國　立　清　華　大　學

碩 士 論 文

題目: 文本相關之語者識別及其不佳輸入之濾除機制

Research on Text-dependent Speaker Identification with Screening Mechanism for Impropriate Speech Inputs

系別　資訊工程學系　　組別

學號姓名　9662565 劉玉情 (Yu-Ching Liu)

指導教授　張智星博士 (Jyh-Shing Roger Jang)

中華民國九十八年六月

摘要

本論文主要基於語者識別應用在門禁系統下所作的研究，為了安全上的考量，測試語音和參考語音皆為文本相關(Text Dependent)，且以動態時間扭曲(Dynamic Time Warping, DTW)的方法作語者識別。而本論文的目標是希望能夠提升語者識別的初始辨識率，因此提出兩大改善的方法，分別為改善端點偵測錯誤、濾除不佳的測試語音。

首先將測試語音和參考語音擷取出 13 維的 MFCC，利用 13 維的 MFCC 作動態時間扭曲比對後，得到初始語者識別的辨識率。根據初始辨識錯誤的音檔作分析，得知端點偵測(Endpoint Detection)錯誤是影響辨識錯誤的主要原因，進而提出改善端點偵測錯誤的方法，使得在特徵擷取前即排除端點標錯的可能。在本論文中，提出改善端點偵測錯誤的方法有兩種，分別為改善端點偵測錯誤接受和改善端點偵測錯誤拒絕。

除此之外，不佳的測試語音也會造成語者識別上的錯誤，因此須在語者識別之前濾除不佳的測試語音，在本論文中，提出兩種濾除不佳測試語音的方法，一種為設定拒絕門檻值，另一種為拒絕不完整的測試語音。其中設定拒絕門檻值的部份，根據特徵擷取出的平均音量、平均音高、平均清晰度和音框數四種特徵參數，分別使用各別特徵參數設定門檻值，以及高斯混合模型設定門檻值。另外，測試語音可能包含錄音內容不完整的情形，若能事先拒絕錄音不完整的測試語音，語者識別的錯誤率就能降低。如此一來，便能在尚未比對參考語音之前，濾除一些可能會造成辨識錯誤的測試語音，使得辨識結果的準確率提升。

Abstract

This thesis proposes several effective methods to improve our initial recognition rate in text-dependent speaker identification by Dynamic Time Warping(DTW). Our goal is to avoid some bad recordings causing false recognition. In order to develop an automatic text-dependent speaker identification system with screening mechanism for impropriate speech inputs, we define several useful speech features for our research and perform several experiments on feature matching methods. There are two parts in this thesis. The first is “Improvement on endpoint detection”, and the other is “The screening mechanism for impropriate speech inputs”.

“Improvement on endpoint detection” is a method that uses volume to identify the important part of an audio segment for achieving the minimum error . We use two approaches for achieving the minimum error, automatic noice segment removal and the shorter voiced sound segment reservation. The experimental results have shown that automatic noice segment removal is useful to improve the initial recognition rate, but the shorter voiced sound segment reservation is leading to decreased recognition rate.

“The screening mechanism for impropriate speech inputs” is another perprocessing in speaker identification by DTW. Through the screening mechanism rejecting bad recordings before identifying the speaker. We use two approaches for determining the threshold, Gaussian Mixture Model(GMM) and the combination of each feature’s threshold in the same rejection rate. In particular, the combination of each feature’s threshold is effective to improve the initial recognition rate. Besides, there is another approach to reject bad recordings, which is rejecting the incomplete speech content. Therefore the overall recognition can be improved.

致謝

這篇論文能夠完成，首先感謝張智星老師，在張智星老師的指導以及指引研究方向之下，才能順利將論文完成，並在老師兩年的教導帶領期間，學到很多語音相關的技術和實務經驗。研究上遇到不懂的地方，感謝實驗室的學長姐們不吝指教，並且給予詳盡的解說。感謝研究所兩年並肩作戰的同學們，一同加油打氣、互相勉勵，還有在我分身乏術之際，熱心幫忙完成一些事項。

除此之外，感謝一直陪伴在我身邊的家人與朋友們，給予我最大的支持和鼓勵，即使遭逢窘境也一路扶持我，讓我能夠順利完成學業。

目錄

圖目錄

表目錄

第1章　緒論

1.1. 研究動機

現今語音訊號的技術已日臻成熟，不論是語音辨識、語音合成或是語者辨認的技術，除了在電腦系統上的應用外，其應用在行動裝置上將會是未來的趨勢。

其中，語者辨認可藉由一個人的聲音特徵辨認出說話者，對於門禁系統、金融交易、犯罪偵防方面等，語者辨認皆可加以應用。而本論文主要基於語者辨認應用在門禁系統下所作的研究，其安全性顯得格外重要，若語者識別的辨識率能夠提高，而語者驗證的準確度也會相對地提升，因此如何提升語者識別的辨識率將是本論文研究的重點。

為了安全性的考量，使用文本相關的語音密碼，相較於非文本相關的語音密碼來說，驗證的可靠性較高。除此之外，當使用者確認身分時，週遭環境的雜訊和使用者聲音的音量、音高、說話的清晰度，以及語音密碼的長短和測試語音的完整性，皆有可能會影響辨識的準確度，若能在辨識之前，作語音訊號的前處理，並且事先拒絕可能造成辨識錯誤的語音密碼，讓使用者重新輸入語音密碼，當作語者識別的門檻，將可改善語者識別的準確度。

