國 立 交 通 大 學

電信工程研究所

碩士論文

智慧型家用機器人使用之語者辨識系統
Speaker Recognition System for Intelligent Home Robot

研 究 生：吳宜樵

指導教授：王逸如 博士

中華民國一百年八月

智慧型家用機器人使用之語者辨識系統

Speaker Recognition System for Intelligent Home Robot

研 究 生：吳宜樵　　Student：Yi-Chiao Wu

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 Departmant 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

August 2011

Hsinchu, Taiwan, Republic of China

中華民國一百年八月

智慧型家用機器人使用之語者辨識系統

研 究 生：吳宜樵　　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

本研究探討語者辨識系統於智慧型家用機器人上以及家用環境內，所會面臨的各式使用情境，並針對情境開發設計我們的語者辨識系統及註冊流程，以期使用者能更方便直覺地使用本系統。此外，因為家用環境中通常富含各式雜訊，且使用者在使用機器人時必定與機器人有一段距離，所以將文本獨立語者辨識前端整合麥克風陣列、波束形成和空間濾波器的技術，以提高在雜訊環境下的強健性。另一方面也考慮到在家用機器人中，與其他系統整合的可能性，在系統輸出端則提供辨識結果有效與否的判定及信心指數，以利與其他使用者辨識系統整合。最後為了更貼近家庭使用者的使用習慣與喜好，並降低所需註冊及測試語料的秒數，而發展出可用任何語言輸入註冊文本資訊及註冊語料的文本相關語者辨識系統。

Speaker Recognition System for Intelligent Home Robot

Student：Yi-Chaio Wu　　Advisor：Dr. Yih-Ru Wang

Department of Communication Engineering
National Chiao Tung University

Abstract

In this paper we present a speaker recognition system, which is specifically designed for intelligent home robot. The enrollment procedure of the system is designed to fit home scenarios and make it easier to be used. Besides, the performance of speaker recognition system degrades significantly in home environment because of reverberation, noise and the distant between speakers and microphone array. The spatial information from the microphone array, which couples with beam forming and spatial voice activity detection, makes the system more robust in the noisy environment. On the other hand, our system provides the confidence scores so as to be fused with other recognition results and achieve the integration of user recognition systems. Finally, in order to increase the convenience of using systems and reduce requirements of enrollment and test data, we develop a text-dependent speaker recognition system, which can be used in any language.

致謝

僅以此致謝記住所有曾在這趟旅途中，對我而言非常重要的人。陳信宏老師、王逸如老師，江振宇、楊智合、黃信德學長，吳文良、賴智誠、許昱超、劉銘傑、林彥邦、劉冠驛、鍾進竹等同學，電信系 98 級的各位同學好友，電信所、電機系的學長姐、學弟妹，我的家人及所有幫助過我的人，沒有你們我不可能完成這本論文。大恩不言謝，大家後會有期。

目錄

中文摘要.....I

Abstract.....II

致謝.....III

目錄.....IV

表目錄.....VII

圖目錄.....VIII

第一章 緒論.....1

1.1 研究動機.....1

1.2 文獻探討.....2

1.2.1 文本獨立語者辨識.....2

1.2.2 文本相關語者辨識.....3

1.2.3 語者識別與驗證.....4

1.3 研究方向.....7

1.4 章結概要說明.....7

第二章 文本獨立語者辨認系統簡介.....9

2.1 語者辨認基本系統.....9

2.1.1 參數抽取.....9

2.1.2 通用背景模型訓練.....10

2.1.3 語者註冊.....11

2.1.4 語者辨識.....13

2.2 整合麥克風陣列.....13

2.2.1 波束形成.....14

2.2.2 語音端點偵測.....14

2.3 實驗結果.....16

2.3.1 語者識別實驗語料及結果.....16
2.3.2 整合麥克風陣列實驗語料與結果.....18
3.1 語者身分驗證與信心指數.....23
3.1.1 通用模型正規化.....24
3.1.2 最大值正規化.....26
3.1.3 幾何平均數正規化.....27
3.1.4 信心指數曲線及門檻值.....29
3.2 實驗結果.....30
3.2.1 語者驗證實驗語料與結果.....30
3.2.2 語者驗證信心指數.....44
第四章 文本相關語者註冊設計.....48
4.1 文本相關語者辨識系統.....48
4.1.1 語者無關聲學模型之建立.....48
4.1.2 語者註冊.....49
4.1.3 語者辨識.....50
4.2 註冊流程設計.....51
4.2.1 語音輸入註冊資訊.....51
4.2.2 混合語言註冊系統.....53
4.3 實驗結果.....53
4.3.1 訓練、註冊與測試語料.....53
4.3.2 基礎文本相關語者辨識系統實驗.....54
4.3.3 語音輸入註冊資訊的語者辨識系統實驗.....56
4.3.4 混合語言語者辨識系統實驗.....59
第五章 結論與未來展望.....61
5.1 結論.....61
5.2 未來展望.....61

參考文獻.....62
附錄一：中文通關密碼文本.....64
附錄二：英文及台語通關密碼文本.....66

表目錄

表 2.1：TCC300 語料資料統計表.....16
表 2.2：語者辨識基礎系統辨識率.....17
表 2.3：乾淨且經過波束形成註冊語料辨識率.....20
表 2.4：乾淨經過波束形成及語音端點偵測處理之註冊語料辨識率.....21
表 2.5：環境雜訊匹配辨識率.....21
表 3.1：通用模型正規化法目標語者與封閉集合冒名頂替者分數分布統計.....36
表 3.2：通用模型正規化法辨識正確及辨識錯誤的分數分布統計.....38
表 3.3：最大值正規化法辨識正確及辨識錯誤的分數分布統計.....40
表 3.4：幾何平均數正規化法辨識正確及辨識錯誤的分數分布統計.....42
表 3.5：1 秒正規化方法選用及辨識結果信心指數平均值.....44
表 3.6：2 秒正規化方法選用及辨識結果信心指數平均值.....45
表 3.7：3 秒正規化方法選用及辨識結果信心指數平均值.....45
表 3.8：正規化方法誤判集合匹配程度.....46
表 4.1：基礎系統文本相關語者辨識率.....54
表 4.2：基礎文本相關語者驗證封閉集合等錯率.....55
表 4.3：基礎文本相關語者驗證開放集合等錯率.....56
表 4.4：有文法限制音素序列辨識器辨識結果統計.....57
表 4.5：語音輸入註冊文本資訊語者辨識系統辨識率.....57
表 4.6：語音輸入註冊文本資訊語者驗證封閉集合等錯率.....58
表 4.7：語音輸入註冊文本資訊語者驗證開放集合等錯率.....58
表 4.8：新文法限制音素序列辨識器辨識結果統計.....59

圖目錄

圖 2.1：通用背景模型訓練方塊圖.....10
圖 2.2：語者註冊方塊圖.....11
圖 2.3：語者辨識方塊圖.....13
圖 2.4：廣義旁瓣消除器.....14
圖 2.5：語者辨識基礎系統辨識率.....18
圖 2.6：麥克風陣列位置圖.....19
圖 3.1：語者確認系統.....23
圖 3.2：信心指數算法.....29
圖 3.3：通用模型正規化法目標語者與冒名頂替者分數分布.....30
圖 3.4：通用模型正規化法接收者操作特徵曲線圖.....31
圖 3.5：辨識正確及辨識錯誤分數分布.....33
圖 3.6：通用模型正規化法接收者操作特徵曲線圖.....33
圖 3.7：最大值正規化法接收者操作特徵曲線圖.....34
圖 3.8：幾合平均數正規化法接收者操作特徵曲線圖.....34
圖 3.9：通用模型正規化法 1、2、3 秒目標語者與封閉集合冒名頂替者分數分布.....36
圖 3.10：通用模型正規化法 1、2、3 秒接收者操作特徵曲線圖.....37
圖 3.11：通用模型正規化法 1、2、3 秒辨識正確及辨識錯誤的分數分布.....38
圖 3.12：通用模型正規化法 1、2、3 秒接收者操作特徵曲線圖.....39
圖 3.13：最大值正規化法 1、2、3 秒辨識正確及辨識錯誤的分數分布.....40
圖 3.14：最大值正規化法 1、2、3 秒接收者操作特徵曲線圖.....41
圖 3.15：幾何平均數正規化法 1、2、3 秒辨識正確及辨識錯誤的分數分布.....42
圖 3.16：幾何平均數正規化法 1、2、3 秒接收者操作特徵曲線圖.....43
圖 4.1：聲學模型之建立流程.....48

圖 4.2：文本相關語者註冊流程.....49
圖 4.3：文本相關語者辨識流程.....50
圖 4.4：註冊語料音節辨識.....51
圖 4.5：註冊語料調適.....52
圖 4.6：決定最佳音節辨識結果.....52
圖 4.7：基礎文本相關語者驗證封閉集合操作者接收曲線圖.....55
圖 4.8：基礎文本相關語者驗證開放集合操作者接收曲線圖.....56
圖 4.9：語音輸入註冊文本資訊語者驗證封閉集合操作者接收曲線圖.....57
圖 4.10：語音輸入註冊文本資訊語者驗證開放集合操作者接收曲線圖.....58
圖 4.11：混合語言開放集合操作者接收曲線圖及等錯率.....60

第一章 緒論

1.1 研究動機

本論文的目的在於能夠以人為出發點，發展出人性化的產品。而家用環境幾乎為每個人共通擁有的使用情境，因此如何就目前各項日新月異的科技技術，確切地發展出符合人們日常家庭生活需求的產品，是目前科技發展的重點。我們為了能提供不同使用者利用最直覺的方式，使用智慧型家用產品裡各項客製化的服務，所以發展語者辨識系統，讓使用者只需透過聲音，即可讓系統得知其身分，並針對使用者而做出各式各樣適當的服務。

最基本的語者辨識系統，可以同時做到語者識別（speaker identification）找出可能的語者排名，以及語者身分驗證（speaker verification）確認語者是否為其所宣稱之人，並提供信心指數。而對於家用環境而言，使用者辨識使用的情境、環境皆相當複雜多變，因此整合各項技術以提供更貼近人們生活的服務是必然的趨勢。例如將語者辨識與語音辨識系統整合，對語音辨識系統提供語者資訊，以增加其對於已知使用者情況下的辨識率；亦或是藉由語者辨識結合人臉辨識系統，以發展更具強健性的使用者辨識系統；更進一步還可與各種個人工具軟體結合，提供使用者客製化服務。

此外將使用範圍擴大到真實生活中，許多為了安全或是便利等因素的情境，例如：車內環境，多人的會議室...，麥克風通常擺放在固定位置，離每個使用者皆有一段距離，因此本論文將前端將整合麥克風陣列及波束形成（beam forming）系統，並藉由探討如何處理經過麥克風陣列及波束形成系統的語料，以期能讓系統在一般環境下能有不錯的辨識率，而未來還能更進一步地應用在例如：會議時不同語者的語音歸檔；車內空間時，駕駛與其他成員的聲控系統；家內環境，每個家庭成員所需客製化的服務提供等實際情況中。

1.2 文獻探討

1.2.1 文本獨立語者辨識

文本獨立（text independent）語者辨識的技術有三個主要的分類【1】，第一類也是最早的技術為使用長期統計（long term statistic）的語音參數，例如頻譜或是音調等做為辨識依據。其概念在於將除了語者相關的聲學因子，其他像是不同音節等所造成的聲學差異藉由平均的方式消除掉，只留下代表語者平均聲道（vocal tract）形狀的長期頻譜平均值等語者相關的聲學參數。然而其壞處在於需要相當長的註冊語料去產生穩定的長期統計模型，且丟棄了許多聲學上有用的語者資訊。

第二類技術為將註冊語料分為幾個語音單元，並由這些單元的語音參數來為每位語者訓練各自的語者模型，而在辨識時藉由比較測試語料中每群語音單元的與每個語者模型中相對應語音單元的相似度，來分辨測試語料屬於哪一個語者。此技術又可以再細分為兩種不同的切割語音單元的方式，分別為顯式分段（explicit segmentation）與隱式分段（implicit segmentation）。顯式分段為在註冊或辨識前，就先做語音辨識並以辨識結果切割出每個語音單元，但在【1】裡提到先做語音辨識不僅增加計算量，且對於語者辨識的幫助不大，因此在文本獨立的語者辨識範圍裡，比較常用的是隱式分段的方法。隱式分段是在訓練或辨識前用非監督式分類法（unsupervised clustering）做語音單元切割，而每個分類是沒有標籤的，所以並不需要依標籤個別訓練模型。而隱式分段下又有幾種形式，像是分類樣板形式的向量量化編碼（vector quantization, VQ），就是將每個語音單元所得的語音參數做分群，並用記錄每群頻譜樣版的碼本（codebook）來代表語者，也就是每位語者的語音參數用其碼本去量化會有最小的量化誤差，並以此條件來做語者辨識。向量量化雖然在有限詞彙裡的語者辨識效果不錯，但因其本質較難以去代表每群內在真實情況裡的變異，所以在較大詞彙、噪音環境或是有通道效應的文本獨立語者辨識裡，我們通常使用機率模型去提供一較佳聲學模型，

例如高斯混和模型（Gaussian mixture model, GMM）或是隱藏式馬可夫模型（hidden Markov model, HMM）就常應用於文本獨立或文本相關的語者辨識。

第三類技術為使用鑑別式類神經網路（discriminative neural network），其特點在於並非為每位註冊語者訓練各自的語者模型，而是找出分辨出所有註冊語者最佳決策方程式。而其好處在於相對於為每位語者各自訓練模型，可以使用較少的參數，但卻達到差不多辨識率。但其缺點在於，每當加入新的註冊語者，則整個辨識模型都要重新訓練產生。

而在本篇論文裡我們選用屬於第二類中隱式分段的高斯混和模型來當作我們的語者模型，因為在【1】裡面有提到，高斯混和模型是大家非常熟悉的且簡單的模型，所以在計算上相當的方便。另一方面高斯混和模型可以簡單的用來模擬任何機率分布，且即使原本的機率分布因資料量較小而不平滑，也可以透過用高斯混和模型模擬的方式使其平滑。此外，許多語者相關的聲學特性，以及真實地反映人類口腔等不同的特性，可以用高斯混和模型簡單的去代表。

因為是即時系統，所以不可能有大量的訓練語料，但很多語音基礎的特性因為訓練語料的稀少，而無法全部涵蓋，使得我們的語者模型無法正確的代表出語者相關與語者共通不同的特性，且少量語料可能對模型造成過適（over fitting）的現象，這些缺點都會對辨識率造成很大的影響。為了解決此一問題，在【2】裡面提到了，在系統建立前，用相對大量的訓練語料先行訓練一個通用背景模型（universal background model, UBM），此模型涵蓋了大部分語者共同的聲學特性，而在系統要註冊語者時，再用最大事後機率法則（maximum a posteriori probability, MAP）調適通用背景模型成為每位語者各自的高斯混和模型，如此語者模型不但包含語者本身的聲學特性也包含語者間共通的聲學特性。

1.2.2 文本相關語者辨識

文本相關（text dependent）語者辨識因為其準確性與針對性，是目前在商業化應用上最被廣泛使用的語者辨識技術。在【3】裡面提到，傳統的文本相關語

者辨識技術可以分為兩類。第一類為動態時軸校準（dynamic time warping, DTW），最典型的方法是由 Furui 在 1981 年提出的頻譜樣版比對（spectral template matching approach），用一序列的特徵參數向量去做為每位語者的樣板，並在測試時藉由比對測試語料的特徵參數向量序列與每位註冊語者特徵參數向量序列樣版的距離，決定辨識的結果；第二類為隱藏式馬可夫模型，在【3】、【4】裡都提到本質上隱藏式馬可夫模型不是直接使用特徵參數向量序列做為樣板，而是對語言中基本的音節或音素訓練成包含多個由高斯混和模型組成的狀態，且狀態之間有方向性及轉移機率的一序列語音模型，所以較不易受到說話快慢等因素的影響，比起動態時軸校準更能有效率的去代表文字相關的聲學模型。

1.2.3 語者識別與驗證

語者辨識系統主要可分為兩個基礎功能，分別為語者識別及語者驗證【3】。語者識別就是在一個已知的註冊語者模型集合內，找出測試語料最有可能來自的那位語者，所以又稱為封閉集合（closed set）辨識。而語者驗證則為確認使用者是否為其所宣稱之語者，而冒名頂替的使用者（impostor）有可能是我們已知集合的其它語者，或是來自於已知集合之外，所以又稱為開放集合（open set）辨識。而其實兩個功能也可以看成同一個問題，也就是每當有使用者辨識時，系統除了已知註冊語者之外還多了一個非已知註冊語者的選項。

而在【5】、【7】裡都提到任何的驗證問題，皆可視為統計假設檢定（statistical hypothesis testing）。統計假設一般有兩種形式，其一為虛無假設（null hypothesis）是我們欲證明其為錯的假設，以 H_0 表示；另一為對立假設（alternative hypothesis）是虛無假設的反面，以 H_1 表示。而我們必須找出足夠的證據否定 H_0，否則就接受 H_0 為真。而在【7】裡也提到不管是在語者辨識或是語音辨識，我們都必須要做離群值偵測（outlier rejection），對語者辨識而言就是驗證是否為冒名頂替的使用者。一般辨識系統通常分成兩個階段，第一階段先將測試資料做模式分類（pattern classification），第二階段再做離群值檢測。所以套用到語者辨識系統

裡，則首先找出與測試語料最相似的已註冊語者模型，接著驗證其是否真的為這個語者（屬於這個分群，H_0），亦或是冒名頂替的使用者（離群值，H_1）。

此外【7】裡提到在我們得知H_0、H_1的機率分布的前提下，根據奈曼-皮爾生引理(Neyman - Pearson lemma)，最佳解為使用概似比檢測（likelihood ratio test, LRT）。概似比檢測的意義在於，比較兩種模型何者較適合詮釋我們的統計資料，應用在語者辨識上，則是比較此語者的模型與非此語者的模型何者與我們的測試語料較相近。因此其決策原則（decision rule）就是

$$\frac{p(X \mid \lambda_{hyp})}{p(X \mid \lambda_{\overline{hyp}})} \begin{cases} > \eta, X \in H_0 \\ \leq \eta, X \in H_1 \end{cases} \qquad (1\text{-}1)$$

而寫成對數型式則為

$$\Lambda(X) = \log p\left(X \mid \lambda_{hyp}\right) - \log p\left(X \mid \overline{\lambda}_{hyp}\right) \qquad (1\text{-}2)$$

而一般情況中H_0模型容易得到，也就是我們的語者模型，但要估計H_1模型卻很困難，我們無法取得足夠的資料去完全估計H_1模型，而在【2】、【6】、【7】裡提到，有兩種主要的方法去估計H_1模型，第一種以通用背景模型來代表，因此我們將語者辨識結果所得到的對數概似值減去語料對於通用背景模型的對數概似值，才是我們概似比檢測的分數。而η一般經由大量的實驗結果，畫出接收者操作特徵(receiver operating characteristic, ROC)曲線，並依此訂定信心指數。在【9】裡面稱此方法為通用模型正規化（World model normalization, WMN），其主要的精神在於冒名頂替的使用者語料對於目標語者模型（target speaker model）與通用背景模型的分數差距，應該是小於目標語者的語料對於目標語者模型與通用背景模型的分數差距。而使用通用背景模型的好處，在於只需訓練一個通用背景模型，即可代表所有或某一特定集合目標語者的對比假設模型，且另一方面此通用模型還可以用於調適語者模型的辨識系統裡。

第二種方法則是使用一非目標語者的語者模型集合去代表H_1，而這集合一般可稱為同儕集合（cohort set）、概似比集合（likelihood ratio sets）或背景語者

