191	

附錄三：音素索引對照表

index	phone	index	phone
1	**FNULL1**	20	**n**
2	**FNULL2**	21	**o**
3	**a**	22	**p**
4	**b**	23	**q**
5	**ch**	24	**r**
6	**c**	25	**s**
7	**d**	26	**sh**
8	**en**	27	**t**
9	**er**	28	**wu1**
10	**e**	29	**wu2**
11	**eh**	30	**wu3**
12	**f**	31	**x**
13	**g**	32	**yi1**
14	**h**	33	**yi2**
15	**j**	34	**yi3**
16	**k**	35	**yu1**
17	**l**	36	**yu2**
18	**m**	37	**zh**
19	**ng**	38	**z**

附錄四：外國人語料音素資料量

phone	amount	phone	amount
a	720	**z**	90
yi1	387	**ch**	81
e	369	**j**	81
en	351	**f**	72
wu3	279	**h**	72
wu1	261	**s**	72
ng	225	**n**	63
eh	198	**FNULL1**	54
o	198	**g**	54
yi3	171	**p**	54
x	153	**zh**	54
sh	117	**FNULL2**	45
yi2	108	**c**	45
b	99	**er**	45
d	99	**k**	45
q	99	**r**	45
l	90	**yu1**	45
m	90	**yu2**	36
t	90	**wu2**	27