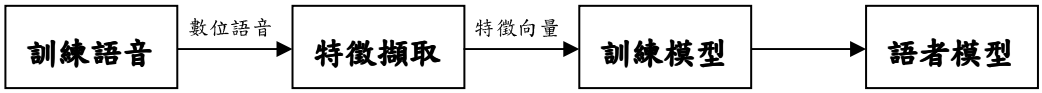
1.2. 語者辨認概述

語者辨認(Speaker Recognition)主要根據使用者聲音的特徵，辨別使用者的身分。在不同的應用層面上，分為語者辨別(Speaker identification)和語者驗證(Speaker verification)兩大類，基本架構如圖 1-1 所示。【1】【3】【4】【8】

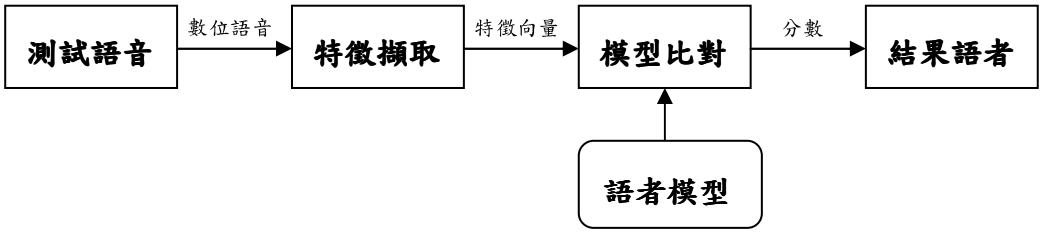
其中，語者識別的目標為測試語音從一群參考語者當中，藉由各種語者模型比對，挑選出最相近的語者。而語者驗證，則為一個驗證測試者所聲稱身分真偽的機制，同樣地根據參考語者的模型比對，決定接受測試者為宣稱者(Claimer)，或是拒絕測試者為冒充者(Imposter)。

至於語者辨認的測試者身分，一種為開放性測試(Open Set)和一種為封閉性測試(Close Set)。開放性測試的測試者，可能不在參考語者裡，為一個未知語者；密閉性測試的測試者，則為已知的參考語者中的一員。

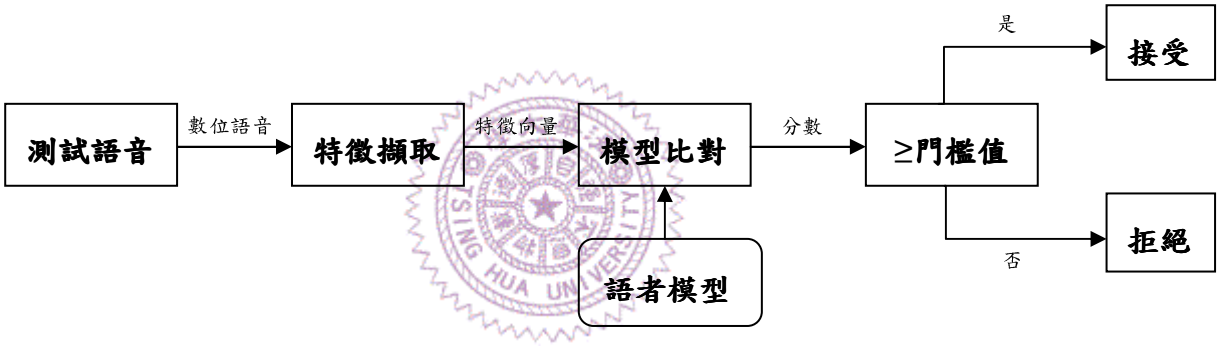
另外，根據測試語音的文字內容，可將語者辨認的模型分為文本相關(Text dependent)和非文本相關(Text independent)兩種。文本相關的測試語句必須限定，並且和參考語句相同；非文本相關的測試語句則不限定語句內容，可為任意語句，但非文本相關的測試語句的準確率相對較低。



(a) 語者模型訓練



(b) 語者識別



(c) 語者驗證

圖 1-1 語者辨認系統架構

1.3. 資料庫

本實驗語料使用 PDA 錄音檔，總共錄製 100 人。每人錄 10 個不同的華語語句，每句重複 3 次，並在不同時間錄製 2 次，2 次錄音時間間隔至少 1 個禮拜。錄音時間為 5 秒，取樣頻率為 16KHz，取樣大小為 16Bits，句子長度 4 個字到 8 個字不等，句子重複的情形如下表 1-1，其餘的句子皆不相同。本實驗中取第二次錄音檔作為參考語音，取第一次錄音檔作為測試語音。

錄音句子	重複次數
清華大學	7
中央大學	2
台北的天空	2
建國中學	2
音訊處理	2
程式設計	2
筆記型電腦	2
新年快樂	2
資料結構	2
電腦學習	2
養樂多綠茶	2