（background speakers）【2】。同儕集合又可分為兩類，第一類為封閉同儕集合（closed cohort set），假設我們現在有 20 位已註冊語者，則每一個語者模型其封閉同儕集合則為其他 19 個語者模型，也就是對於目標語者而言，其他 19 位已註冊語者為冒名頂替的使用者；第二類為開放同儕集合（open cohort set），如果我們現在有 20 位已註冊位語者，則每個語者模型其開放同儕集合皆不與這 20 個已註冊語者模型重疊，也就是對於目標語者而言，所有非註冊語者為冒名頂替的使用者。應用到語者辨識系統上，則為系統所面對的兩類問題，第一類為已註冊語者使用時，其辨識結果的正確性；第二類為非註冊語者使用時，系統是否可偵測出其為冒名頂替的使用者。

兩種方法的比較上，因為目標語者與冒名頂替的使用者對於一包含完整語音特性的通用背景模型所得之分數差異較小，因此用同儕集合來估計 H_1 模型，或許能估計得較細微，而有較好的分辨效果。而在【10】裡面也提到，較有參考價值的同儕集合是那些與我們的目標語者模型較靠近的模型，因其帶有較大的資訊量，較能做為我們檢測的依據。在【11】裡面則提到，同儕集合人數數量在 10 個人以下，等錯率（equal error rate，EER）是小於通用背景模型，而在 15 個人以後則高於通用背景模型。因人數太多反而造成混淆，語者模型之間的關係變得複雜，所以反而不如用通用背景模型來得好。在【6】裡面也提到，同儕集合大小的選法，是根據最大概似法則來選擇,假設我們希望用一大小為 M 的同儕集合來估計 H_1 模型，則找出在同儕集合中最接近目標語者的 M 個模型做為同儕集合。而同儕集合的大小對於開放同儕集合的影響較封閉同儕集合顯著，封閉同儕集合的等錯率隨著同儕集合內語者模型的數量上升，很快即達到一收斂值，且等錯率與模型數量少時差別不大。

此外【8】、【9】裡都提到了測試分數正規化（testing normalization, T-norm）與平均值正規化（zero mean normalization, Z-norm）。平均值正規化是在系統運作前利用同儕集合的語料當作冒名頂替使用者的語料，訓練出用來正規化的平均值

及標準差，目的是希望能獲得較好的整體臨界值（global threshold）η。而測試分數正規化則是在系統運作時，利用同儕集合裡的語者模型當作冒名頂替使用者的模型，並使測試語料經過冒名頂替使用者的模型，訓練出用來正規化的平均值及標準差，目的是希望藉此找出目標語者模型與冒名頂替使用者的模型之間的距離是否夠遠，又或是多少的距離才算是夠遠。而在【8】裡面也提到，測試分數正規化與平均值正規化相比的好處就是不會有訓練語料及測試語料間，可能因麥克風、環境等因素而造成不匹配的問題。

1.3 研究方向

本論文目標為將系統建構在一般家用個人電腦上，包含完整的語者註冊及語者辨識功能，且搭配設計過的註冊流程及文本， 使其能輕易的與其他需要得知使用者身分的軟體整合。

基礎系統為文本獨立語者辨識系統，每位註冊語者在註冊時，將系統內的通用背景模型調適成語者各自的高斯混和模型。辨識時，則用最大概似機率（maximum likelihood）法則，找出最有可能辨識結果，並提供本次測試是否達有效測試門檻及其信心指數。語者辨識系統前端與麥克風陣列及波束形成（beam forming）系統整合，並測驗其在各種情況下的辨識率，以期能增加語者辨識系統的強鍵性。

接著本論文為更進一步增強使用者與系統的互動關係，且引入同樣常用於使用者身分辨識的通關密碼，將使用隱藏式馬可夫模型發展文本相關語者辨識系統，使用者可以用任意姓名註冊或是替系統取名字，並藉由此特殊通關密碼來做語者辨識。

1.4 章結概要說明

本論文的內容共分為五章：

第一章 緒論：介紹本論文之研究動機、研究方向、語者辨認的基礎方法。

第二章 文本獨立語者識別系統簡介：介紹本論文的文本獨立語者識別系統

及其實驗數據。

第三章 文本獨立語者確認系統簡介：介紹本論文的文本獨立語者確認系統使用之方法及其實驗數據。

第四章 文本相關語者註冊設計：介紹基於文本相關語者辨識系統，所設計之註冊流程。

第五章 結論與未來展望。

第二章 文本獨立語者辨認系統簡介

文本獨立語者辨認為最基礎的語者辨認系統，可以廣泛地應用於各種情況。本章將描述本論文所使用之文本獨立語者辨認系統，以及為了適應家用環境場而將麥克風陣列整合入系統。2.1 節介紹本研究所採用的使用最大事後機率法則調適語者通用模型的語者辨認系統；2.2 節介紹基礎系統前端整合線性麥克風陣；2.3 節介紹語者識別及整合麥克風陣列實驗結果。

2.1 語者辨認基本系統

整個語者辨識系統主要可分為註冊及辨識兩個階段，而註冊又包含了背景通用模型的訓練及註冊語者模型，本節將依序介紹參數抽取、註冊及辨識。

2.1.1 參數抽取

梅爾倒頻係數 (Mel-frequency cepstrum coefficients, MFCCs)，因為其接近人耳對語音區別性的特性，所以在語音或語者辨識系統裡被普遍地使用。而本論文為了在即時系統裡可以有基本的抗噪效果，因此將所求出的梅爾倒頻係數再經過RASTA濾波器處理，詳細過程如下：我們由麥克風收錄取樣頻率8000赫茲（Hertz）、取樣位元數為16位元的語料，音框大小（frame size）取240點、音框位移（frame shift）取80點，首先將語音訊號做預強處理(pre-emphasis)，如下式所示：

$$H_p(z) = 1 - \alpha z^{-1} \qquad (2\text{-}1)$$

其中為了降低運算量，令α=0.9375=1-2^{-4}。接著做256點的快速傅立葉(FFT)轉換將訊號轉至頻域，然後通過一組20個「三角帶通濾波器」(triangular band pass filter)，再經餘弦轉換(discrete cosine transform)後得到12維的梅爾倒頻係數。之後將倒頻係數通過RASTA濾波器:

$$H_r(z) = 0.1z^4 \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \qquad (2\text{-}2)$$

因為特徵參數軌跡(time sequence of spectral parameter)低頻部份通常受到通道效應汙染較為嚴重，而RASTA為一帶通濾波器可用來壓抑特徵參數軌跡的低頻部份，使其能達成去除部分通道效應的目的。最後求得26維的語音特徵參數，包含了12維的RASTA-based MFCC，12維一階差量RASTA-based MFCC、1維一階差量log energy以及1維二階差量log energy。

2.1.2 通用背景模型訓練

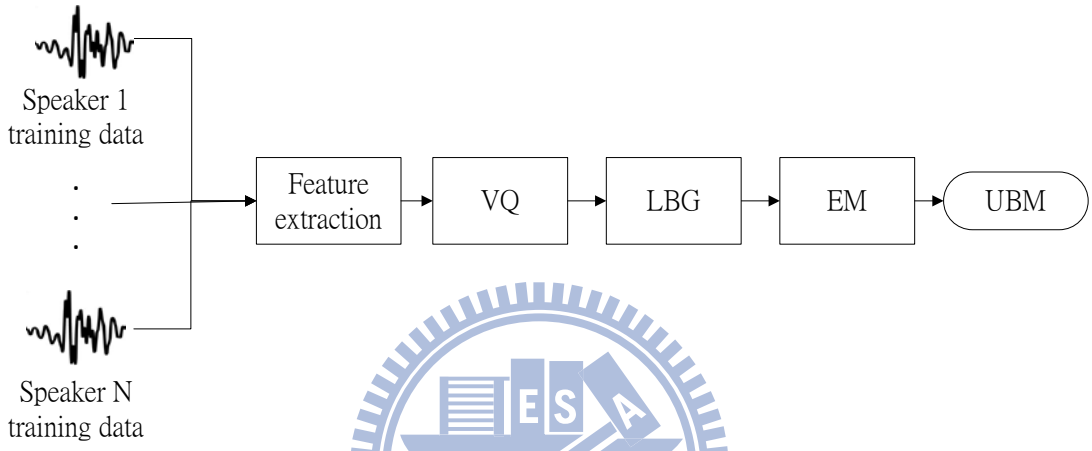


圖 2.1：通用背景模型訓練方塊圖

一般的文本獨立語者辨識系統是利用高斯混合模型來做為每一個語者之語音參數分布統計模型，但是通常無法對每位語者都取得大量的註冊語料，所以先訓練一個通用背景模型，並使通用背景模型盡可能涵蓋語音信號中所有音素，以求其能真正表現出語音訊號的語者無關的特性，再將通用背景模型調適為每位語者的語音參數分布模型，如此可避免單一語者訓練及測試語料涵蓋音素不同時，效能下降的問題。如【2】裡所示，一般論文通常用 512 維或是 1024 維的通用背景模型，而每位語者再由其中挑選 64 維對於註冊語料概似值較高的高斯分布代表，但在本文系統中為了降低運算量，通用背景模型直接訓練為 64 維。具體訓練過程如圖 2.1 所示，先將 TC300 語料庫中 300 位語者，每位取 24 秒訓練語料抽成語音參數後，做向量量化編碼做為 64 維通用背景模型的初始值，再依據 LBG（Linde-Buzo-Gray）演算法反覆更新碼本直到收斂為止。最後再使用最大期望（Expectation-Maximization, EM）演算法對模型各項參數做調適。

2.1.3 語者註冊

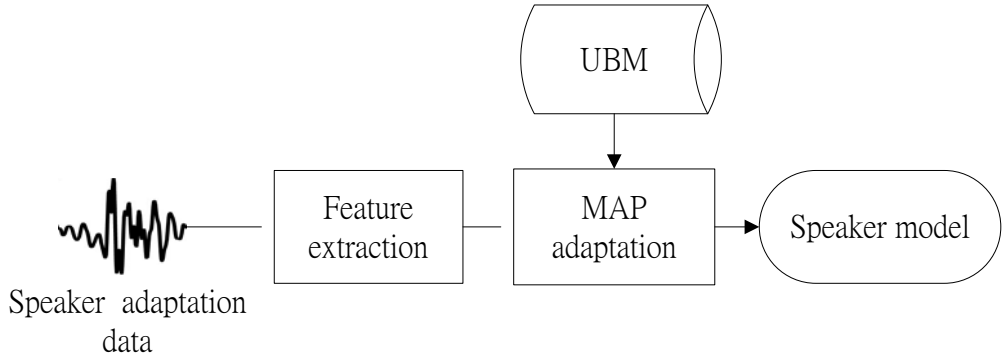


圖 2.2：語者註冊方塊圖

如圖 2.2 所示，每位使用者註冊時，用系統麥克風錄下 24 秒註冊語料，語料經過特徵參數抽取為 26 維 RASTA 參數後，對預先訓練好存在系統中的 64 維通用背景模型，進行最大事後機率調適為使用者的 64 維語者高斯混和模型儲存於系統中，並更新註冊名單以完成註冊。

假設我們要依觀察到的資料 x 去評估一未知的母體參數（unobserved population parameter）θ，而資料 x 的抽樣分布(sampling distribution)為 $f(\cdot)$ 且存在一 θ 的事前分布（prior distribution）$g(\cdot)$，則 θ 的事後分布可表示為：

$$f(\theta \mid x) = \frac{f(x \mid \theta) g(\theta)}{\int_{\theta' \in \Theta} f(x \mid \theta') g(\theta') d\theta'} \qquad (2\text{-}3)$$

Θ 為 $g(\cdot)$ 的定義域。而最大事後機率法則可表示為：

$$\hat{\theta}_{\text{MAP}}(x) = \underset{\theta}{\operatorname{argmax}} \frac{f(x \mid \theta) g(\theta)}{\int_{\theta' \in \Theta} f(x \mid \theta') g(\theta') d\theta'} = \underset{\theta}{\operatorname{argmax}}\, f(x \mid \theta) g(\theta) \qquad (2\text{-}4)$$

而在本論文中最大事後機率調適法則的做法是首先計算註冊語料對於通用背景模型中每一個高斯分布的充份統計量（sufficient statistics），接著用新的充份統計量及新的高斯混和模型參數去更新通用背景模型裡包含舊的充份統計量資訊的各項參數，更新的方法為以一定的比例混合新的參數及舊的參數，比例則是由一資料相關的係數 α 決定。α 與新的充分統計量成正比，代表每一個高斯分布新的充份統計量的可信度，如果充分統計量越高，則越多的註冊語料落在這一個高斯

分布，可信度越高，新參數占得比例也越重。

具體的調適過程如下：

$D-$dimensional feature vector X ， $D\times 1$ feature vector μ_i ，

$D\times D$ covariance matrix Σ_i ，因此高斯分布機率為

$$p_i(X)=\frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}(X-\mu_i)'(\Sigma_i)^{-1}(X-\mu_i)\right\} \quad (2\text{-}5)$$

而高斯混和模型和通用背景模型皆為 M mixtures ， feature vector $X=\{x_1,...,x_T\}$ ，

$$\text{我們得到 } \Pr(i\mid x_t)=\frac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)} \quad (2\text{-}6)$$

接著利用 $\Pr(i\mid x_t)$ 和 x_t 得到每一個高斯分布的充分計算量跟新的平均值（mean）及方差（variance）：

$$n_i=\sum_{t=1}^{T}\Pr(i\mid x_t) \quad (2\text{-}7)$$

$$E_i(x)=\frac{1}{n_i}\sum_{t=1}^{T}\Pr(i\mid x_t)\cdot x_t \quad (2\text{-}7)$$

$$E_i(x^2)=\frac{1}{n_i}\sum_{t=1}^{T}\Pr(i\mid x_t)\cdot x_t^{\,2} \quad (2\text{-}9)$$

為了控制新參數語就參數的平衡，我們利用充份統計量及比例因子 γ 來計算調適系數 α_i^w（用於權重）、α_i^m（用於平均值）、α_i^v（用於方差）。因為在【2】中提到三個調適系數找出各自的最佳值與共用一個值對結果影響的程度微乎其微，所以本系統將三個調適系數設為一致，且比例因子 γ 設為經驗法則所得到的常數 16。

$$\alpha_i^w=\alpha_i^m=\alpha_i^v=\alpha_i=\frac{n_i}{n_i+\gamma} \quad (2\text{-}10)$$

得到調適系數後，根據下列算式，我們就可以得到新的高斯混和模型參數：

$$w_i^{new}=\left[\frac{\alpha_i n_i}{T}+(1-\alpha_i)w_i\right] \quad (2\text{-}11)$$

$$\mu_i^{new}=\alpha_i E_i(x)+(1-\alpha_i)\mu_i \quad (2\text{-}12)$$

$$\left(\sigma_i^{new}\right)^2 = \alpha_i E_i\left(x^2\right) + \left(1-\alpha_i\right)\left(\sigma_i^2 + \mu_i^2\right) - \left(\mu_i^{new}\right)^2 \qquad (2\text{-}13)$$

然後因為我們假設高斯混和模型的每一個高斯分布之間是獨立的,所以他們的協方差矩陣（covariance matrix）是一對角矩陣,所以展開後是一連串相乘,而取對數之後是連加：

$$\det = \frac{1}{2}\sum_{d=1}^{26}\log\left(\sigma_d^2\right) \qquad (2\text{-}14)$$

而這些新的參數則組成新註冊語者的高斯混和模型。

2.1.4 語者辨識

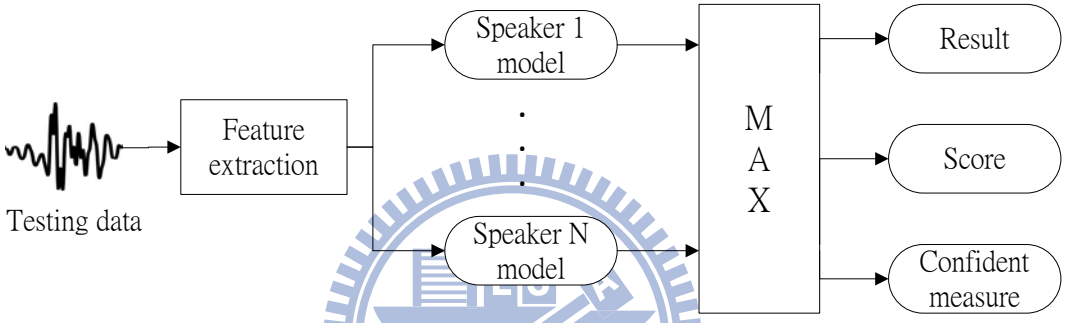


圖 2.3：語者辨識方塊圖

如圖 2.3 所示，辨識時使用者用系統麥克風錄下任意秒數的測試語料。語料經過特徵參數抽取為 26 維 RASTA 參數後，對存在系統中的每個語者模型計算概似值（likelihood），並找出最大概似值（maximum likelihood）的語者模型當作辨識結果。系統輸出辨識結果排名，最大概似值分數以及信心指數，而信心指數的算法將在第三章詳細介紹。

2.2 整合麥克風陣列

考慮系統實際應用的環境中，必然會受到環境雜訊的干擾，因此我們嘗試將系統前端輸入由原本的單一麥克風改為由四支麥克風組成的線性麥克風陣列，並藉由麥克風陣列所提供的空間資訊加入波束形成技術，使得系統接收到經過純化處理的聲音訊號，以期能降低環境雜訊所造成之影響。本節將介紹系統所結合的波束形成和語音端點偵測（voice activity detection, VAD）技術及其應用於受雜訊干擾的語者識別的實驗結果。

2.2.1 波束形成

麥克風陣列常使用波束形成的方法為廣義旁瓣消除器（generalized sidelobe canceller），如圖 2.4 所示此方法為在獲得抵達方向偵測（direction of arrival, DOA）提供的所需訊號的方向資訊後，藉由將麥克風接收訊號中的干擾部分分離出來，並應用可適性權重向量將分離出來的干擾訊號與原始麥克風訊號相消，以達到聲音純化的目的。而在【12】裡則提出，傳統廣義旁瓣消除器是在假設線性陣列模型為裡想均勻的前提，即陣列訊號環境滿足窄頻訊號及遠場平面波的假設下，所推導出來的最佳解，但在實際的語音訊號環境中，使用者大多離麥克風很近且訊號多為球面波，並非如假設般單純，因此必須藉由估計聲源於空間中所經過的轉移函數來調適廣義旁瓣消除器，以適應實際的語音訊號環境。但在實際應用中，空間轉移函數很難得到，因此本系統採用在【12】中藉由估計較容易得到的空間轉移函數比值來調適空間濾波器的方法。

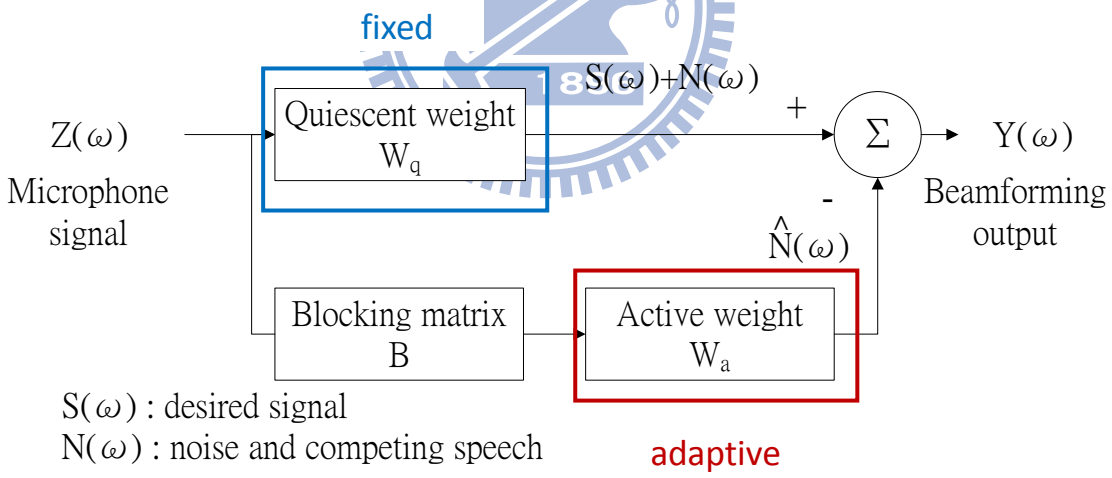


圖 2.4：廣義旁瓣消除器

2.2.2 語音端點偵測

傳統的語音端點偵測作法為利用訊號短時間能量最大值與最小值之間差距來訂定動態的偵測門檻值，並根據此值判定語音端點，但傳統方法在低訊雜比（signal-to-noise ratio, SNR）或是常有突發性雜訊的環境，性能會大受影響，比