表 1-1 重複的錄音句子統計表

1.4.研究方向與主要成果

本論文以應用在行動裝置上的門禁系統為目標，基於此目標的語者識別系統，其安全性的考量格外重要，因此研究方向著重於如何提升語者識別的辨識率。

為了達到較高的準確度，使用文本相關的資料庫，不同於非文本相關的資料庫，將每個語者的聲學特徵訓練一個語者模型，由於語音內容為限定，所以利用動態時間扭曲(Dynamic Time Warping, DTW)的方式作語者識別。並且本論文致力於研究如何在辨識之前，將有雜訊的語音訊號經過前處理，以避免端點測試的錯誤，以及拒絕可能造成辨識錯誤的測試語句，進而得到辨識率的改善。

首先以動態時間扭曲的方法作語者識別，得到初始的辨識結果。根據初始的辨識結果，分析得知週遭環境雜訊所造成的端點偵測錯誤是主要影響辨識錯誤的原因。因此若能在辨識之前將端點偵測的錯誤排除，便可以提升辨識率。除了端點偵測會使辨識錯誤外，測試語句的音量大小、音高，以及聲音的清晰度和語句的長度，皆有可能會造成辨識錯誤。藉由上述四種特徵設定門檻值，當小於門檻值則拒絕測試語句，大於門檻值則接受測試語句，並進行語者識別比對。實驗設定兩種門檻值的方法，其一使用各別特徵參數設定門檻值，其二利用高斯混合模型設定門檻值。另外，測試語句的完整性也會影響辨識的準確度，因此判斷測試語音是否完整，也可避免語者識別辨識錯誤。

經過本論文提出的改善方法，實驗結果顯示，經過排除端點偵測錯誤後，可改善語者識別的辨識率，並且接近理想人工標端點的辨識率。進而加上使用拒絕門檻值的設定，和判斷測試語音是否完整，排除可能造成辨識錯誤的語句後，錯誤率明顯降低。

1.5. 章節摘要

本論文各章節安排如下:

第一章為緒論，闡明研究動機，語者辨認的概述、所使用的資料庫，以及研究的方向與主要的成果。

第二章將介紹語者辨認的基本技術，包含語者識別的相關研究，以及簡單描述本論文所使用的特徵參數，和使用動態時間扭曲作語者識別的方法。

第三章則是根據動態時間扭曲作語者識別的初始結果，提出降低語者識別錯誤率的方法，分別介紹改善端點偵測錯誤的方法，和設定門檻值的方法，以及拒

絕不完整的測試語音。

第四章為本論文的實驗結果與分析，從初始辨識結果分析，經過三個階段提升語者識別辨識率後，得到的實驗結果與比較，最後作錯誤分析。

第五章對本論文提出的方法作總結，和未來可能改進的方向。

第2章　語者辨認的基本技術

語者辨認大致上可分為訓練和辨認兩個階段，訓練階段主要由訓練語音特徵擷取後，建立各別的語者模型；辨認階段則是將訓練語音經過特徵擷取，與已建立好的語者模型比對，達到辨認的效果。由於語者的聲音特徵可代表語者的身分，因此特徵參數的選取相當重要，以及拒絕條件也必須根據聲音特徵作門檻值的設定。本章節主要介紹，語者辨認的相關研究、特徵選取的方法，以及本論文語者識別所使用的方法-動態時間扭曲。

2.1.　語者辨認的相關研究

目前已有的語者辨認的技術有【1】:向量化編碼(Vector Quantization, VQ)、動態時間扭曲(Dynamic Time Warping, DTW) 【2】【3】、類神經網路(Artificial Neural Networks, ANN) 【9】、隱藏式馬可夫模型(Hidden Markov Model, HMM)、最近鄰居分類法(Nearest Neighbor Rule, NNR) 【14】，以及高斯混合模型(Gaussian Mixture Model, GMM) 【6】【7】【8】【12】。

依據測試語句的文本相關性，挑選語者辨認系統的技術也不盡相同。一般來說，動態時間扭曲和類神經網路這兩種方法，對於文本相關的語者辨認系統較有效率;而向量化編碼、隱藏式馬可夫模型、最近鄰居分類法，和高斯混合模型這三種方法，則適用於非文本相關的語者辨認系統。

文本相關的語者辨認系統，所需要語者的語料較少，反觀非文本相關的語者辨認系統，則需要語者大量的語料，才能訓練出有效的模型。因此根據語料的多寡以及系統所需的條件，挑選適合語者辨認系統的技術。

2.2. 特徵參數擷取

每個語者皆有不同的聲音特徵，藉由特徵參數擷取，從語音中萃取出一些能夠區別聲音特質的參數值，以進行語者辨認的工作。由於人的聲音特性會隨著時間變化，屬於時變(time-varying)的訊號，而無法以線性非時變的方法分析長時域的語音訊號，因此使用短時域(short term)的頻譜特徵來代表聲學訊息，本論文在語者識別階段，採用目前最廣泛使用的特徵參數—梅爾倒頻譜參數(Mel-Frequency Cepstral Coefficients, MFCC)。至於在拒絕門檻值設定階段，採用四種特徵參數，分別為平均音量、平均音高、平均清晰度，和音框數。

2.2.1. 梅爾倒頻譜參數

梅爾倒頻譜參數(Mel-Frequency Cepstral Coefficients, MFCC) 【11】考量到人耳對不同頻率的感受程度，對於低頻的聲音感受程度較強，而高頻的聲音感受程度較弱的特性，因此在求取特徵參數時採取低頻多取、高頻少取的方式，其擷取方式說明如下:

為了消除發聲過程中聲帶和嘴唇的效應，以補償語音訊號受到發音系統所壓抑的高頻部分。因此首先將語音訊號作預強調(Pre-emphasis)的處理， 16KHz 的語音訊號 s(n)通過一個高頻濾波器: $H(z)=1-a*z^{-1}$ ，其中 a 介於 0.9 和 1.0 之間。若以時域的運算式來表示，預強調後的訊號 $s_2(n)$ 為 $s_2(n) = s(n) - a*s(n-1)$。

然後將預強調後的訊號取音框化(Frame blocking)，通常以 256 或是 512 個取樣點為一個觀測單位，涵蓋時間約為 20~30ms 左右。為了避免相鄰兩音框的變化過大，讓兩相鄰音框之間有一段重疊區域，此重疊區域包含了 M 個取樣點，

通常 M 的值約是 N 的一半或 1/3。

接著將每一個音框乘上漢明窗(Hamming windows)，以增加音框之間的連續性，假設音框化的訊號為 S(n), n = 0,...N-1。那麼乘上漢明窗後為 S'(n) = S(n)*W(n)，此 W(n) 形式如下：

$$W(n,\alpha) = (1-\alpha) - \alpha\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1$$

一般取 $\alpha = 0.46$。

由於訊號在時域(Time domain)上的變化通常很難看出訊號的特性，所以通常將它轉換成頻域(Frequency domain)上的能量分佈來觀察，不同的能量分佈，就能代表不同語音的特性。所以在乘上漢明窗後，每個音框還必需再經過快速傅利葉轉換(Fast Fourier Transform, FFT)，以得到在頻譜上的能量分佈。

將能量頻譜乘以一組 20 個三角帶通濾波器(Triangular Bandpass Filters)，求得每一個濾波器輸出的對數能量（Log Energy），且這 20 個三角帶通濾波器在「梅爾頻率」(Mel Frequency）上是平均分佈的，而梅爾頻率和一般頻率 f 的關係式如下：

$$\text{mel(f)} = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \text{ 或是 } \text{mel(f)} = 1125 * \ln\left(1 + \frac{f}{700}\right)$$

利用上述的 20 個對數能量 E_k 帶入離散餘弦轉換(Discrete cosine transform, DCT)，求出 L 階的 Mel- scale Cepstrum 參數，在此論文中取 L=12，即是 12 維的梅爾倒頻譜參數。離散餘弦轉換公式如下：

$$C_m = \sum_{k=1}^{N} \cos\left[m * (k-0.5) * \frac{\pi}{N}\right] * E_k, \ m = 1,2,\dots,L$$

其中 E_k 是由前一個步驟所算出來的三角濾波器和頻譜能量的內積值，N 是三

角濾波器的個數。由於之前作了 FFT，所以採用 DCT 轉換是期望能轉回類似 Time Domain 的情況來看，又稱 Quefrency Domain，其實也就是 Cepstrum。由於之前採用 Mel- Frequency 來轉換至梅爾頻率，因此稱之為 Mel-scale Cepstrum。

除此之外，在語音方面的應用上，通常會加上音框的對數能量（Log Energy），對數能量即是一個音框的音量，計算方式為一個音框內訊號的平方和，再取以 10 為底的對數值，再乘以 10。因此使用 1 個對數能量和 12 個倒頻譜參數，使得每一個音框基本的語音特徵為 13 維。

而在實際應用於語音辨識時，通常會再加上差量倒頻譜參數，以顯示倒頻譜參數對時間的變化。它的意義為倒頻譜參數相對於時間的斜率，也就是代表倒頻譜參數在時間上的動態變化，公式如下：

$$\Delta C_m(t) = \frac{\left[\sum_{\tau=-M}^{M} C_m(t+\tau)\tau\right]}{\sum_{\tau=-M}^{M} \tau^2}$$

其中 M 的值一般取 2 或 3。因此如果加上差量運算，就會產生 26 維的特徵向量；如果再加上差差量運算，就會產生 39 維的特徵向量。一般我們在 PC 上進行的語音辨識，就是使用 39 維的特徵向量，在此論文中沒有使用差量倒頻譜參數，只使用 13 維的特徵向量。

2.2.2. 平均音量

音量(Volume) 【11】代表聲音的強度、力度或是能量，可由一個音框內的訊號振幅大小來類比。對於語者識別而言，測試語音的音量大小會影響辨識的準確度，若音量太大則會造成爆音，音量太小則無法清楚判別所說的語句，而在端

點偵測時，也無法準確估測正確的起始點和終點。首先將端點偵測過後的所有語音片段，計算出每一個音框的音量，音量計算的方式為音框減去音框訊號的平均值，且取絕對值的總和。然後加總每一個音框的音量，除以所有語音片段的音框總數，即是平均音量。主要針對平均音量較小的測試語音，需要在語者識別前事先拒絕，排除可能造成辨識上的錯誤以提升辨識率。

2.2.3. 平均音高

音高(Pitch) 【11】代表聲音頻率的高低，此頻率指的是「基本頻率」（Fundamental Frequency），也就是「基本週期」（Fundamental Period）的倒數。若聲音穩定，即可觀察出基本週期的存在，在觀察音訊波形時，每一個基本週期的開始點，稱為「音高基準點」（Pitch Marks，簡稱 PM），PM 大部分是波形的局部最大點或最小點。