方說在家用環境中，除了目標使用者之外還有其他家庭成員不時說話聲的干擾，則傳統方式易將家庭成員的聲音也當作我們想要的語音訊號。此外一般的語音端點偵測的方法，並非使用麥克風陣列做為前端輸入，所以無法擁有空間上的資訊，而空間資訊的好處在於可以更有效的消除雜訊的影響，以及在頻譜與時域之外再加入空間亂度的資訊以幫助語音端點偵測。因此在【13】裡提出了基於廣義旁瓣消除器的麥克風陣列處理結構，而發展出的空間語音端點偵測方法（spatial voice activity detection, SVAD），藉由估算目標對干擾比（target-to-jammer ratio, TJR）做為語音端點偵測的門檻值。

由圖 2.4 可看出廣義旁瓣消除器的結構分為上下兩個分支，而上分支包含想要訊號的能量及干擾的人聲或雜訊的能量，下分支為分離出來的干擾人聲及雜訊的能量及非常少量或是幾乎沒有的想要訊號能量，因此我們可以將想要目標訊號的短時間能量近似為（$\{\cdot\}$ 表示為短時間平均）：

$$P_t \approx \left\{ \left\| \mathrm{Y}(\omega) \right\|^2 \right\} \qquad (2\text{-}15)$$

而干擾訊號的短時間能量近似為：

$$P_j \approx \left\{ \left\| \hat{\mathrm{N}}(\omega) \right\|^2 \right\} \qquad (2\text{-}16)$$

因此可得到目標對干擾比：

$$\mathrm{TJR} = 10\log_{10} P_t - 10\log_{10} P_j \qquad (2\text{-}17)$$

並以此值做為語音端點偵測門檻值。利用此方法的好處在於，其所切割出的語音訊號區段，是目標訊號能量大於干擾訊號能量的區段，而這些語音訊號區段對於辨識而言是較可靠的區段。根據【13】裡的實驗結果可發現，此方法即使在訊雜比-10dB 的情況下依然可以有相當準確的語音端點偵測，因此可適用於嘈雜且含有非穩態干擾（non-stationary）雜訊的環境。

2.3 實驗結果

2.3.1 語者識別實驗語料及結果

本論文中使用 TCC-300 麥克風語音資料庫是由國立交通大學、國立成功大學、國立台灣大學所共同錄製，中華民國計算語言學學會所發行，此語料庫屬於麥克風朗讀語音，主要目的是為提供語音辨認研究，檔案統計資料如表 2.1 所示。台灣大學語料庫主要包含詞以及短句，文字經過設計，考慮音節與其相連出現之機率，共 100 人，每人錄製一句而成；成功大學及交通大學為長文語料，其語句內容由中研院提供之 500 萬詞詞類標示語料庫中選取，每篇文章包含數百個字，再切割成 3 至 4 段，每段至多 231 字，分別各 100 人，每人錄製一句朗讀來錄製，且每人所朗讀之文章皆不相同。每個學校之語句取樣頻率皆為 16000 赫茲，取樣位元數為 16 位元。音檔檔頭為 4096 位元組（byte），副檔名為*.vat。

表 2.1：TCC300 語料資料統計表

學校名稱	文章屬性	語者總數		總音節數		檔案總數	
台灣大學	短文	男	50	男	27541	男	3425
		女	50	女	24677	女	3084
		總數	100	總數	52218	總數	6590
交通大學	長文	男	50	男	75059	男	622
		女	50	女	73555	女	616
		總數	100	總數	148614	總數	1238
成功大學	長文	男	50	男	63127	男	588
		女	50	女	68749	女	582
		總數	100	總數	131876	總數	1170

本實驗從語料庫中的交通大學及成功大學的部分隨機挑選 90 位語者男女各半，並依其切割位置檔案去除音檔裡大於 0.5 秒的短停頓（short pulse）及靜音（silence），接著將處理過的音檔轉為取樣頻率為 8000 赫茲，取樣位元數為 16

位元的無檔頭 PCM 音檔供實驗之用。接著將 90 位語者分為 9 組，10 男、10 女各 4 組及 5 男 5 女 1 組。每位語者各有 50 筆 1~6 秒測試語料，和一組 5~40 秒的註冊語料。實驗時以各組為單位，做 10 人的語者識別，實驗結果如下表 2.2 及下圖 2.5 所示。圖 2.5 的橫軸為測試秒數，縱軸為辨識率，每條曲線代表不同的註冊秒數。

表 2.2：語者辨識基礎系統辨識率

註冊秒數 / 測試秒數	5 秒	10 秒	20 秒	30 秒	40 秒
1 秒	55.96%	67.51%	75.38%	79.07%	80.08%
1.5 秒	65.76%	76.78%	83.62%	86.71%	89.00%
2 秒	72.13%	82.98%	88.51%	90.87%	93.07%
3 秒	79.07%	89.47%	93.22%	94.82%	95.69%
4 秒	84.56%	91.80%	95.24%	96.36%	97.40%
5 秒	87.18%	93.64%	96.40%	97.22%	98.04%
6 秒	88.84%	95.02%	97.07%	97.76%	98.67%

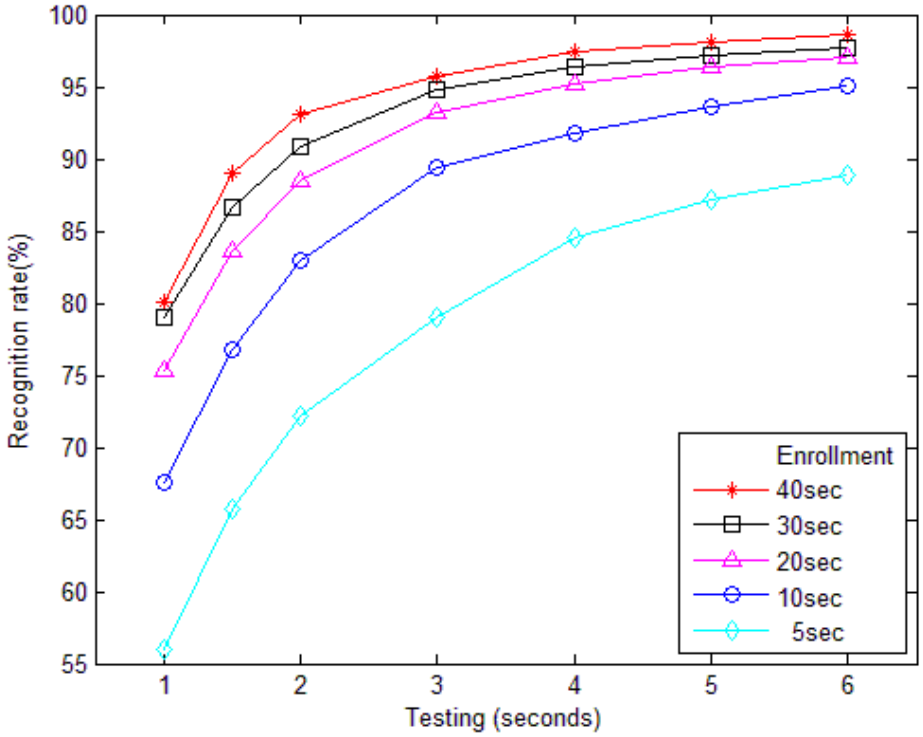


圖 2.5：語者辨識基礎系統辨識率

由表 2.2 及圖 2.5 可以發現註冊秒數在超過 20 秒之後對辨識率所造成的影響趨緩，而測試秒數在註冊秒數較長時大約在 3 秒之後對辨識率影響趨緩，但在註冊秒數較短時辨識率隨著大於 3 秒的測試秒數上升較為明顯。由實驗結果可以發現本論文基礎系統最有效率的操作點大約是註冊秒數 20 多秒、測試秒數 3 秒左右。另一方面即使註冊秒數不長，只要測試秒數夠長依然可以有八成以上的辨識率，對於在即時系統上的應用而言，使用者一開始並不需要太多時間註冊，而在使用時多說幾次即可有不錯的辨識率。但若是在對反應時間有較高要求的應用裡，增加註冊秒數可以達到一定效果，但還是無法大幅提升辨識率，因此對於這些應用可能必須使用文本相關或其他方式來增加其強健性。

2.3.2 整合麥克風陣列實驗語料與結果

麥克風陣列的擺放方式如圖 2.6 所示，四顆數位麥克風彼此相距 7 公分排成一線性陣列，而聲源方向為陣列正前方 30 公分處。麥克風收錄的訊號為取樣頻率 8000 赫茲，取樣位元數為 16 位元的訊號。

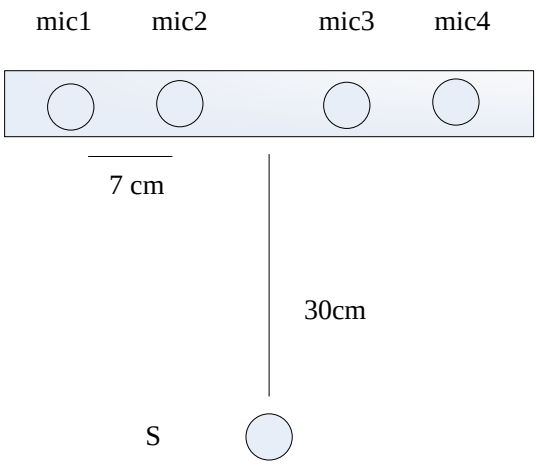


圖 2.6：麥克風陣列位置圖

原始語料由 9 位男性語者各念"阿凡達"（機器人的初始名字）5 次。因為語料稀少，所以採用交叉測試的方式，每位語者依序由五個音檔中挑選四個音檔串聯在一起做為註冊語料，剩下的那一個音檔則做為那組註冊語料的測試語料。而我們將註冊語料經過處理後分為 4 大類，因此每個分類的語料會有 5 組註冊語料，4 類分別為：1.乾淨且經過波束形成處理但未經過語音端點偵測處理的語料；2.乾淨且經過波束形成及語音端點偵測除去靜音的語料；3.混人聲（babble）雜訊、訊雜比為 15dB 經過波束形成及語音端點偵測除去靜音的語料；4. 混飛機聲（F16）雜訊、訊雜比為 15dB 經過波束形成及語音端點偵測除去靜音的語料。因為經過語音端點偵測除去靜音的語料長度幾乎變為原本的一半，而最大事後機率調適法則會依據調適語料的長短調整新舊資料的權重，所以我們將有除去靜音的註冊語料複製為原本兩倍長度以示公平。

測試語料則分為 6 大類，每類再分為 0dB、5dB、10dB、15dB 等 4 種訊雜比語料，每種語料有 5 組共 45 筆測試語料。6 類分別為：1.混人聲雜訊但未經過波束形成及語音端點偵測處理的第二支麥克風語料，代號 Babble；2.混人聲雜訊經過波束形成但未經過語音端點偵測處理的語料，代號 Babble+BF；3.混人聲雜訊經過波束形成且經過語音端點偵測除去靜音的語料，代號 Babble+BF+VAD；4.混飛機聲雜訊但未經過波束形成及語音端點偵測處理的第二支麥克風語料，代號 F16；5.混飛機聲雜訊經過波束形成但未經過語音端點偵測處理的語料，代號

F16+BF；6.混飛機聲雜訊經過波束形成且經過語音端點偵測除去靜音的語料，代號 F16+BF+VAD。

而實驗共分為四個階段，第一階段為利用乾淨註冊語料及乾淨測試語料做語者辨識，因為註冊與測試的音檔內容皆為"阿凡達"，所以實驗結果的辨識率為 100%；第二階段為利用乾淨且經過波束形成處理但未經過語音端點偵測處理的註冊語料，及六類測試語料做語者辨識，下表為實驗的結果。

表 2.3：乾淨且經過波束形成註冊語料辨識率

NOISE / SNR	Babble	Babble +BF	Babble +BF +VAD	F16	F16+BF	F16+BF +VAD
0dB	22.2%	31.1%	35.6%	20.0%	33.3%	33.3%
5dB	20.0%	33.3%	75.6%	24.4%	35.6%	68.9%
10dB	31.1%	35.6%	91.1%	26.7%	44.4%	88.9%
15dB	35.6%	42.2%	97.8%	40.0%	55.6%	100.0%

由表 2.3 的結果可以看出波束形成是有些微的效果，由其是在訊雜比在低時較為明顯，而語音端點偵測除去靜音的效果則較為顯著，在 0dB 時雖然提升辨識率的幅度較小，但在 5dB 之後幾乎都提升了一倍的辨識率，這結果可以看出波束形成對於語者辨識系統的幫助有限，但使用麥克風陣列所提供之空間資訊去做語音端點偵測在訊雜比不高的情況下依然能對辨識率有顯著的幫助。接著第三階段我們探討利用乾淨且經過波束形成及語音端點偵測處理除去靜音的註冊語料，及六類測試語料做語者辨識的結果，以期能更深入了解空間語音端點偵測對於語者辨識的影響。

表 2.4：乾淨經過波束形成及語音端點偵測處理之註冊語料辨識率

NOISE / SNR	Babble	Babble +BF	Babble +BF +VAD	F16	F16+BF	F16+BF +VAD
0dB	17.8%	17.8%	46.7%	13.3%	13.3%	40.0%
5dB	31.1%	24.4%	95.6%	15.6%	15.6%	77.8%
10dB	44.4%	31.1%	100.0%	15.6%	22.2%	95.6%
15dB	53.3%	44.4%	100.0%	26.7%	28.9%	100.0%

由表 2.4 的結果可以看出靜音含量較少的註冊語料，對於同樣是靜音含量較少的測試語料有著較好的辨識率，但對於含有靜音的測試語料則會更容易受到不同的雜訊的影響而有差異度相當大的辨識率。因此由結果可以推測，如果系統的語音端點偵測做得非常準確，即使訊雜比只有 5dB 依然能有不錯的辨識，但如果測試語料含有雜訊及較多靜音時，幾乎不含靜音的註冊語料強健性較差，且在不同雜訊間的辨識率變化相當大。最後第四階段我們探討利用與測試語料混相同雜訊且訊雜比為 15dB、經過波束形成及語音端點偵測處理之註冊語料，及六類測試語料做語者辨識的結果，以期能了解環境雜訊匹配與否對語者辨識的影響。

表 2.5：環境雜訊匹配辨識率

NOISE / SNR	Babble	Babble +BF	Babble +BF +VAD	F16	F16+BF	F16+BF +VAD
0dB	17.8%	20.0%	28.9%	11.1%	17.8%	33.3%
5dB	20.0%	20.0%	84.4%	17.8%	20.0%	93.3%
10dB	22.2%	22.2%	100.0%	22.2%	28.9%	100.0%
15dB	28.9%	24.4%	100.0%	26.7%	33.3%	100.0%

由表 2.5 可以看出不論使用混人聲雜訊或是飛機聲雜訊之註冊語料，對於相

對應雜訊且未經過語音端點偵測除去靜音之測試語料的辨識率，明顯不如乾淨的註冊語料，而對於相對應雜訊且經過語音端點偵測去除靜音之測試語料辨識率，則隨著雜訊不同與乾淨的註冊語料互有高低。由這三個實驗結果，我們可以得到在雜訊環境下使用麥克風陣列做為前級之語者辨識系統，空間語音端點偵測有無是影響其辨識率關鍵的結論，且有麥克風陣列提供空間之資訊，及使在訊雜比只有 5dB 的嘈雜環境裡，使用空間語音端點偵測對語料做處理後，依然可以有不錯的語者辨識率。

第三章 文本獨立語者確認系統

完整的語者辨識系統除了包含對於已註冊語者辨別身分的功能之外，還必須包含偵查是否為非註冊語者的能力。更進一步地，語者辨識系統與其他使用者身分識別系統（例：人臉辨識系統）結合時，還必須賦予辨識結果一個量化的分數，也就是所謂的信心指數，以利多個系統融合出一個共同的辨識結果。3.1 節將介紹系統計算門檻及信心指數的方法；3.2 節記錄實驗結果。

3.1 語者身分驗證與信心指數

在本論文中如圖 3.1 所示，我們將語者身分驗證問題分成兩個階段：

圖 3.1：語者確認系統

第一階段，系統依據辨識結果判定此次辨識是否有效。必須被判定為無效的情況包括非註冊語者使用系統，以及辨識錯誤的結果。判定無效的結果則跳過第二階段，直接輸出無法辨識使用者身分。判定有效的結果，則在第二階段賦予每個排名一個信心指數，信心指數為一個介於 0 到 1 之間的數值，代表系統對於這辨識結果的信心程度。

而我們將這兩階段視為統計假設檢定，H_0 代表已註冊語者使用系統，或是辨識正確的假設，而 H_1 則代表其對立假設。在已知 H_0、 H_1 的機率分布條件下，則根據奈曼-皮爾生引理，最佳解為概似比檢測。而其決策規則（decision rule）為 ：

$$\frac{p(X \mid \lambda_{hyp})}{p(X \mid \lambda_{\overline{hyp}})} \begin{cases} > \eta, X \in H_0 \\ \leq \eta, X \in H_1 \end{cases} \quad (3\text{-}1)$$

其對數形式為：

$$\Lambda(X) = \log p(X \mid \lambda_{hyp}) - \log p(X \mid \overline{\lambda}_{hyp}) \quad (3\text{-}2)$$

註冊語者為已知，因此我們可以藉由註冊語料得到 H_0 的機率分布，但如何去估計 H_1 就成了重點。本節我們將介紹三種估計 H_1 的方法及其實驗的結果，3.1.1 介紹使用通用背景模型做為 H_1，稱為通用模型正規化；3.1.2 介紹以第二名的語者做為同儕集合去估計 H_1，稱為最大值正規化（maximum normalization）；3.1.3 介紹以除了第一名以外所有註冊語者做為同儕集合去估計 H_1，稱為幾合平均數正規化（geometric mean normalization）。而我們先分別利用這三種方法及實驗語料得到辨識正確與辨識錯誤語料的分數 $\Lambda(X)$，抑或是目標語者與非目標語者的分數 $\Lambda(X)$，並藉由分析分數分布而得到最佳的門檻值以及賦予辨識結果信心指數的函數以供即時系統使用。

3.1.1 通用模型正規化

對於本論文的即時系統而言，使用者使用時，系統並無任何的使用者資訊，完全憑其語料及系統內事先註冊的模型來判斷每位語者的身分，因此首先必須判定使用者是否為已註冊語者，接著判定辨識結果是否為正確，如果為錯誤的辨識結果，其實也可以視為封閉集合的冒名頂替者。而通用背景模型正規化的概念，就是認為辨識正確的語料對於目標語者模型及通用背景模型的相似度差異，是比辨識錯誤的語料和非註冊語者的語料對於目標語者模型及通用背景模型的相似度差異來得較為明顯。

具體的正規化流程如下，假設現在共有 N 位已註冊語者，且第 M 位註冊語者為目標語者，而語者模型為 λ_n，$n = 1 \cdots S \cdots N$；目標語者模型為 λ_S；通用背景模型為 λ_{UBM}；觀測資料 $O = \{O_1, ..., O_t, ..., O_T\}$ 共 T 個音框，則我們可以得到每筆測

試語料對語者模型 λ_n ，經過通用模型正規化後的對數概似值分數為：

$$\Lambda_n(O)=\frac{\sum_{t=1}^{T}\left(\log p(O_t\mid\lambda_n)-\log p(O_t\mid\lambda_{UBM})\right)}{T} \quad (3\text{-}3)$$

測試的實驗語料分成三類，第一類為目標語者的語料 $O^S=\{O_1^S,...,O_t^S,...,O_T^S\}$ ；第二類為封閉集合的冒名頂替者語料 $O^n=\{O_1^n,...,O_t^n,...,O_T^n\}$ ， $n=1\cdots N\cap n\neq S$ ；第三類為開放集合的冒名頂替者語料（假設現在共有 K 位開放集合的冒名頂替者） $O^k=\{O_1^k,...,O_t^k,...,O_T^k\}$ ， $k=N+1\cdots N+K$ 。則我們可以得到三類分數，第一類為目標語者語料對於目標語者模型得到的分數：

$$\Lambda_S^S(O^S)=\frac{\sum_{t=1}^{T}\left(\log p(O_t^S\mid\lambda_M)-\log p(O_t^S\mid\lambda_{UBM})\right)}{T} \quad (3\text{-}4)$$

第二類為封閉集合的冒名頂替者語料對於目標語者模型得到的分數：

$$\Lambda_S^n(O^n)=\frac{\sum_{t=1}^{T}\left(\log p(O_t^n\mid\lambda_S)-\log p(O_t^n\mid\lambda_{UBM})\right)}{T},\ n=1\cdots N\cap n\neq S \quad (3\text{-}5)$$

第三類為開放集合的冒名頂替者語料對於目標語者模型得到的分數：

$$\Lambda_S^k(O^k)=\frac{\sum_{t=1}^{T}\left(\log p(O_t^k\mid\lambda_S)-\log p(O_t^k\mid\lambda_{UBM})\right)}{T},\ k=N+1\cdots N+K \quad (3\text{-}6)$$

而根據第一類分數是應該被判定有效，二三類分數是應該被判定為無效的原則，我們可以得到接收者操作特徵曲線圖，並藉由分析圖及找出其等錯率點來訂定門檻值。

此外，更進一步地，除了冒名頂替者是該被判定無效的辨識結果外，辨識錯誤的結果一樣是應該被判定為無效。因此我們將目標語者的語料對於目標語者模型的分數，再分成辨識正確的分數：

$$\Lambda_S^{SR}(O^{SR})=\frac{\sum_{t=1}^{T}\left(\log p(O_t^{SR}\mid\lambda_S)-\log p(O_t^{SR}\mid\lambda_{UBM})\right)}{T} \quad (3\text{-}7)$$

及辨識錯誤的分數：

$$\Lambda_S^{SE}\left(O^{SE}\right)=\frac{\sum_{t=1}^{T}\left(\log p\left(O_t^{SE}\mid\lambda_S\right)-\log p\left(O_t^{SE}\mid\lambda_{UBM}\right)\right)}{T} \qquad (3\text{-}8)$$

兩類，並依此原則訂定最終供即時系統使用的信心指數及門檻值。而辨識錯誤即代表系統將使用者誤認為其他已註冊語者，因此對於被誤認語者而言，此辨識錯誤也可以視為封閉集合的冒名頂替者。

3.1.2 最大值正規化

最大值正規化的概念，即是認為辨識正確的語料對於目標語者模型的相似度會大於辨識錯誤的語料和非註冊語者的語料對於目標語者模型的相似度。因此若辨識結果第一名與第二名的分數差距越大，系統對於此結果為辨識正確結果的信心度也越高。最大值正規化法也是討論第一名分數與其他名次分數分布的關係，因此其出現封閉集合的冒名頂者的情況也即為辨識錯誤的情況。

具體的正規化流程如下，假設共有 N 位已註冊語者，則語者模型為 λ_n，$n=1\cdots S\cdots N$；觀測資料 $O=\{O_1,...,O_t,...,O_T\}$ 共 T 個音框，我們可以得到測試語料對於語者模型 λ_n 的對數概似值分數：

$$\Lambda_n\left(O\right)=\frac{\sum_{t=1}^{T}\left(\log p\left(O_t\mid\lambda_n\right)\right)}{T} \qquad (3\text{-}9)$$

首先找出最高的分數：

$$M=\arg\max_{n}\Lambda_n\left(O\right) \qquad (3\text{-}10)$$

$$\Lambda_{1st}\left(O\right)=\Lambda_M\left(O\right) \qquad (3\text{-}11)$$

接著找出第二高的分數：

$$\Lambda_{2nd}\left(O\right)=\max_{n=1\cdots N\cap n\neq M}\left[\Lambda_n\left(O\right)\right] \qquad (3\text{-}12)$$

最後我們可以得到經過最大值正規化的分數：

$$\Lambda_{MAX}\left(O\right)=\Lambda_{1st}\left(O\right)-\Lambda_{2nd}\left(O\right) \qquad (3\text{-}13)$$

而在這裡我們將語料分為兩類，第一類為已註冊語者語料

$$O^n = \left\{ O_1^n, ..., O_t^n, ..., O_T^n \right\} \text{，} n = 1 \cdots N$$

第二類為開放集合的冒名頂替者的語料（假設我們現在有 K 位開放集合冒名頂替者）

$$O^k = \left\{ O_1^k, ..., O_t^k, ..., O_T^k \right\} \text{，} k = N+1 \cdots N+K$$

因此可以得到三類分數，第一類為辨識正確的分數：

$$\Lambda_{MAX}^{R}\left(O^{nR}\right) = \Lambda_{1st}\left(O^{nR}\right) - \Lambda_{2nd}\left(O^{nR}\right) \qquad (3\text{-}14)$$

第二類為辨識錯誤（對於被誤認的語者，則視為封閉集合的冒名頂替者）的分數：

$$\Lambda_{MAX}^{E}\left(O^{nE}\right) = \Lambda_{1st}\left(O^{nE}\right) - \Lambda_{2nd}\left(O^{nE}\right) \qquad (3\text{-}15)$$

第三類為開放集合的冒名頂替者的分數：

$$\Lambda_{MAX}^{I}\left(O^{k}\right) = \Lambda_{1st}\left(O^{k}\right) - \Lambda_{2nd}\left(O^{k}\right) \qquad (3\text{-}16)$$

也是根據第一類分數是應該被判定有效，二三類分數是應該被判定為無效的原則，並藉由分析接收者操作特徵曲線圖及找出其等錯率點來訂定門檻值。

3.1.3 幾何平均數正規化

幾何平均數正規化的概念，即是認為辨識正確的語料對於目標語者模型及其他已註冊語者模型的相似度差異，是比辨識錯誤的語料和非註冊語者的語料對於目標語者模型及其他已註冊語者模型的相似度差異來得較為明顯。因此若第一名的分數大於其他名次分數的平均越多，系統對於此結果為辨識正確結果的信心度也越高。此外由於此正規化法討論的是第一名分數與其他名次分數分布的關係，因此其出現封閉集合的冒名頂者的情況即為辨識錯誤的情況，例：已有註冊語者模型的使用者 A 使用系統時，被系統辨識為註冊語者 B，因此對於 A 而言此辨識結果為錯誤的，而 A 同時也成為了冒名頂替 B 的使用者。

具體的正規化流程如下，假設共有 N 位已註冊語者，語者模型為 λ_n，$n=1\cdots S\cdots N$；觀測資料 $O=\{O_1,...,O_t,...,O_T\}$ 共 T 個音框，我們可以得到測試語料對於語者模型 λ_n 的對數概似值分數：

$$\Lambda_n(O)=\frac{\sum_{t=1}^{T}\left(\log p(O_t\mid\lambda_n)\right)}{T} \qquad (3\text{-}17)$$

首先找出最高的分數：

$$M=\arg\max_n \Lambda_n(O) \qquad (3\text{-}18)$$

$$\Lambda_{1st}(O)=\Lambda_M(O) \qquad (3\text{-}19)$$

接著找出幾合平均數正規化項：

$$\Lambda_G(O)=\frac{\sum_{n=1\cap n\neq M}^{N}\Lambda_n(O)}{N-1} \qquad (3\text{-}20)$$

最後我們可以得到經過幾合平均數正規化的分數：

$$\Lambda_{GM}(O)=\Lambda_{1st}(O)-\Lambda_G(O) \qquad (3\text{-}21)$$

而在這裡我們將語料分為兩類，第一類為已註冊語者語料

$$O^n=\{O_1^n,...,O_t^n,...,O_T^n\}\ ,\ n=1\cdots N \qquad (3\text{-}22)$$

第二類為開放集合的冒名頂替者的語料（假設我們現在有 K 位開放集合冒名頂替者）

$$O^k=\{O_1^k,...,O_t^k,...,O_T^k\}\ ,\ k=N+1\cdots N+K \qquad (3\text{-}23)$$

因此可以得到三類分數，第一類為辨識正確的分數：

$$\Lambda_{GM}^{R}(O^{nR})=\Lambda_{1st}(O^{nR})-\Lambda_G(O^{nR}) \qquad (3\text{-}24)$$

第二類為辨識錯誤（對於被誤認的語者，則視為封閉集合的冒名頂替者）的分數：

$$\Lambda_{GM}^{E}(O^{nE})=\Lambda_{1st}(O^{nE})-\Lambda_G(O^{nE}) \qquad (3\text{-}25)$$

第三類為開放集合的冒名頂替者的分數：

$$\Lambda_{GM}^{I}\left(O^{k}\right)=\Lambda_{1st}\left(O^{k}\right)-\Lambda_{G}\left(O^{k}\right) \qquad (3\text{-}26)$$

同樣根據第一類分數應該被判定有效，二三類分數應該被判定為無效的原則，並藉由分析接收者操作特徵曲線圖及找出其等錯率點來訂定門檻值。

3.1.4 信心指數曲線及門檻值

本論文中訂定第一階段門檻值的方式為，藉由每個正規化方法三種秒數的等錯率點的分數值做為門檻值。而當實際使用時，系統將測試語料區分為小於 1.5 秒、介於 1.5 秒到 2.5 秒、大於 2.5 秒三個區間，不同區間的測試語料分別對應到 1、2、3 秒測試結果等錯率點所算出的門檻值。

而第二階段賦予辨識結果信心指數的方式則如下圖所示，先利用大量實驗的結果得到辨識正確分數的累積分布函數 $F_R(x)$，以及辨識錯誤分數的累積分布函數 $F_E(x)$。而信心指數曲線為經過 $F_R(x)$-(1- $F_E(x)$)運算後得到的曲線，再將 Y 軸正規化到 0 到 1 之間。曲線圖的 X 軸代表經過正規化方法後的辨識結果分數，Y 軸代表信心指數（一個 0 到 1 的值）。實際使用時，每次辨識結果所得到的語者分數經過正規化後，依其落在 X 軸的點找相對應的 Y 軸的點即為其信心指數。

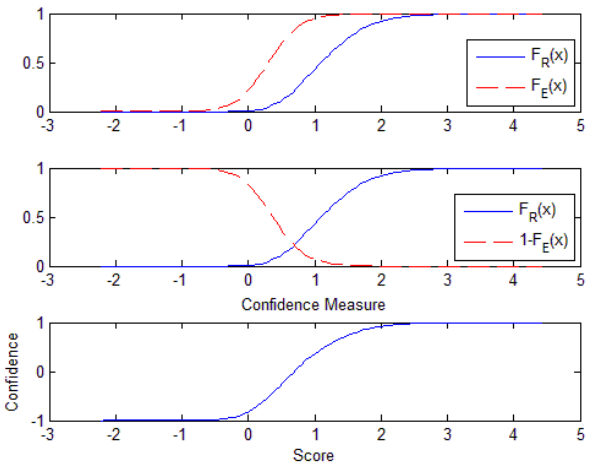


圖 3.2：信心指數算法

3.2 實驗結果

3.2.1 語者驗證實驗語料與結果

實驗語料為與語者識別實驗語料相同的 10 男、10 女各 4 組及 5 男 5 女 1 組共 9 組的語料。每位語者使用 24 秒註冊語料，及 50 筆 1 秒測試語料，每筆語料之間不重疊。

實驗分為兩個階段，第一階段先挑選其中 10 男、10 女各 1 組及 5 男 5 女 1 組共 3 組語料，在通用模型正規化法裡探討目標語者語料分數與冒名頂替者語料分數的分布，及更進一步地在三種正規化方法下探討辨識正確的分數與辨識錯誤（封閉集合的冒名頂替者）和開放集合冒名頂替者的分數分布。

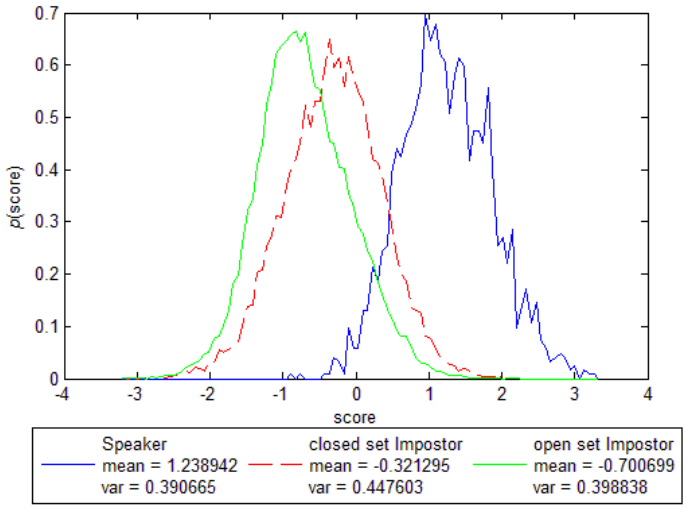


圖 3.3：通用模型正規化法目標語者與冒名頂替者分數分布

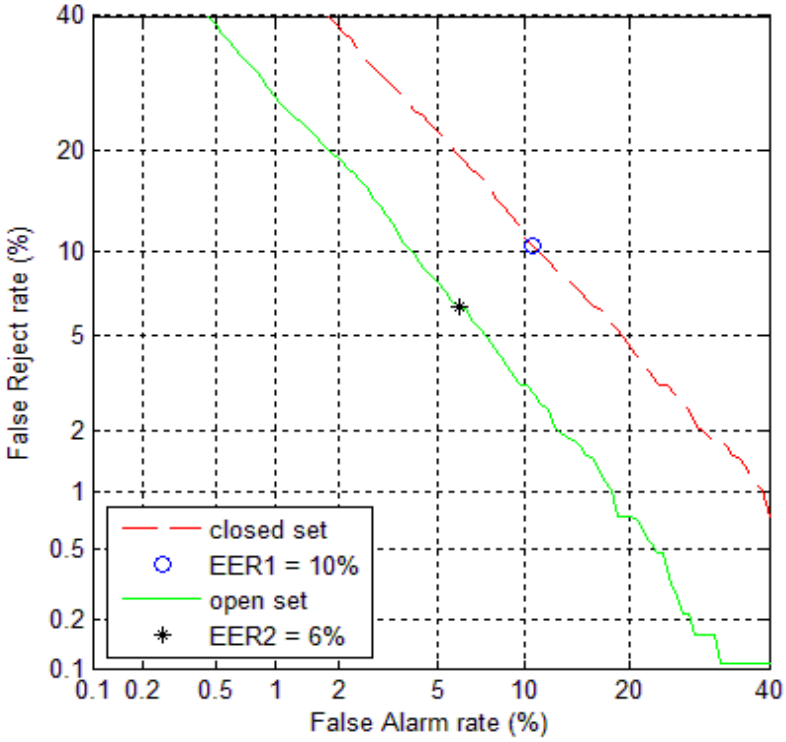
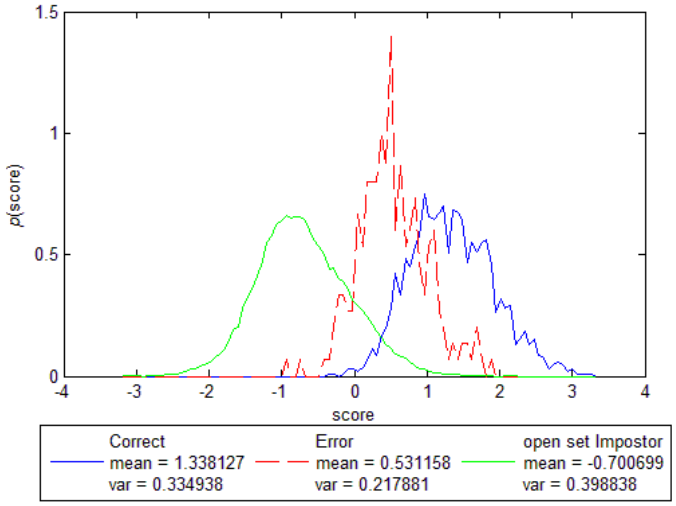


圖 3.4：通用模型正規化法接收者操作特徵曲線圖

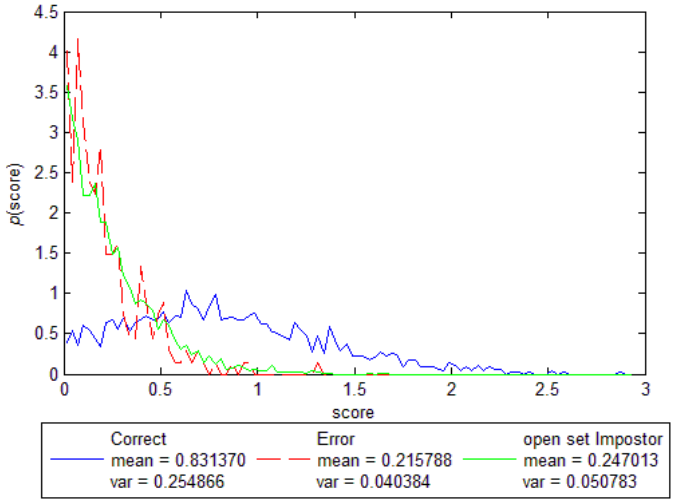
圖 3.3 為 3 組所有目標語者語料分數以及封閉集合和開放集合冒名頂替者語料分數經過通用模型正規化之後的分布圖。在我們系統的使用情境裡，因為文本獨立的關係，系統並不知道使用者說話的內容，而每位語者使用時也並無宣稱自己身分，因此在這個實驗裡所謂的封閉集合冒名頂替者，就定義為每次語者識別後目標語者測試語料對於除了自己以外其他所有已註冊語者模行得到的分數，也就是假設目標語者使用系統時宣稱自己為其他已註冊語者時的情況。由圖可以發現封閉集合的冒名頂替者分數分布較為靠近目標語者分數，據推測是因為在全部男生或女生的實驗組裡封閉集合的冒名頂替者為同樣性別，而開放集合的冒名頂替者則有 3/4 為不同性別，才造成此差異，而這也影響到了兩者的等錯率，分數分布距離較遠的開放集合冒名頂替者因其與目標語者分數分布重疊部分較少，所以有比較低的等錯率。

對於家用機器人上所使用之文本獨立語者辨識系統而言，由於較常用於生活幫手或是客制化服務的情境中且文本獨立的關係，因此系統並無任何關於使用者的資訊，只能單憑其測試語料來辨別身分，而對於系統而言真正會出現的冒名頂

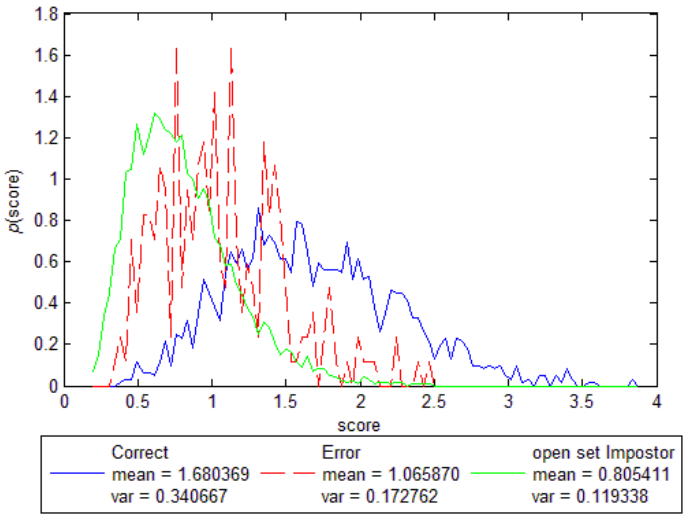
替者即為語者識別錯誤的結果及非註冊語者。接著我們將探討在三種正規化方法下辨識正確、辨識錯誤和開放集合冒名頂替者之間分數分布的情況，以期系統能偵測錯誤。而辨識正確的分數則是來自每筆目標語者測試語料對於其語者模型所得到的分數中辨識結果正確的部分，辨識錯誤的分數則是來自每筆目標語者測試語料對於其他已註冊語者模型所得到的分數中，超過該筆語料對於其自身語者模型所得到的分數，而造成辨識錯誤的部分。