在本論文中，由於 ACF（Autocorrelation function）【11】的運算量少，因此選用 ACF 作為音高追蹤的方法。ACF 為一個時域的音高追蹤方法，且找出波形的局部最大點，假設 n 為取樣點總數，令一個音框中的取樣點 s(i)，i = 1,…,n， 其 ACF 的計算方式如下：

$$\mathrm{acf}(\tau) = \sum_{i=1}^{n-\tau} s(i)s(i+\tau)$$

其中 τ 代表以取樣點為單位，音框向右平移的距離，經過平移後的音框，與原本的音框重疊部分作內積，加總之後即是 ACF 值。

以音檔內容為“you”為例，計算出的 ACF 向量如圖 2-1 所示，根據 ACF 的向量發現 ACF 的最大值發生在第一點，若將此點的值設定為 0，則可找到第

二最高點的索引值，而第二最高點則是「ACF 音高點」，因此利用 ACF 音高點可自動找出音高。

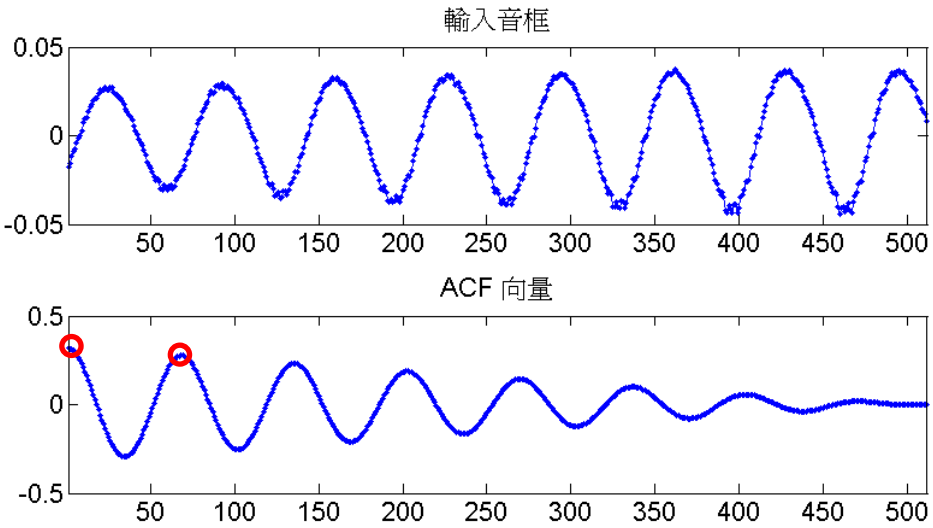


圖 2-1 ACF 示意圖一

將端點偵測後，測試語音的所有語音片段，計算每一個音框的音高，加總所有的音框的音高，除上所有音框的總數，即是平均音高。利用平均音高的特性，排除音高過高或音高過低，以致於辨識上不易的情形。

2.2.4. 平均清晰度

清晰度(Clarity) 【10】代表音訊片段的音高穩定度，在語者識別的應用上，能夠利用清晰度判別測試語音是否發音清晰，避免發音含糊不清的測試語音，造成辨識上的錯誤。在本論文中，所使用清晰度的計算方式，接續前面介紹的音高追蹤法，利用 ACF 向量找到第一最高點和第二最高點，且取出兩者的 ACF 高度(如圖 2-2)，以第二最高點的高度除以第一最高點的高度，即是清晰度，數學示表示如下：

$$\text{clarity} = \frac{h_1}{h_0}$$

而平均清晰度的計算方法，首先將端點偵測後的所有語音片段，每一個音框計算各自的清晰度，加總所有音框的清晰度，除以所有音框的總數，即是平均清晰度。

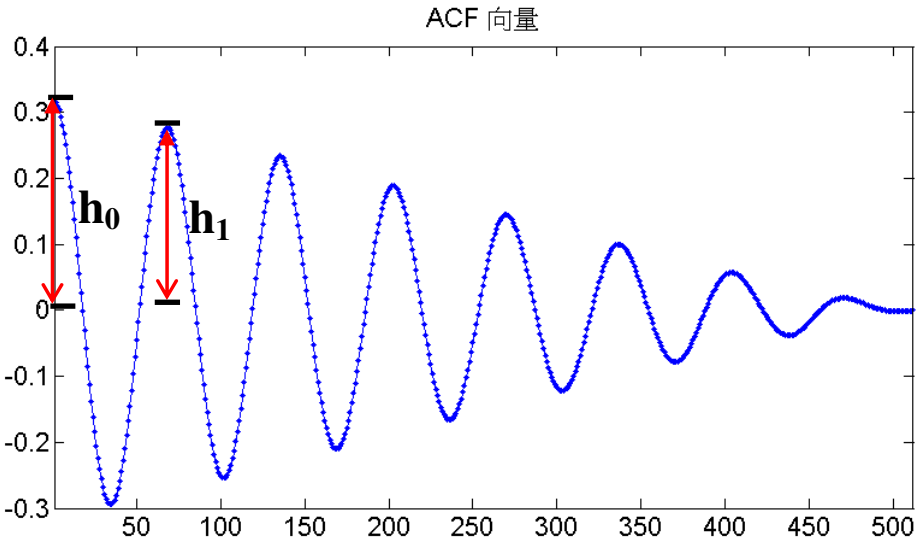


圖 2-2 ACF 示意圖二

2.2.5. 音框數

音框數目代表語音訊號的長短，對於語者識別而言，音框數目過少容易造成辨識上的錯誤，因為音框數目少相對地特徵向量也較小，在使用動態時間扭曲比對特徵向量時，不易辨識出正確的語者。另外，當測試語音完全沒有聲音時，透過音框數的判別，若音框數為 0 則可事先拒絕此測試語音。因此使用音框數目作門檻，拒絕低於音框數的門檻值的測試語音，即在語者識別前拒絕測試語音，使得錯誤率降低。