（a）通用模型正規化法



（b）幾合平均數正規化法



（c）最大值正規化

圖 3.5：辨識正確及辨識錯誤分數分布

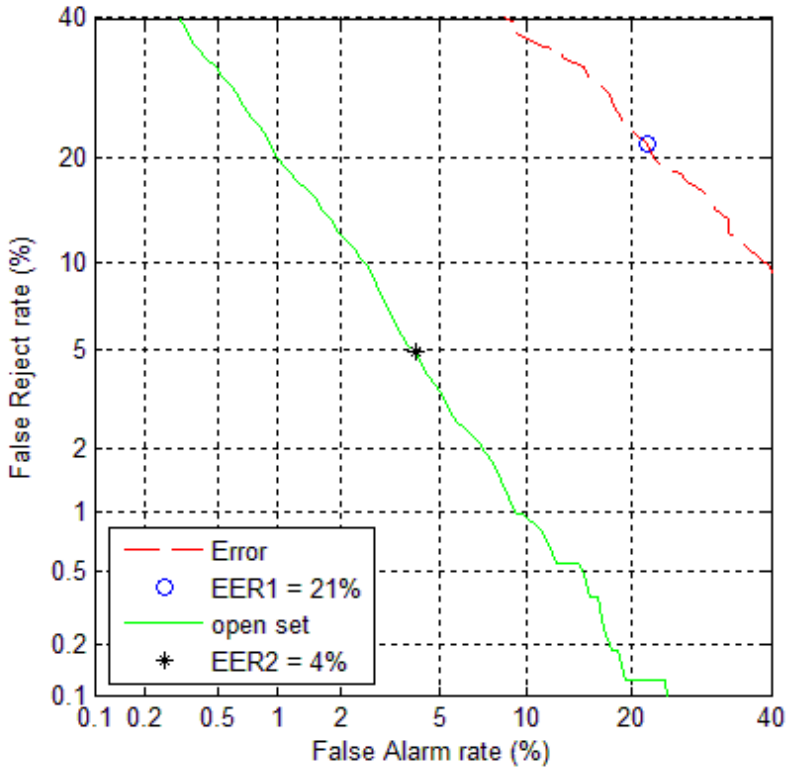


圖 3.6：通用模型正規化法接收者操作特徵曲線圖

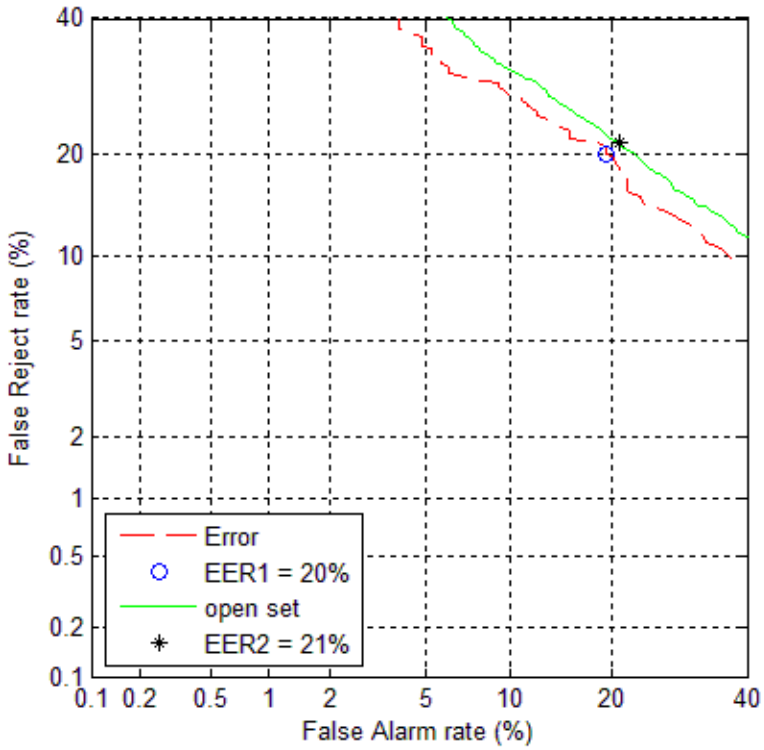


圖 3.7：最大值正規化法接收者操作特徵曲線圖

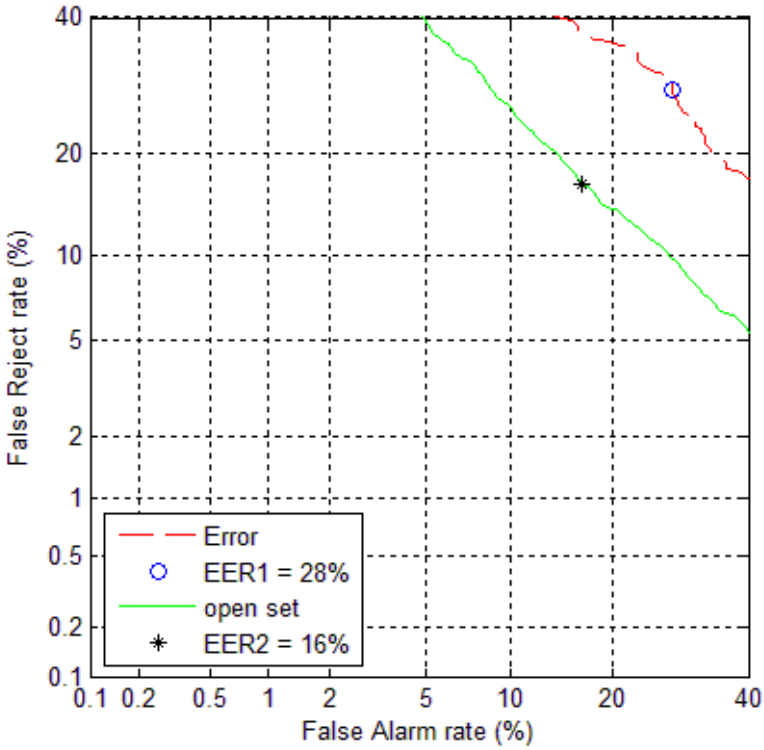


圖 3.8：幾合平均數正規化法接收者操作特徵曲線圖

由圖 3.5（a）可以看出辨識錯誤的分數確實與辨識正確的分數有所差別，但重疊的部分卻比之前封閉集合冒名頂替者分數與目標語者分數重疊部分大得

多，因此由圖 3.6 可看出等錯率提高了不少。而開放集合的冒名頂替者與辨識正確的分數分佈重疊的更少了，因此等錯率下降了些許。由這結果可以看出在通用模型正規化法裡若以辨識正確與否當做判斷準則，則依照等錯率點所找出的門檻值對於開放集合的冒名頂替者是相當有鑑別力的，而對於判斷是否辨識錯誤則有 20%左右的機率判斷錯誤。

由圖 3.7 及圖 3.8 實驗的結果可以看出幾何平均數正規化法對於判別辨識錯誤的能力似乎較最大值正規化法和通用模型正規化法差上許多，但其辨別非註冊語者的能力似乎略優於最大值正規化法，不過也還是遠輸於通用模型正規化法。此外，由圖 3.5（b）我們可以發現最大值正規化法辨識錯誤以及非註冊語者的分數分布非常相似，這也顯示了最大值正規化法的背後意義，在於如果第一名的分數超過第二名越多，則辨識結果為正確的可能性越大。

綜合第一階段實驗的結果，通用模型正規化法對於辨識錯誤的判斷能力與其他兩個方法相差不遠，而對於非註冊語者的判斷則優於其餘兩個方法。但非註冊語者的分數分布較難以掌握，因為其語料變化較大也未知，在實驗時也無法完全正確地去估計真正非註冊語者的模型，所以在真實情況時，是否能確實的拒絕未註冊使用者，可能會受到許多未能事先掌控的因素大幅度影響。而且對於家用系統的使用情境而言，大部分時間皆為已註冊語者使用系統，所以如何將辨識錯誤的結果判定無效比拒絕非註冊語者的情境更常見，因此在第二階段實驗裡我們將重點擺在三個方法的偵錯能力。此外在第一階段實驗確定這三個方法有一定的偵錯能力之後，為了印證對於結果的推論具有一般性，我們將測試組數增加為 9 組，並測試不同秒數所造成的結果變異。每位語者一樣使用 24 秒註冊語料，而測試語料則增加為 1、2、3 秒各 50 筆。

在第二階段實驗，我們首先觀察通用模型正規化法對於目標語者語料分數與封閉集合冒名頂替者語料分數的分布，接著觀察三種正規化方法辨識正確與辨識錯誤的分數分布。觀察通用模型正規化法對於目標語者語料分數與封閉集合冒名頂替者語料分數的分布，是為了藉由了解目標語者對於自己的語者模型及對其他

註冊名單裡的語者模型所得到的分數分布，做為之後分析辨識正確及辨識錯誤分數分布的參考。

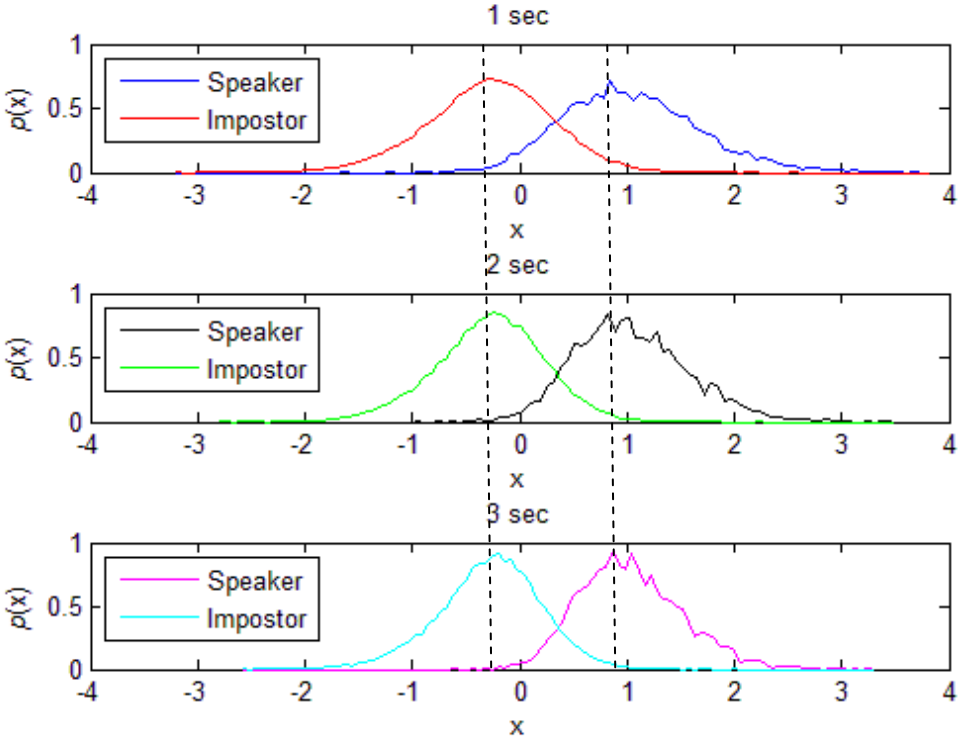


圖 3.9：通用模型正規化法 1、2、3 秒目標語者與封閉集合冒名頂替者分數分布

表 3.1：通用模型正規化法目標語者與封閉集合冒名頂替者分數分布統計

測試秒數 / 分數	1 秒	2 秒	3 秒
目標語者平均值	1.04	1.04	1.04
目標語者方差	0.42	0.29	0.24
冒名頂替者平均值	-0.28	-0.28	-0.28
冒名頂替者方差	0.35	0.26	0.23

由圖 3.9 及表 3.1 我們可以發現不論測試語料的秒數為多少，目標語者與封閉集合冒名頂替者分數分佈的平均值都差別不大，但方差則明顯的隨著秒數增加而變小。這結果代表了當測試秒數越高，則兩類分數的分布越往其各自的平均值

集中，而重疊造成混淆的部分越來越小，因此等錯率也應該越來越小。此外重疊部分在一定程度上與辨識錯誤率相關，因此隨著秒數上升辨識錯誤率下降，重疊部分減少也是合理的情況。由下圖 3.10 也可以看出等錯率確實如預期般隨著秒數上升而下降，但下降的幅度也隨著秒數趨緩，對照於語者識別率上升速度的結果，也確實隨著秒數上升而趨緩。

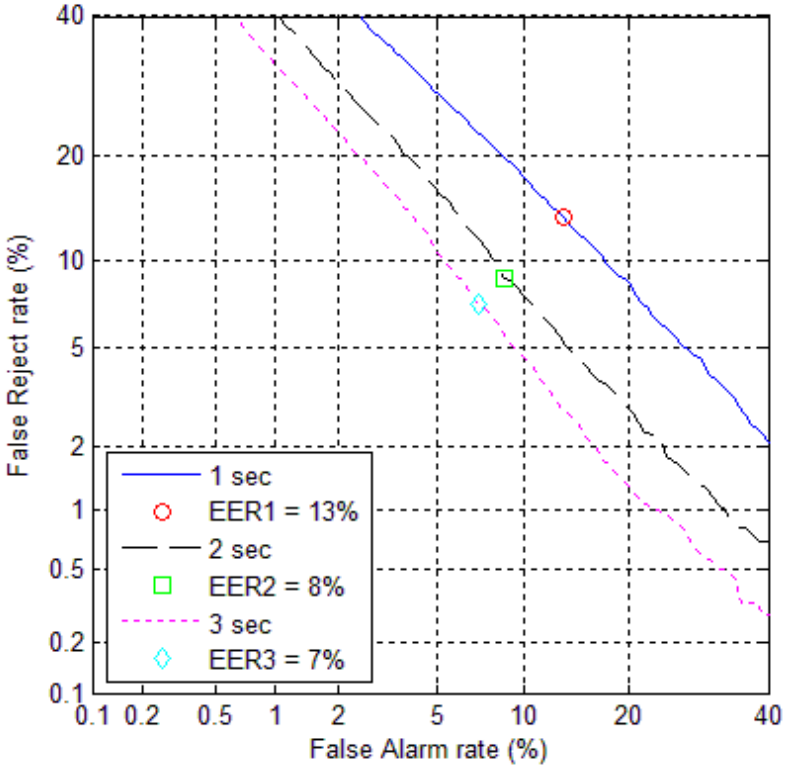


圖 3.10：通用模型正規化法 1、2、3 秒接收者操作特徵曲線圖

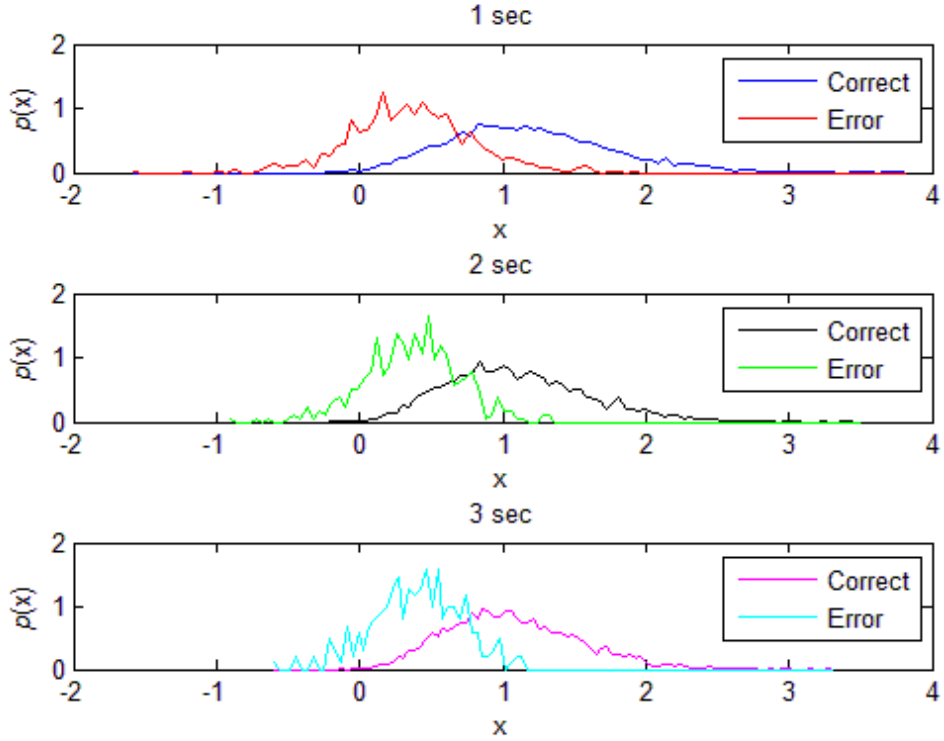


圖 3.11：通用模型正規化法 1、2、3 秒辨識正確及辨識錯誤的分數分布

表 3.2：通用模型正規化法辨識正確及辨識錯誤的分數分布統計

測試秒數 / 分數	1 秒	2 秒	3 秒
辨識正確平均值	1.20	1.11	1.08
辨識正確方差	0.34	0.25	0.22
辨識錯誤平均值	0.36	0.37	0.40
辨識錯誤方差	0.19	0.11	0.10

接著觀察通用模型正規化法下辨識正確與錯誤的分數分布，由圖 3.11 及表 3.2 可以看出從目標語者分數分佈中拆出部分的辨識正確分數，及從封閉集合冒名頂替者分數分布中拆出部分的辨識錯誤分數，也有著方差隨著秒數增加而變小的特性，但兩分數平均值的距離卻隨著秒數增加而縮短。據推測應該是因為會造成語者識別錯誤的語料，其分數必定是落於封閉集合冒名頂者分數分布與目標語

者分數分布重疊的區間，而隨著秒數上升重疊部分越來越小，所以依然辨識錯誤結果的分數也必然越來越高。另一方面由表 3.2 可看出隨著秒數上升，目標語者與封閉集合冒名頂替者各自分數的平均值幾乎不變，但目標語者分數中屬於辨識正確的部分則越來越多，因此辨識正確分數的平均值則會隨著秒數上升而下降，所以造成辨識正確語辨識錯誤分數平均值間的距離隨著秒數上升而縮短。由下圖 3.12 可以發現三種秒數等錯率幾乎是差不多的，會造成這樣的結果就在於雖然辨識正確與辨識錯誤的分數分佈都因為秒數上升而方差下降，但其平均值也越來越靠近，所以造成其重疊的部分並沒有因此減少。

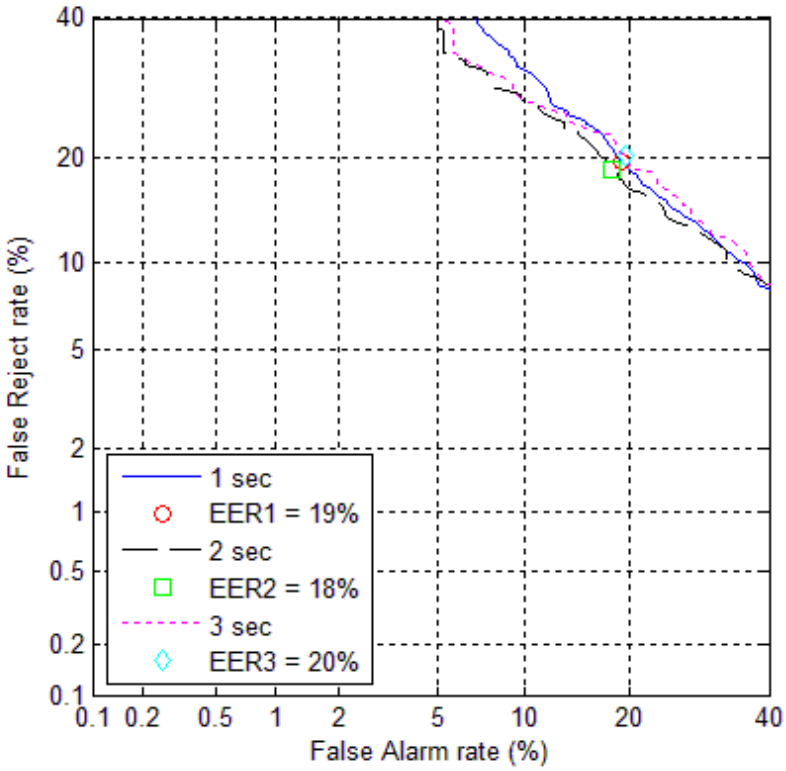


圖 3.12：通用模型正規化法 1、2、3 秒接收者操作特徵曲線圖

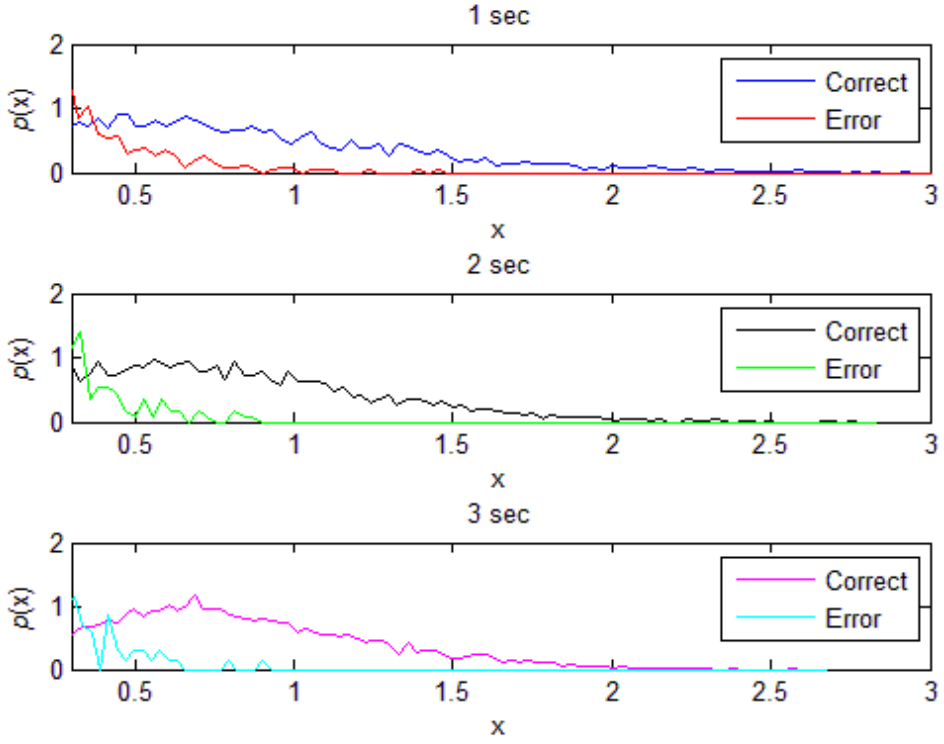


圖 3.13：最大值正規化法 1、2、3 秒辨識正確及辨識錯誤的分數分布

表 3.3：最大值正規化法辨識正確及辨識錯誤的分數分布統計

測試秒數 / 分數	1 秒	2 秒	3 秒
辨識正確平均值	0.78	0.78	0.78
辨識正確方差	0.28	0.21	0.19
辨識錯誤平均值	0.21	0.17	0.16
辨識錯誤方差	0.04	0.03	0.02

由圖 3.13 及表 3.3 我們可以看出在最大值正規化法下辨識正確的分數分佈的平均值在三種秒數下差別不大，但是方差就有比較明顯的下降，根據通用模型正規化的實驗結果，目標語者分數與封閉集合冒名頂替者的分數隨著秒數增加方差跟著變小，但平均值則相差不大，因此可以推測在辨識正確的情況下第一名（目標語者）與第二名（封閉集合冒名頂替者之一）分數的方差也隨著秒數變小，平均值則應該變化不大，因此兩者相減得到的最大值正規化後分數的方差也會隨著

秒數增加而變小，平均值則也相差不大。另一方面辨識錯誤的分數的平均值與方差則都隨著秒數上升而下降，其原因據推測應該是因為可能造成辨識錯誤的封閉集合冒充頂替者分數必定落在目標語者及封閉集合分數分布重疊的區間，而且因為隨著秒數增加辨識錯誤的機率也越來越低，所以識錯誤時目標語者所在的名次的平均值也更接近第二名，因此第二名分數應該也是落在此區間。而此區間根據前面實驗果將會隨著秒數上升而變小，表示辨識錯誤情況下第一、二名的分數分布將更集中，這將造成一二名分數的差值越來越小，且其差值變動的範圍也變小，因此辨識錯誤的分數經過最大值正規化後，平均值及方差均隨著秒數上升而下降。由下圖 3.14 可以看出由於秒數增加，辨識正確分數與辨識錯誤分數平均值間的距離變大，且兩群分數的方差皆變小因此其重疊的部分也變小，等錯率也就隨著秒數增加而變小。

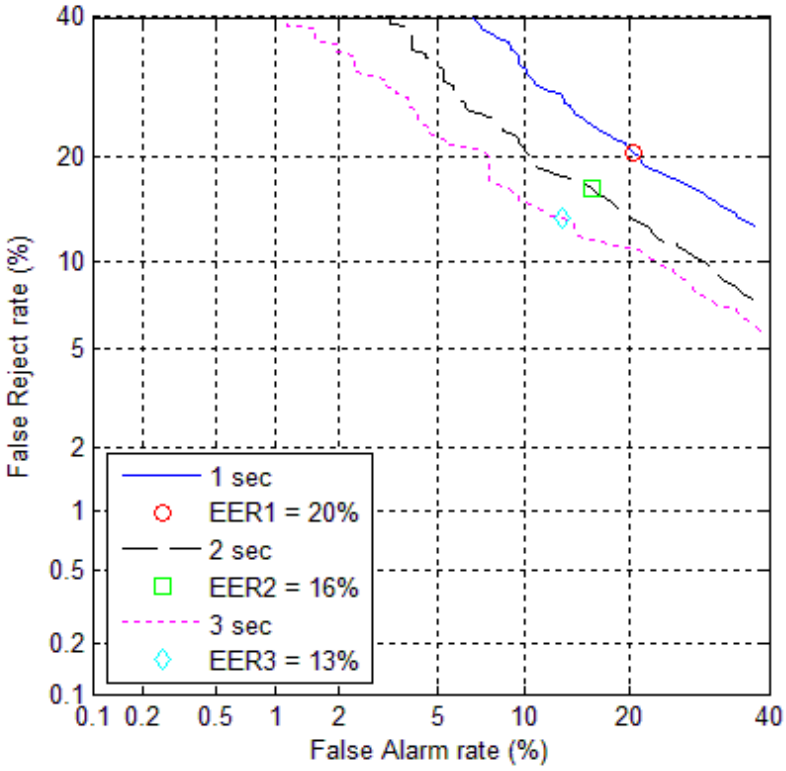


圖 3.14：最大值正規化法 1、2、3 秒接收者操作特徵曲線圖

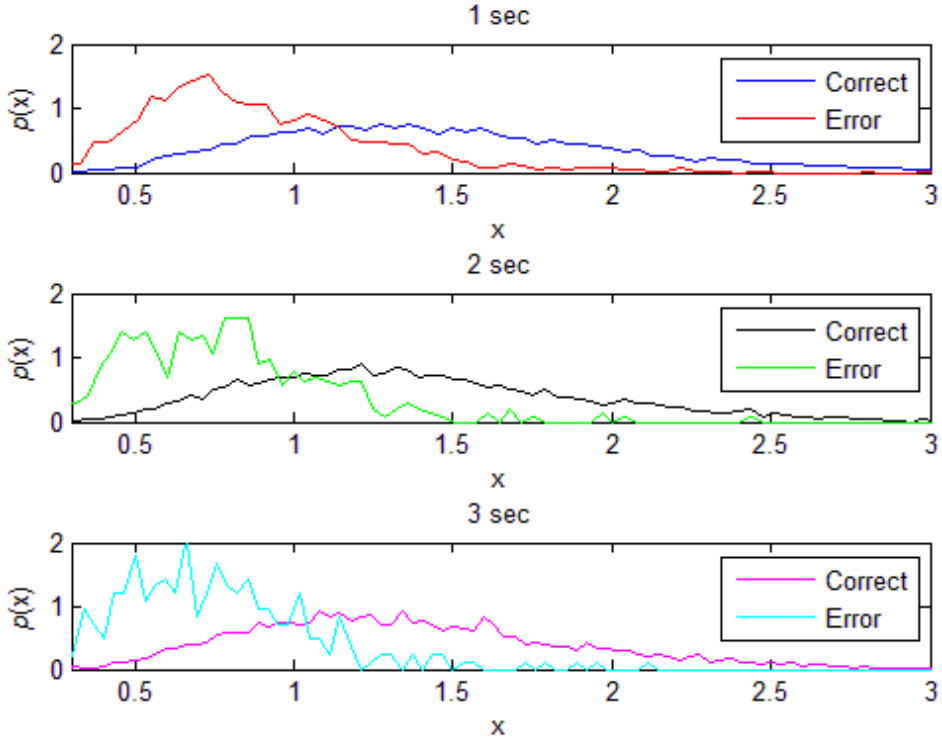


圖 3.15：幾何平均數正規化法 1、2、3 秒辨識正確及辨識錯誤的分數分布

表 3.4：幾何平均數正規化法辨識正確及辨識錯誤的分數分布統計

測試秒數 / 分數	1 秒	2 秒	3 秒
辨識正確平均值	1.50	1.40	1.37
辨識正確方差	0.36	0.29	0.25
辨識錯誤平均值	0.90	0.78	0.76
辨識錯誤方差	0.14	0.10	0.10

如下圖 3.16 所示，幾何平均數正規化法雖然也隨著秒數增加等錯率變小，但等錯率下降的幅度明顯不如最大值正規化法。觀察圖 3.15 及表 3.4 裡辨識正確與辨識錯誤的分數的分佈可以發現，隨著秒數增加兩者的平均值與方差都下降，因此兩群分數平均值間的距離只有些微的變化，等錯率會下降主要是因為方差下降使得重疊部分變少的緣故。而最大值正規化法則是辨識正確分數的平均值幾乎不變，辨識錯誤的分數隨著秒數增加而變小，因此兩者平均值的距離隨著秒數增

加而變大，且兩者方差又都隨著秒數變小，因此其等錯率下降幅度明顯比幾何平均數正規化法大上許多。推測應該是因為鑑別度的原因，相對接近的第二名比起一直都很遠的後面幾名更有鑑別第一名屬於哪一分類的能力。例如假設目標語者 A 的某一音檔取 1~3 秒皆辨識為語者 B，則依據之前的觀察第一名 B 與辨識結果為第二名或其他名次的 A 分數的距離會越來越小，但對於一些已註冊模型，此音檔所得到的分數可能一直就是很低，隨著秒數增加分數變化並不大，因此使用最大值正規化法時隨著秒數增加，我們發現等錯率越來越小，而在幾何平均數正規化法下雖然第二名越來越靠近第一名，但受到第二名分數還必須與其他名次相加而平均的結果，所以鑑別程度就不如最大值正規化法來得佳。

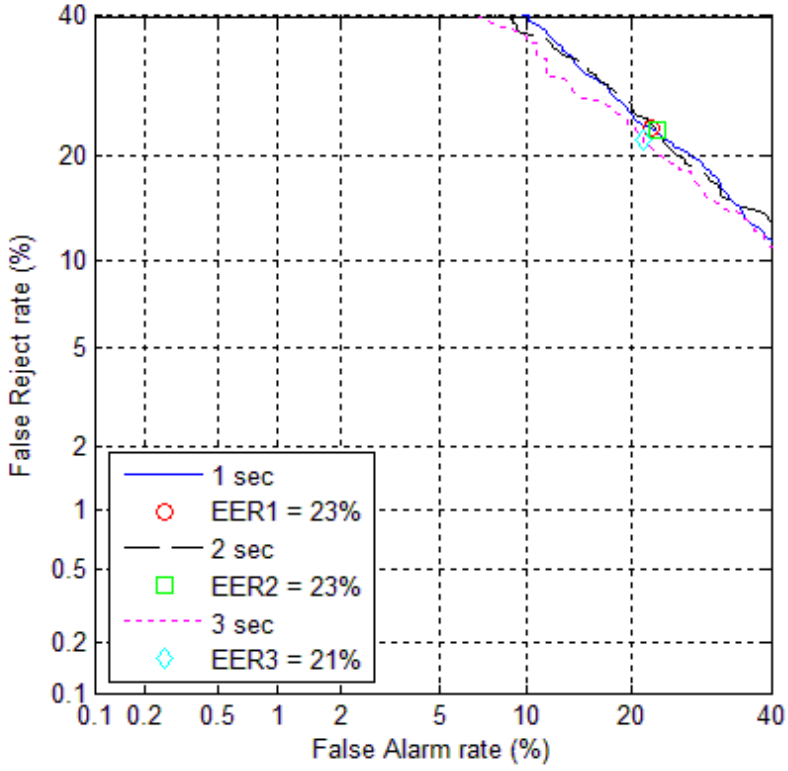


圖 3.16：幾何平均數正規化法 1、2、3 秒接收者操作特徵曲線圖

3.2.2 語者驗證信心指數

語者驗證系統分為兩個階段，第一階段希望將非註冊語者與辨識錯誤的結果判定無效，第二階段則對於辨識結果給於一信心指數以利其與其他系統進行整合。而為了確認系統該如何挑選兩階段所用之正規化方法，我們挑出於前面辨識正確與錯誤分數分布實驗結果等錯率較低的最大值正規化法及通用模型正規化法，以四種可能挑選的方式：兩階段皆為最大值正規化法，以 MN+MN 表示；第一階段為最大值正規化法，第二階段為通用模型正規化法，以 MN+WMN 表示；兩階段皆為通用模型正規化法，以 WMN+WMN 表示；第一階段為通用模型正規化法，第二階段為最大值正規化法，以 WMN+MN 表示。探討於 1、2、3 秒時辨識正確與錯誤分數的平均值，以及更進一步地觀察系統可能出現的四種結果信心指數的平均值：辨識正確且系統判定有效，以 Correct &accept 表示；辨識正確但系統判定無效，以 Correct &reject 表示，而其機率以 FR rate 表示；辨識錯誤但系統判定有效，以 Error &accept 表示，而其機率以 FA rate 表示；辨識錯誤且系統判定無效，以 Error &reject 表示。可以得到下面三個表。

表 3.5：1 秒正規化方法選用及辨識結果信心指數平均值

	MN+MN	MN+WMN	WMN+WMN	WMN+MN
	mean	mean	mean	mean
Correct	0.68	0.70	0.70	0.68
Correct &accept	0.78	0.75	0.78	0.74
Correct &reject	0.27	0.48	0.33	0.42
FR rate	19.6%		17.7%	
Error	0.31	0.30	0.30	0.31
Error &accept	0.64	0.30	0.66	0.30
Error &reject	0.22	0.30	0.23	0.31
FA rate	20.7%		16.6%	

表 3.6：2 秒正規化方法選用及辨識結果信心指數平均值

	MN+MN	MN+WMN	WMN+WMN	WMN+MN
	mean	mean	mean	mean
Correct	0.71	0.70	0.70	0.71
Correct &accept	0.78	0.75	0.78	0.76
Correct &reject	0.31	0.46	0.31	0.45
FR rate	15.9%		16.1%	
Error	0.28	0.30	0.30	0.28
Error &accept	0.63	0.32	0.63	0.33
Error &reject	0.23	0.29	0.22	0.27
FA rate	13.1%		17.9%	

表 3.7：3 秒正規化方法選用及辨識結果信心指數平均值

	MN+MN	MN+WMN	WMN+WMN	WMN+MN
	mean	mean	mean	mean
Correct	0.72	0.70	0.70	0.72
Correct &accept	0.77	0.74	0.79	0.77
Correct &reject	0.30	0.41	0.33	0.48
FR rate	12.1%		19.2%	
Error	0.27	0.30	0.30	0.27
Error &accept	0.60	0.37	0.62	0.30
Error &reject	0.22	0.29	0.23	0.26
FA rate	12.4%		18.2%	

由這三個表可以看出兩種第一階段與第二階段選用不同正規化方法的挑選方式，在系統將辨識錯誤結果誤判為有效時，其信心指數的平均值幾乎是另外兩種兩階段挑選相同正規化方法的挑選方式的一半，也就代表說如果系統選用這兩

種兩階段選用不同正規化法的挑選方式，即使在第一階段將辨識錯誤的結果誤判為有效的辨識結果，則依然可以給予此結果一較低的信心指數，以利與其他使用者辨識系統整合時，將此誤判更正。而為了探討兩階段選用不同正規化方法的好處，我們假設所有測試語料中辨識正確但在通用模型正規化法裡被判定為無效的集合為 FR_W ，而在最大值正規化法裡被判定為無效的集合為 FR_M ，而兩個集合匹配程度則定義為

$$\frac{FR_M \cap FR_W}{FR_M \cup FR_W} \qquad (3\text{-}27)$$

而所有測試語料中辨識錯誤但在通用模型正規化法裡被判定為有效的集合為 FA_W，在最大值正規化法裡被判定為無效的集合為 FA_M，而兩個集合匹配程度則定義為

$$\frac{FA_M \cap FA_W}{FA_M \cup FA_W} \qquad (3\text{-}28)$$

則可以得到表 3.8，並可由其結果看出兩種正規化方法各自誤判集合中，重疊的部分相當少，所以某部分辨識錯誤的結果，即使系統在第一階段並未成功地將其偵測出來並判定無效，在第二階段使用另一正規化方法時，依然有相當大的機會給予較正確的信心指數值。

表 3.8：正規化方法誤判集合匹配程度

	FA match	FR match
1 sec	10.5%	35.6%
2 sec	13.9%	35.0%
3 sec	9.7%	34.1%

根據以上的實驗結果，本論文系統在第一階段選用最大值正規化法做為判定辨識結果是否有效的方法，因為考慮到家用使用情況，比較常見為已註冊使用者使用的情況，而最大值正規化法在測試語料一秒時等錯率雖然與通用模型正規化法相差不多，但隨著秒數增加最大值正規化法的等錯率則越來越低，通用模型正規化法則無明顯下降，幾合平均數正規化法則是整體等錯率皆比較高。而對於開

放集合冒名頂替者的鑑別能力，依據實驗結果最大值正規化法不如通用模型正規化法，因此系統在第二階段選用通用模型正規化法。如此一來雖然在某些情況下系統無法在第一階段就將辨識錯誤或是冒名頂替者判定無效，但可透過第二階段給予較低的信心指數，以用於與其他使用者是別的系統整合時更正誤判。

第四章 文本相關語者註冊設計

基於隱藏式馬可夫模型建立文本相關語者辨識系統，是目前被廣泛使用的方法，在本章中將利用 TCC300 語料庫以及劍橋大學開發之 HTK（HMM Tool Kit）軟體【14】建立一文本相關語者辨識系統，並針對家用環境設計整體註冊流程。4.1 節介紹本研究所採用的有限制的最大概似線性迴歸法則（constrained maximum likelihood linear regression, CMLLR）調適語者無關的隱藏式馬可夫模型的語者辨認系統；4.2 節介紹註冊流程的設計及想法；4.3 節介紹實驗結果。

4.1 文本相關語者辨識系統

在基於隱藏式馬可夫模型之文本相關語者辨識系統裡，依然使用梅爾倒頻係數做為特徵參數，但為了能更確切的描述語音音節的特性，我們選擇使用 13 維梅爾倒頻係數加入其一階及二階差量及除去能量那一維共 38 維特徵參數，且不經過 RASTA 濾波器以減少對於語音音節特性造成的失真。整個流程主要可分為註冊及辨識兩個階段，而註冊又包含了語者無關聲學模型建立及語者模型註冊，本節將依序介紹聲學模型建立、註冊及辨識。