2.3. 動態時間扭曲

一般來說，文本相關的語者辨認大多使用動態時間扭曲(Dynamic Time Warping, DTW) 【11】【13】的方法，最主要藉由測試語句和參考語句相比對，即可計算距離找到最相近的語者，其好處是所需的語料少，辨識率就很高，但其缺點是計算時間較長。若基於門禁系統的考量下，設定語音密碼可增加驗證的可靠性，但語音密碼一旦被冒充者得知，或是冒充者欲用測錄的方式企圖登入，其安全性將受到考驗。

因此在本論文中使用動態時間扭曲作為兩特徵值比對的方法，此方法為動態規畫(Dynamic Programming, DP)的一種，藉由動態時間扭曲找出兩個向量之間的最近距離。假設有一個測試語音特徵向量 t，長度為 i，且參考語音特徵向量 r，長度為 j(如圖 2-3 所示)，期望找到一條路徑距離總和 D 最小，在此使用的動態時間扭曲必須端點對應，即頭對頭、尾對尾的方式，而點對點之間計算最短距離的方法為 0˚-45˚-90˚，若以計算 D(i, j) 距離為例，則考慮來自不同的三個方向的前一個點距離，找出 D(i-1, j)、D(i-1, j-1)、D(i, j-1)中最短距離者，且加上| t(i) – r(j) |的距離(數學式表示如下)。

$$D(i,j) = |t(i) - r(j)| + \min\left(D(i-1,j), D(i-1,j-1), D(i,j-1)\right)$$

計算兩點距離的方式使用最常見的歐基里德距離(Euclidean Distance)，在此使用的特徵向量為梅爾倒頻譜參數，根據測試語音和每一個參考語音比對得到的最短距離總和，挑選出與所有參考語音中最短的距離即辨別出的語句，進而得知測試語者的身分。

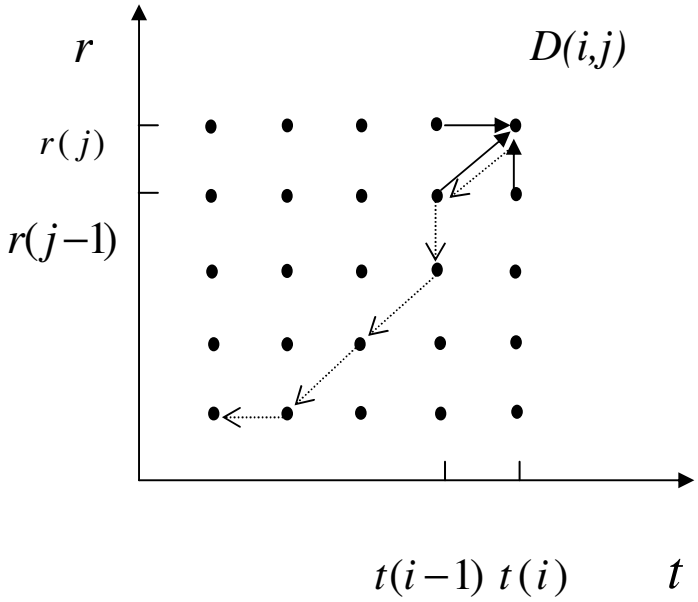


圖 2-3 動態時間扭曲比對示意圖

第3章　提升語者識別辨識率的方法

原先以動態時間扭曲作語者識別，其初始的辨識率已經很高，但是若將語者識別應用在門禁系統上，必須以安全性為考量，且提升語者識別的準確率。因此本論文的研究，著重於如何提升以動態時間扭曲作語者識別的初始辨識率，整體的實驗系統流程如圖 3-1 所示。在本論文中，提出兩大提升語者識別辨識率的方法：改善端點偵測錯誤、濾除不佳的測試語音，其中濾除不佳的測試語音的部分，又可分為設定拒絕門檻值，以及拒絕不完整的測試語音兩種，分別在 3.1、3.2、3.3 節中詳細介紹。