4.1.1 語者無關聲學模型之建立

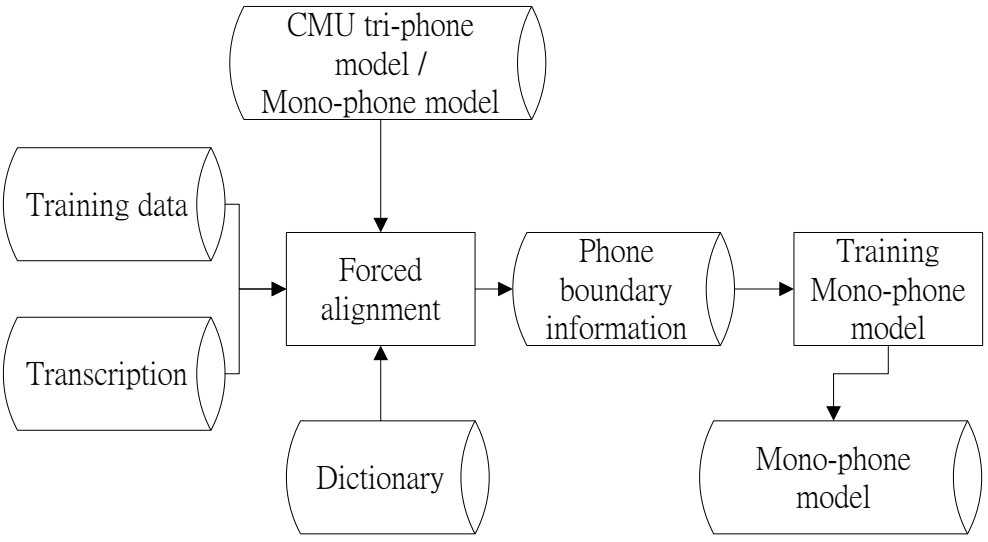


圖 4.1：聲學模型之建立流程

建立流程如圖 4.1 所示。一開始利用卡內基美隆大學（Carnegie Mellon University, CMU）所提供的英文三連音素模型（tri-phone model）對訓練語料做音素的強迫對位（forced alignment），得到訓練語料的音素切割位置，並依此建立初始的聲學模型。得到初始的中文單音素聲學模型之後，再次對訓練語料進行音素切割，以得到較好的切割位置，並使用新的切割位置重新訓練較好的單音素聲學模型。本系統使用之聲學模型為單音素模型（mono-phone model），而每一個音素的隱藏式馬可夫模型採用 3 個由左至右的狀態表示。

4.1.2 語者註冊

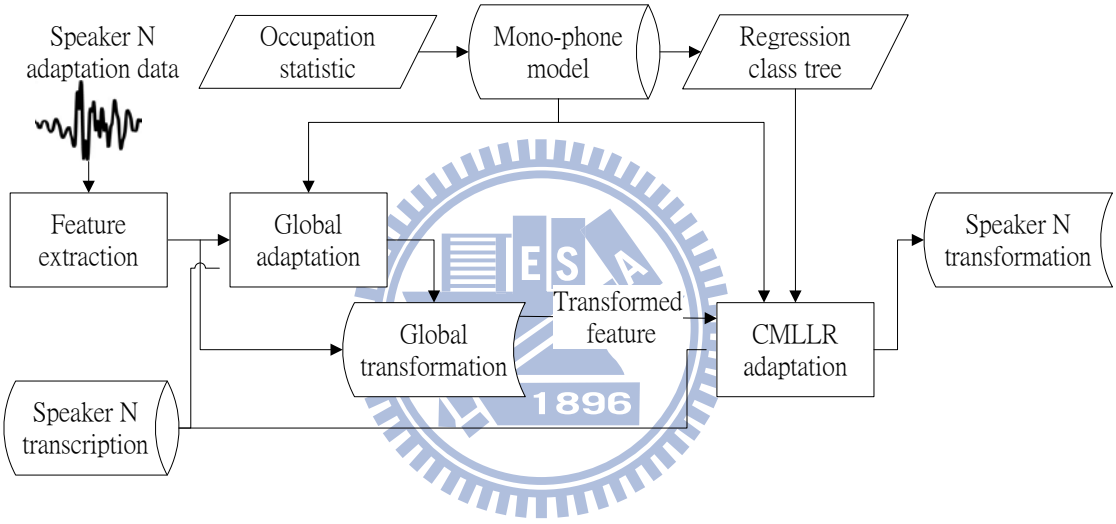


圖 4.2：文本相關語者註冊流程

語者註冊如圖 4.2 所示分為兩個步驟，步驟一為先將註冊語料經過參數抽取為 38 維梅爾倒頻譜參數後，對單音素模型進行全域調適（global adaptation）以得到全域轉換函式。步驟二為先利用訓練單音素聲學模型所產生的統計資料（occupation statistic）去產生分類迴歸樹（regression class tree），接著將註冊語料的參數經過步驟一的全域轉換函式，以得到較好的音框及狀態對位（alignment），並使用此對位結果和分類迴歸樹去得到每個分類更準確的轉換函式。

先做全域調適的意義在於，全域轉換函式適用於所有的在我們要調適的模型集合裡的高斯分布，所以當我們沒有足夠的調適語料時就只用全域轉換函式，當

我們有較多的調適語料時就可以更進一步去將高斯分布分群，並找出其相對應的轉換函式。而分群的目的在於找出聲學空間（acoustic space）上距離相近的高斯分布，使得分在同一群的高斯分布使用同一個轉換函式。分群的好處在於，因為不是固定將哪些高斯分布分於某群，而是依據其調適語料落於各高斯分布的量而決定，所以即使某些高斯分布並沒有調適語料，依然可以與其相近的高斯分布使用同一個轉換函式，而當其有足夠的調適語料則又可以再重新找出其對應的轉換函式，如此一來不管有無足夠的調適語料，所有的聲學模型都可以做調適。

4.1.3 語者辨識

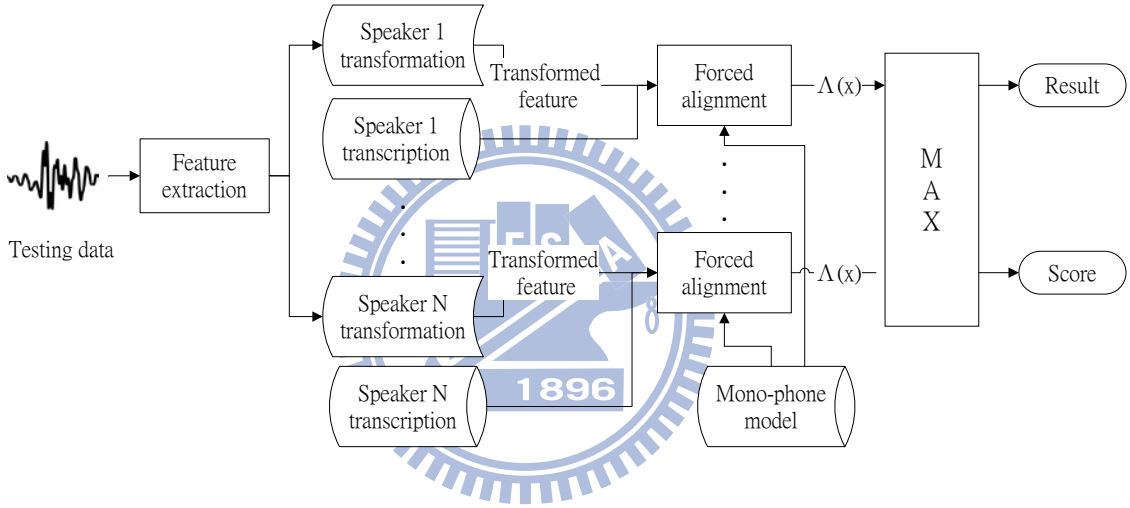


圖 4.3：文本相關語者辨識流程

語者辨識時，測試語料在抽成參數後分別經過註冊名單裡每位語者的轉換函數進行參數轉換，並以該位語者的註冊文本資訊做音素的強迫對位。接著得到音素序列切割後每個音素的概似分數，除去靜音及短停頓的概似分數後，將每個音素所得到的概似分數除以其音框數並加總，概似分數最高者即為語者辨識的結果。而之所以除去靜音及短停頓的概似分數，就意義上來說可視為一種語音端點偵測，在前面整合麥克風陣列的語者辨識實驗裡，我們也發現經過語音短點偵測後的辨識率是較高的，因為除去靜音及短停頓後的語料將留下真正屬於該名語者的聲音資訊，而靜音及短停頓則不但沒有包含語者資訊還有會影響語者辨識的通道效應。

4.2 註冊流程設計

對於家用情境而言，使用者希望能以最自然簡單的方式操作系統，因此註冊系統應要能避免必須使用鍵盤輸入任何資訊的情況，所以本節將先介紹針對完全靠語音輸入資訊之系統的註冊流程設計。此外，由於家庭中有各年齡層成員，彼此間使用之語言偏好及內容也不盡相同，本節將接著介紹混雜英文、台語及國語註冊系統。

4.2.1 語音輸入註冊資訊

針對家用機器人上的語者辨識系統設計，我們將情境設定為，每位使用者可以藉由呼叫機器人的名字或是自己的名字來做語者辨識。一般情況使用者必須由鍵盤輸入其所想要做為辨識依據的特定名字，以提供系統註冊文本的資訊，但為了可以讓使用者更自然地與機器人互動，我們在此設計一音節辨識器，系統只需由麥克風擷取使用者的註冊語料，即可自動轉成註冊所需的文本資訊。而音節辨識器分為註冊語料音節辨識、註冊語料調適與決定最佳音節辨識結果三個主要部分，將依序介紹如下。

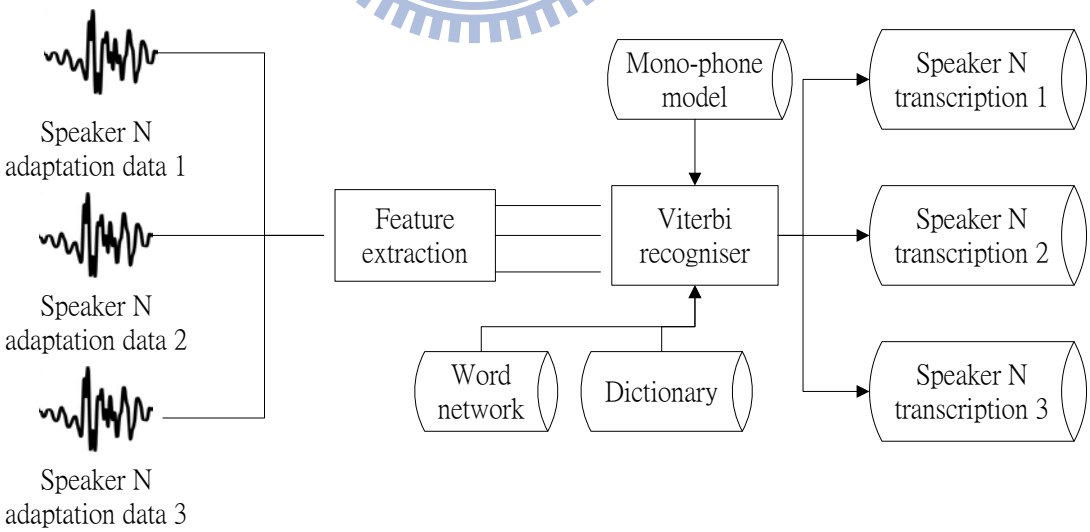


圖 4.4：註冊語料音節辨識

首先如圖 4.4 所示，系統將擷取使用者念其所選定之特定名字三次的註冊語

料，接著將這三次語料用語者無關的單音素聲學模型做語音辨識。由於一開始我們先考慮中文使用的情況，因此在系統中加入文法的限制，規定其辨識結果必為中文語音的 411 個音節之一。如此一來系統將得到三組可能相同或不盡相同的音節序列，做為可能當做此語者註冊文本的候選音節序列。

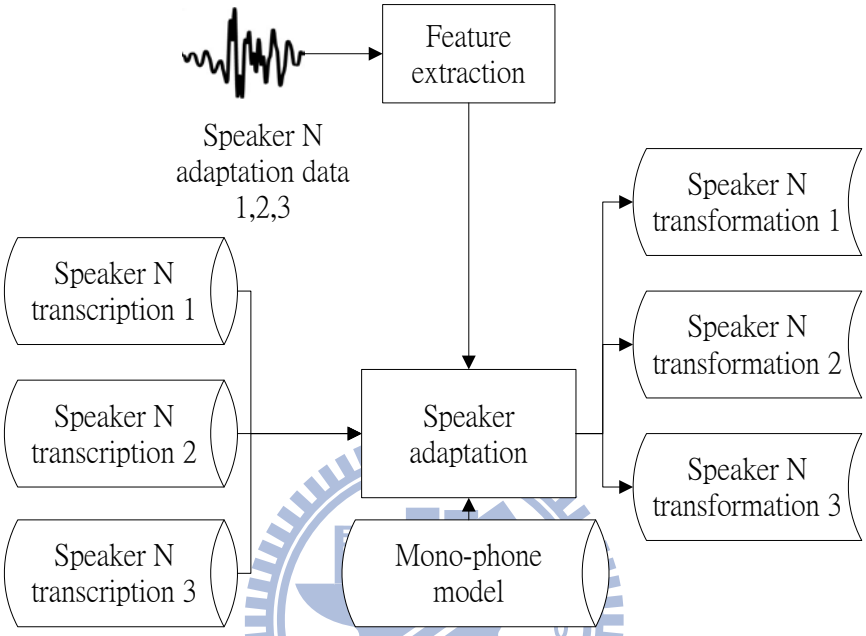


圖 4.5：註冊語料調適

第二部分如圖 4.5 所示，我們接著將所有的註冊語料分別以三組候選音節序列為註冊文本資訊，送入文本相關語者註冊系統裡進行語者調適，而註冊系統將輸出三組相對應的語者轉換函式。

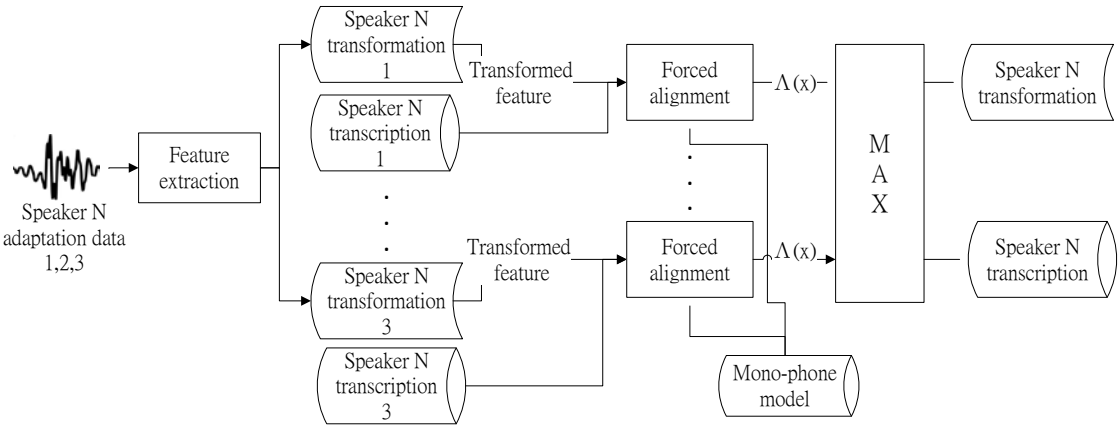


圖 4.6：決定最佳音節辨識結果

第三部分如圖 4.6 所示，在得到三組語者轉換函式之後，我們將所有的註冊

語料分別經過三組轉換函式並以相對應的文本資訊做音素的強迫對位。在得到音素序列切割所得到的概似分數之後，一樣除去靜音及短停頓所得到的概似分數，並以概似分數最高那組轉換函式及文本資訊做為該位語者的標準轉換式及正確文本，供之後語者辨識時使用。

4.2.2 混合語言註冊系統

考慮到家用環境裡各年齡層的成員很可能習慣使用的語言不盡相同，且希望做為語者辨識的特定名字能有較大的自由度，一方面可以更貼近使用者的使用習慣及喜好，另一方面由於名字的選擇性增加，使得辨識系統的安全性也有所提升，因此我們更進一步地改進系統，讓使用者可以沒有任何限制的使用任何他想要的語言、發音的名字做為辨識用的通關密碼。因為我們無法針對所有使用者可能使用的語言去訓練相對應的聲學模型，所以我們依然採用系統原本的單音素聲學模型，但修改文法的限制及字典內容，使得音素之間的連接不再限制於中文的 411 個音節裡的音素組合，只限制 38 個音素中的聲母及介音後面不可以接短停頓（因為人類實際上的發聲情況，聲母無法單獨成聲）。如此一來使用者不論用任何唸法或語言做為註冊語料，系統依然可以找出相近的音素序列做為該名語者的文本資訊，並以此做語者辨識。

4.3 實驗結果

4.3.1 訓練、註冊與測試語料

本系統訓練單音素模型的訓練語料為 TCC300 語料庫中從中挑選 9/10 的檔案，其中包含 271 個語者，136 男 135 女，音節總數為 300836 個。此外我們依據附錄一的文本，錄製 16 位男生語者的實驗語料，每位語者語料內容為包含 2~5 個字的機器人名字，每個字數有 10 個名字，每個名字錄 8 遍，其中前 3 遍為註冊語料，後 5 遍為測試語料。所以每位語者有 2~5 字，每種字數各有 10 組內容不同的註冊語料，以及每種字數各有 50 筆測試語料。

4.3.2 基礎文本相關語者辨識系統實驗

基礎文本相關語者辨識系統實驗包含三種操作情境的實驗。情境一為十位語者在註冊時，給系統一組通關密碼的文本資訊以及三次調適語料，每個語者之間的通關密碼皆不相同，而使用時每位使用者念自己的通關密碼來進行語者識別。為了得到情境一的辨識率資訊，我們挑選 2~5 字，每個字數 10 組通關密碼的組合來做實驗，得到數據如下表 4.1。

表 4.1：基礎系統文本相關語者辨識率

2 syllables	3 syllables	4 syllables	5 syllables
97.2%	100.0%	100.0%	100.0%

由表可以看出，除了 2 個字的通關密碼以外，其他字數的通關密碼辨識率皆為百分之百。可見得在文本相關的語者辨識系統裡，只需要很少量的調適語料，語料文本內容，及很短的測試語料即可達到不錯的語者辨識率。

接著情境二為封閉集合的冒名頂替者使用系統，註冊語料與情境一相同，但此時測試語料分為兩類，第一類為目標語者但念其他語者之通關密碼之情形，第二類為非目標語者但卻念目標語者的通關密碼，這兩類皆為必須被拒絕的辨識結果，我們在此統計出其概似分數分布，並藉由統計情境二與情境一分數分布得到的接收者操作曲線及等錯率點，實驗結果一樣分為 2~5 個字的測試語料。

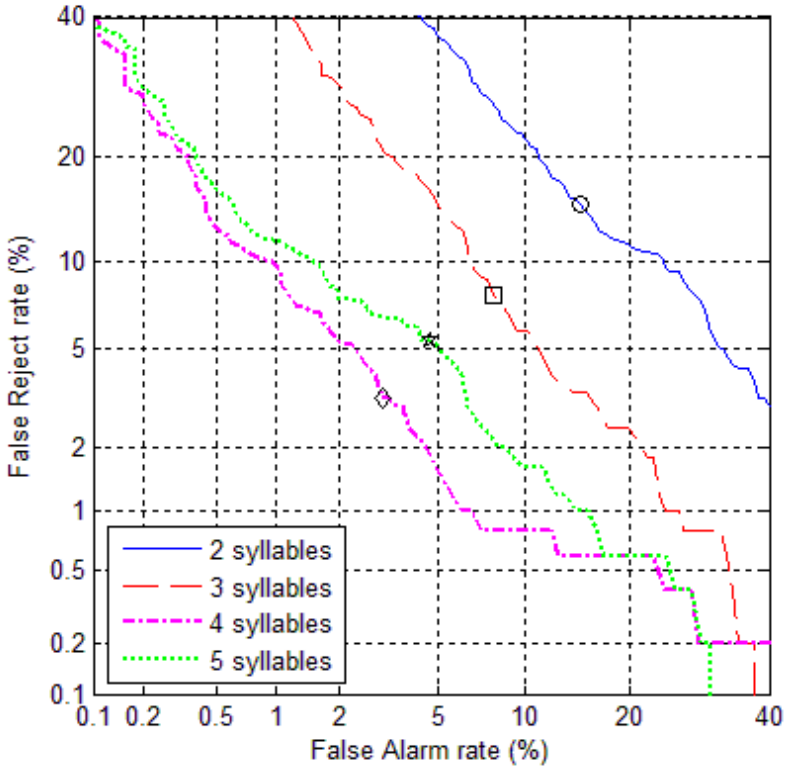


圖 4.7：基礎文本相關語者驗證封閉集合操作者接收曲線圖

表 4.2：基礎文本相關語者驗證封閉集合等錯率

2 syllables	3 syllables	4 syllables	5 syllables
15.4%	7.8%	3.0%	4.6

由此結果我們可以看出，等錯率隨著字數的上升而下降，表示註冊時使用越多字做為通關密碼，語者驗證的誤判率就越低。但四字詞卻稍微低於五字詞，可能是語料內容造成或是當超過四個字後系統強建性反而變比較差。

最後情境三為開放集合的冒名頂替者使用系統，，註冊語料與情境一相同，但此時測試語料為六位非註冊名單之語者，念註冊這十個通關密碼。這所有測試結果皆為必須被拒絕的辨識結果，我們在此統計出其概似分數分布，並藉由統計情境三與情境一分數分布得到的接收者操作曲線及等錯率點，實驗結果一樣分為2~5 個字的測試語料。

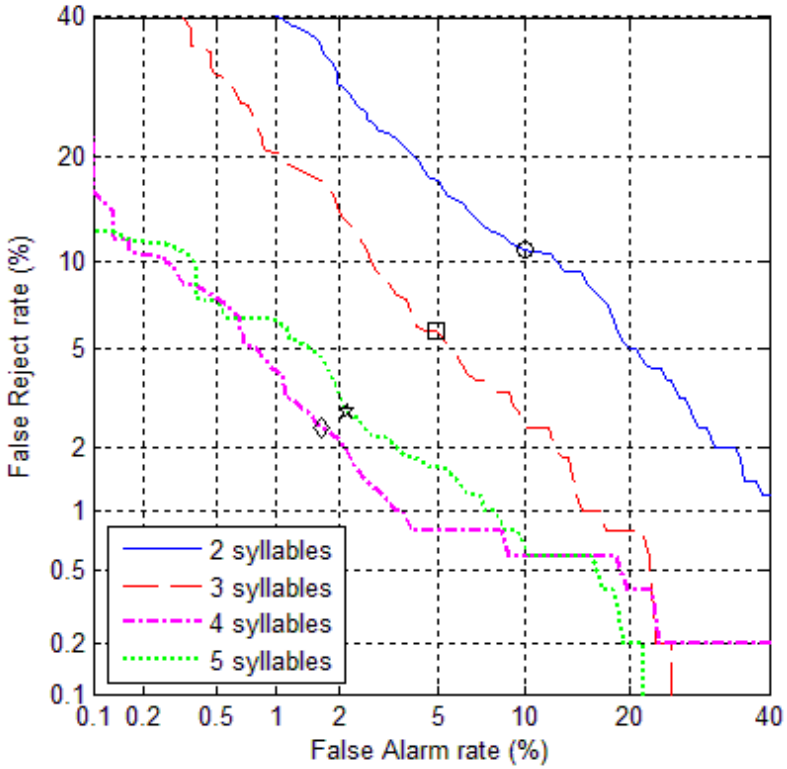


圖 4.8：基礎文本相關語者驗證開放集合操作者接收曲線圖

表 4.3：基礎文本相關語者驗證開放集合等錯率

2 syllables	3 syllables	4 syllables	5 syllables
11.0%	5.8%	2.4%	2.8%

由此結果我們可以看出，等錯率一樣隨著字數的上升而下降，表示註冊時使用越多字做為通關密碼，語者驗證的誤判率就越低。此外 2~5 字詞的等錯率也明顯的比封閉集合的冒名頂替者低，據推測應該是因為封閉集合的冒名頂替者包含目標語者並非念本身正確的通關密碼與其它語者念目標語者的通關密碼，而開放集合的冒名頂替者雖然也與註冊語者念一樣的通關密碼，但其全部皆為非目標語者，因此系統對於開放集合的冒名頂替者較不容易混淆。

4.3.3 語音輸入註冊資訊的語者辨識系統實驗

為了給予使用者更直覺地操作方式，我們將不需要使用者手動輸入文本資訊，而是由系統提供音素序列辨識器結果來做為註冊文本輸入。而對於在有限制

其文法為中文語音的 411 個音節中音素可連接方式的情況下，10 位註冊名單裡語者的註冊與測試語料進行音素序列辨識，得到如下表的統計資料。

表 4.4：有文法限制音素序列辨識器辨識結果統計

總音素個數	相同音素個數	刪除音素個數	改變音素個數	插入音素個數
49040	42648	1232	5160	2512

由結果可以看出音素序列辨識器，所辨識出來的音素序列，與標準答案有八成左右的內容是相同的，也就是辨識結果相當接近使用者所念之音。接著我們同樣探討若註冊文本為音素序列辨識器輸出時三個使用情境下， 語者驗證的情形。

表 4.5：語音輸入註冊文本資訊語者辨識系統辨識率

2 syllables	3 syllables	4 syllables	5 syllables
96.8%	99.6%	100.0%	100.0%

由結果可以看出辨識率與直接輸入註冊文本資訊的基礎系統差別不大，代表在前端改為用語音輸入註冊文本資訊是可行的辦法。接著觀察情境二的結果。

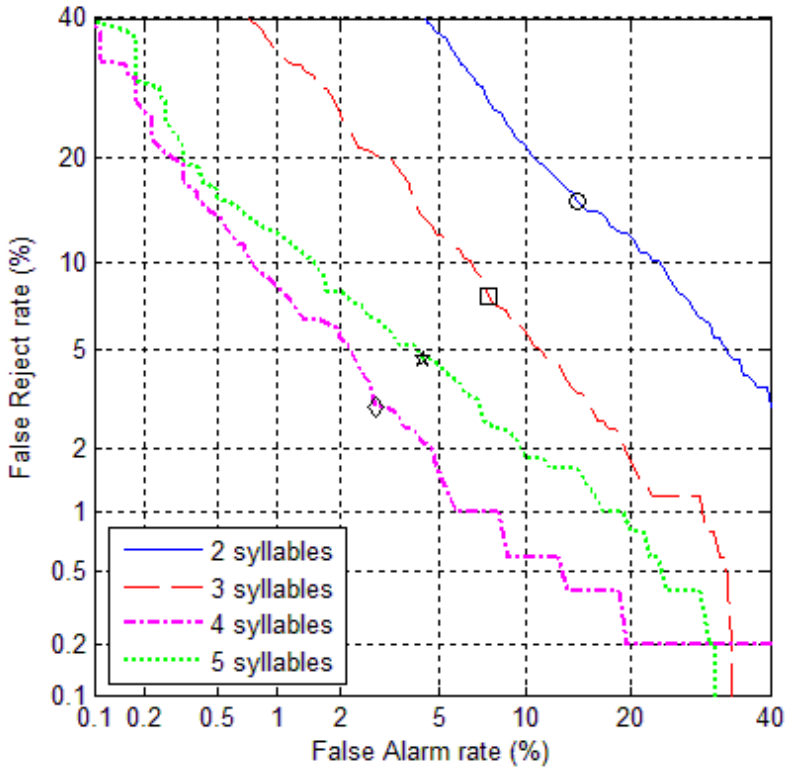


圖 4. 9：語音輸入註冊文本資訊語者驗證封閉集合操作者接收曲線圖

表 4.6：語音輸入註冊文本資訊語者驗證封閉集合等錯率

2 syllables	3 syllables	4 syllables	5 syllables
15.4%	7.8%	3.0%	4.6

由圖 4.9 及表 4.6 的結果比對基礎系統的結果可以發現與語者識別辨識率一樣差別不大，接著觀察情境三下的結果。

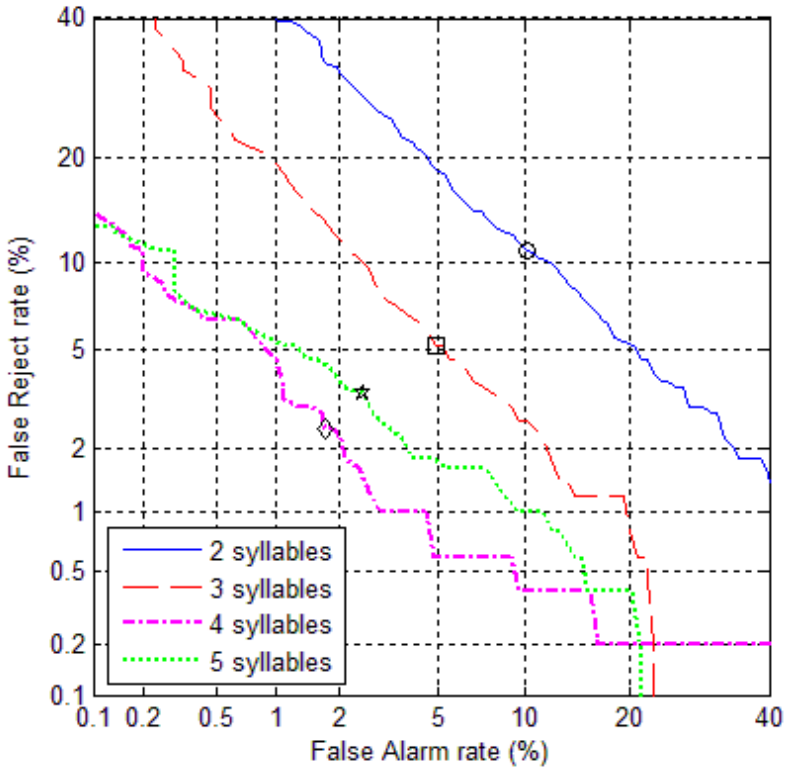


圖 4.10：語音輸入註冊文本資訊語者驗證開放集合操作者接收曲線圖

表 4.7：語音輸入註冊文本資訊語者驗證開放集合等錯率

2 syllables	3 syllables	4 syllables	5 syllables
11.0%	5.2%	2.4%	3.4%

情境三的結果也與基礎系統的結果差不多，由此可知將前端文本輸入改為由音素序列辨識器的辨識結果輸入，對於系統的性能影響不大，但卻可以大大提升系統在使用上的便利性。

4.3.4 混合語言語者辨識系統實驗

為了針對家用環境中使用者使用上的方便與喜好，我們希望可以將系統應用於多種語言同時混用的情境。因此我們將文法限制改為只限制 38 個音素中的聲母及介音後面不可以接短停頓。而對於在新限制下，10 位註冊名單裡語者的註冊與測試語料進行音素序列辨識，得到如下表的統計資料。

表 4.8：新文法限制音素序列辨識器辨識結果統計

總音素個數	相同音素個數	刪除音素個數	改變音素個數	插入音素個數
49040	38800	4200	6040	5480

由表 4.8 可以看出辨識結果較文法限制較多的方法略差一點，表示如果有使用者使用語言的資訊，則可以較好地限制文法使其辨識結果正確率較高。接著為了考慮到更真實使用系統的情形，我們重新分配五組註冊語料，每組註冊語料中含有兩位語者使用 2~6 音節不等的英文通關密碼，另兩位語者使用 2~5 個字不等的台語通關密碼，以及 2 個字、5 個字通關密碼的語者各一位和 3 個字、4 個字通關密碼的語者各兩位共十位語者。首先我們一樣用三筆語料註冊，而每位語者每個通關密碼用五筆語料做測試，五組註冊語料得到的語者識別辨識率平均為 98.4%，與原本系統辨識率相差不多。接著我們考慮在真實情況中，使用者之間並不容易猜到彼此的通關密碼，因此在測試封閉集合冒名頂替者實驗時，每組註冊語料使用每位語者念五個完全不屬於這組註冊通關密碼的任意語言通關密碼做測試，得到的果如下圖所示，等錯率大概是 8%左右。

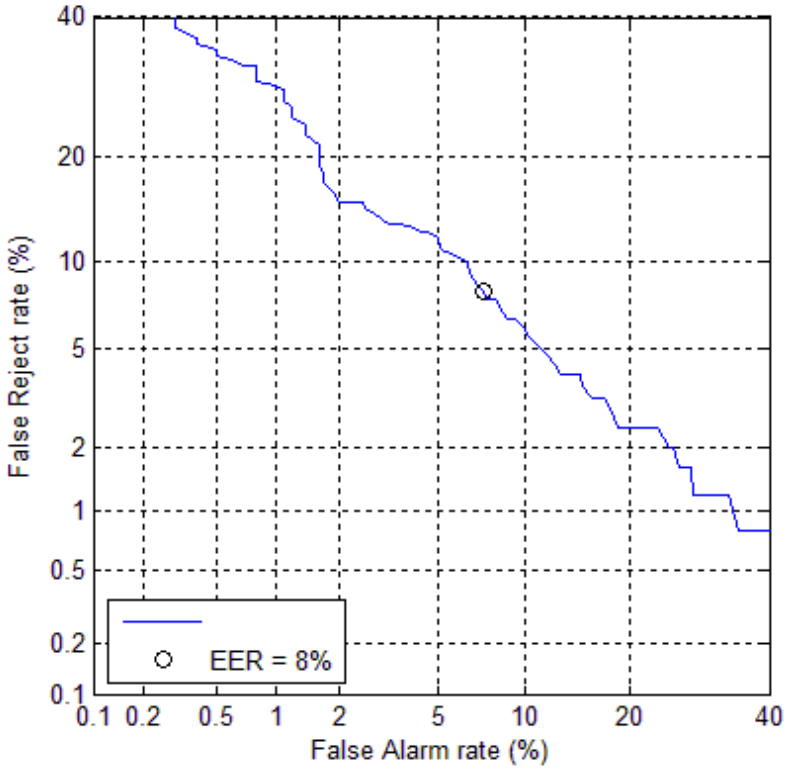


圖 4.11：混合語言開放集合操作者接收曲線圖及等錯率

由結果可以看出，本系統確實可以用於混合語言的情境下，且只需要不到五秒的註冊語料，及一秒左右的測試語料，即可以達到不錯的語者識別率及語者驗證功能。

第五章 結論與未來展望

5.1 結論

本論文實作出了一套文本獨立的高斯分布模型的語者辨識系統及一套結合通關密碼與語者辨識的文本相關的隱藏式馬可夫模型語者辨識系統。高斯分布模型系統在乾淨的 20 秒註冊語料，以及 3 秒乾淨的測試語料下，可以有 93.22%的語者識別率，以及誤判率大約為 13%的辨識結果正確性偵測。更進一步地，對於基礎的高斯分布模型系統，還整合了麥克風陣列以及波束形成及空間濾波器的處理，以在訊雜比 5dB 的吵嘈環境下有 75%以上的語者識別率。而文本相關的隱藏式馬可夫模型語者辨識系統，則是整合不限定語言之小詞彙語音辨識系統及語者辨識系統，使其可以在完全語音輸入文本資訊及註冊語料且不限語言的情況下，且使用相較於高斯分布模型語者辨識系統而言較短的註冊和辨識語料，就有 96%以上的語者識別率，以及 8%左右的語者驗證等錯率。而兩套語者辨識系統，皆針對應用於家用環境的智慧型機器人可能面對的情境，設計相對應的註冊流程及整合相關技術，使其可讓使用者在家用環境中直覺地使用強鍵的語者辨識系統。

5.2 未來展望

對於文本相關的語者辨識系統而言，除了文本及聲學上的資訊以外，因為在註冊時系統擁有註冊語料切割位置的資訊，所以還可以再加入每個字的音長音高等等韻律上的資訊，以期再降低語者驗證的等錯率。此外，在未來還可以針對經過麥克風陣列及波束形成處理後所造成的訊號失真，對通用背景模型或是語者模型進行補償，以期提高於嘈雜環境中的辨識率。本論文的語者辨識系統在未來也可以更進一步地與人臉辨識等其他使用者身分辨識系統進行整合，讓使用者身分辨識系統能在更多元的環境下保持其準確性。

參考文獻

【1】 DA Reynolds and RC Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models.” IEEE Trans. Speech Audio Process. 3, pp. 72–83, 1995.

【2】Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models.” Digital Signal Processing 10, pp. 19–41, 2000.

【3】Sadaoki Furui, “Recent advances in speaker recognition.” Pattern Recognition Letters 18, pp.859-872, 1997.

【4】Rosenberg A Lee C Soong F, “Sub-word unit talker verification using hidden Markov models.” In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, pp. 269-272, 1990.

【5】D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” Speech Communication, vol. 17, no. 1-2, pp. 91–108, 1995.

【6】C.-S. Liu, H.-C. Wang, and C.-H. Lee, “Speaker verification using normalized log-likelihood score,” IEEE Trans. Speech Audio Processing, vol. 4, pp. 56–60, Jan 1996.

【7】Jiang, H., Deng, L.. A Bayesian approach to the verification problem: Applications to speaker verification. IEEE Trans. Speech Audio Process. 9 (8), pp.874–884, 2001.

【8】Roland Auckenthaler, “Score Normalization for Text-Independent Speaker Verification Systems.” Digital Signal Processing 10, pp. 42–54 , 2000.

【9】 A. M. Ariyaeeinia, “Verification effectiveness in open-set speaker identification.” IEE Proceedings -Vision, Image and Signal Processing,_vol. 153,issue 5,pp.618, 2006.

【10】RA Finan, ” Impostor cohort selection for score normalisation in speaker verification.” Pattern Recognition Letters18, pp.881-888, 1997.

【11】R Auckenthaler, “Score normalization for text-independent speaker verification systems.” Digital Signal Processing, 2000.

【12】S Gannot, “Signal Enhancement Using Beamforming and Nonstationarity with Applications to Speech.” IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 49, NO. 8, AUGUST 2001.

【13】Michael W. Hoffman, “GSC-Based Spatial Voice Activity Detection for Enhanced Speech Coding in the Presence of Competing Speech.“IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 9, NO. 2, MARCH 2001.

【14】S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P.Woodland, The HTK Book Version 3.0. Cambridge, U.K.: Cambridge Univ. Press, 2000.

附錄一：中文通關密碼文本

0. 查理

1. 魯夫

2. 來福

3. 旺旺

4. 小黑

5. 艾咪

6. 吉米

7. 胖虎

8. 丹尼

9. 湯姆

10. 加菲貓

11. 阿凡達

12. 派大星

13. 海賊王

14. 寶貝熊

15. 小叮噹

16. 艾利斯

17. 任我行

18. 艾瑞克

19. 安德魯

20. 海綿寶寶

21. 變形金剛

22. 小熊維尼

23. 蠟筆小新

24. 兩津勘吉

25. 哈利波特

26. 巴斯光年

27. 萬用鑰匙

28. 懦夫救星

29. 完美情人

30. 毛利小五郎

31. 魔鬼終結者

32. 摳比布萊恩

33. 俠客歐尼偶

34. 天龍特功隊

35. 交大鬍子哥

36. 飛天小女警

37. 布魯斯威利

38. 天才交響曲

39. 大雷神索爾

附錄二：英文及台語通關密碼文本

0.志明

1.春嬌

2.劉文聰

3.秦假仙

4.阿弟仔

5.緣頭阿桑

6.黑白郎君

7.我的麻及

8.台灣水姑娘

9.鐵獅玉玲瓏

10.Mayday

11.Merry Jiang

12.Jacky Wu

13.Eason Chen

14.Transformer

15.Linkin park

16.Lady Gaga

17.Lebron James

18.Michael Jackson

19.Penny Hardaway