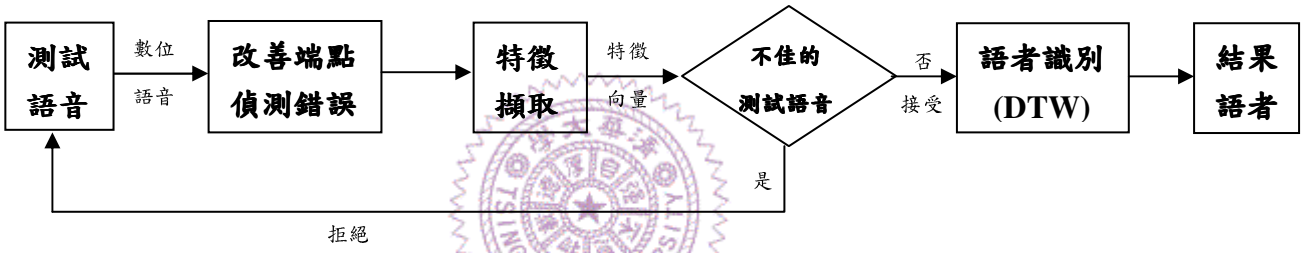


圖 3-1 本實驗系統的流程圖

3.1.　改善端點偵測錯誤

測試語音所取得的語音訊號，首先會經過音框化，在本論文中，選擇音量作為端點偵測的依據，藉由端點偵測標示出每一個音訊片段的起始點和結束點。若端點偵測標示錯誤，則會造成語者識別上的辨識錯誤，因此若能改善端點偵測的錯誤，便能提升語者識別的辨識率。

在本論文中，改善兩種端點偵測的錯誤，分別為錯誤拒絕(False rejection)，以及錯誤接受(False acceptance)，希望藉由這兩種方法提升語者識別的辨識率。

在端點偵測的過程中，當語音片段被誤判為靜音或雜訊片段，稱之為錯誤拒絕；相反地，當靜音或雜訊片段被誤判為語音片段，則稱之為錯誤接受。分別改善這兩種端點偵測錯誤，初步求得端點偵測的辨識率，得知經由改善端點偵測過後端點偵測的錯誤降低率。

首先介紹端點偵測錯誤接受的改善方法，由於本論文所使用的資料庫內含偶發性的雜訊，因此容易將雜訊片段誤判為語音片段，所以主要針對此種雜訊提出改善的方法。而此種雜訊的特性為時間短且與語音片段之間的間隔時間較長(如圖 3-2 所示)，根據這些特性提出改善端點偵測錯誤接受的方法，改善方法的步驟如下:

(1) 挑選出兩音訊片段之間的間隔時間最大者，且滿足兩音訊片段之間的間隔時間必須大於 0.48 秒。

(2) 判斷此兩個音訊片段何者時間較短，且滿足較短的音訊片段時間必須小於 1.6 秒，則視此音訊片段為雜訊。

(3) 重複(1)、(2)步驟，直到音訊片段之間的間隔時間小於 0.48 秒則停止。

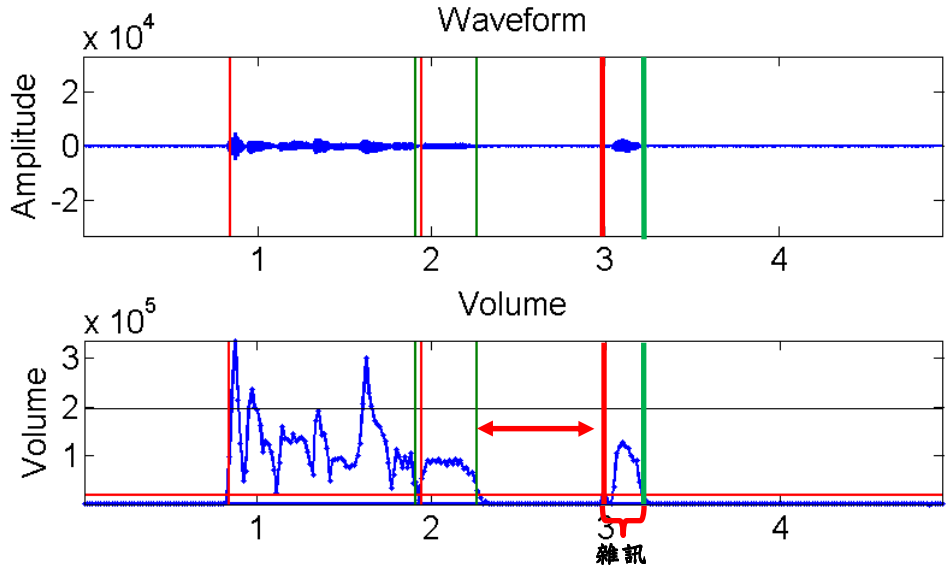


圖 3-2 端點偵測錯誤接受的示意圖

以圖 3-2 為例，第二和第三個音訊片段的時間間隔為最大，且滿足兩音訊片段之間的間隔時間大於 0.48 秒的限制條件，兩個音訊片段相比較之下，以第三個音訊片段時間最短，且滿足較短的音訊片段時間小於 1.6 秒的限制條件，因此將第三個音訊片段視為雜訊，而保留下來的音訊片段之間，沒有滿足時間間隔大於 0.48 秒的兩音訊片段，所以經過改善端點偵測錯誤接受後，保留了第一和第二個音訊片段，移除了第三個雜訊片段。

接著介紹端點偵測錯誤拒絕的改善方法，由於原始端點偵測的方法有一個限制條件，若音訊片段的時間小於 0.064 秒，則一律排除此音訊片段視為雜訊或靜音。但是有些時間較短的音訊片段為語音片段，以圖 3-3 為例，最後一個音訊片段為短促的語音片段，會被誤判為雜訊或靜音，因此造成保留下來的語音訊號不完整，使得語者識別容易辨識錯誤。從觀察得知，此種較短的語音片段與前一個語音片段之間的間隔時間較短，因此若要改善端點偵測的錯誤拒絕，須另外加入一個限制條件，當音訊片段的時間小於 0.064 秒時，判斷音訊片段之間的間隔時間是否小於 0.48 秒，若大於等於則視為語音片段，若小於則視為雜訊片段。經過判斷音訊片段之間的間隔時間後，範例圖 3-3 的第三個語音片段可保留下來。

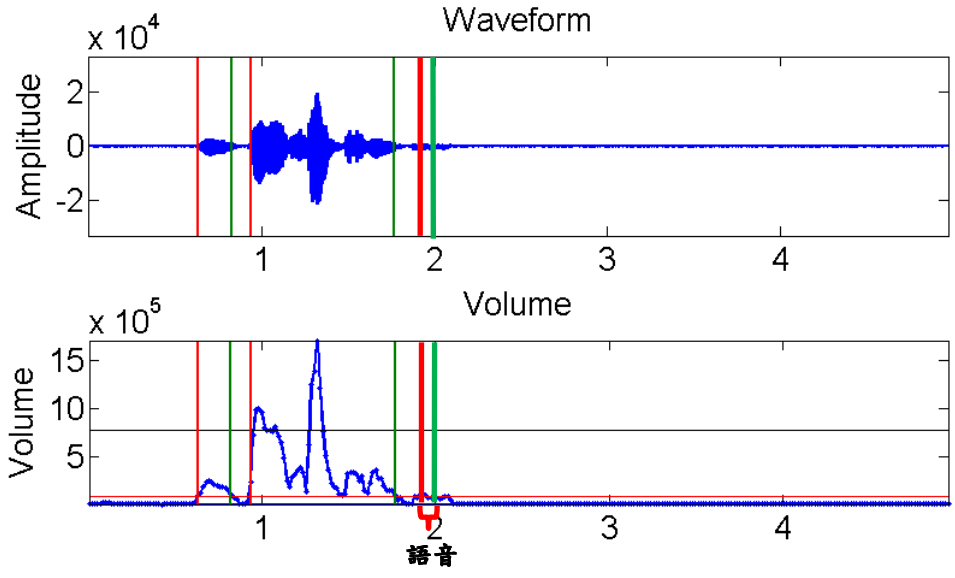


圖 3-3 端點偵測錯誤拒絕的示意圖

這兩種改善端點偵測錯誤的方法，為了得知對於本實驗所使用的資料庫，是否能夠真正地提升端點偵測的辨識率，須先單獨計算端點偵測的辨識率。因此將所有的測試語音和參考語音總共 6000 個音檔，使用人工標端點的方式，作為理想端點偵測的結果，然後和自動端點偵測的結果相比對。若端點偵測取出的語音片段總數不同，則視為端點偵測錯誤；若語音片段總數相同，還須考慮理想人工標示的端點，將理想人工標示的起始點和結束點與自動端點偵測的語音片段相比較，滿足理想人工標示的起始點和自動端點偵測的第一個語音片段起始點最接近，且理想人工標示的結束點和自動端點偵測的最後一個語音片段結束點最接近，以及容許誤差範圍在 0.128 秒以內，即視為端點偵測正確；其它情形則視為端點偵測錯誤。若改善的方法能夠提升端點偵測的辨識率，同樣地對於語者識別的辨識率提升也會有幫助。

3.2. 設定拒絕門檻值

測試語音經過特徵擷取過後，藉由擷取到的四種特徵參數：平均音量、平均音高、平均清晰度、音框數，設定拒絕的門檻值，希望在語者識別之前濾除一些不佳的測試語音，以降低語者識別的錯誤率。在本論文中，使用兩種設定拒絕門檻值的方法，分別為設定各別特徵參數的拒絕門檻值，以及利用高斯混合模型設定門檻值。而在本實驗中，以理想情況下設定拒絕門檻值的標準，即假設測試語音的端點偵測皆標示正確，因此採用理想人工標端點過後的測試語音，作為設定拒絕門檻值的標準。

3.2.1. 各別特徵參數的拒絕門檻值

在本論文中，擷取出的特徵參數總共有四種，而這四種特徵參數當中， 希望藉由拒絕門檻值的設定，分別拒絕掉平均音量較小、平均音高較低、平均清晰度較低，或是音框數較少的測試語音，因此藉由各自的特徵參數作濾除機制，找出主要影響語者識別辨識率的特徵參數組合，以及各自的拒絕比例和拒絕門檻值。

各別特徵參數設定的方法如下，使用理想人工標端點後的測試語音，將特徵擷取出四種特徵參數：平均音量、平均音高、平均清晰度、音框數，各自依照不同的拒絕比例，濾除可能不佳的測試語音。在本實驗中，測試語音的拒絕比例範圍選定在 0%至 30%，而不同的拒絕比例，分別對應到四種特徵參數的拒絕門檻值，通過拒絕門檻值的測試語音，可得到語者識別的辨識結果。根據各別特徵參數在不同拒絕比例下的語者識別辨識率，挑選出對於語者識別辨識提升有幫助的特徵參數，進而選擇測試語音的拒絕比例，設定此特徵參數的拒絕門檻值。測試語音經由各別特徵參數設定的拒絕門檻值，只要低於其中一種拒絕門檻值即拒絕，因此所有濾除的測試語音取聯集，最後得到濾除過後的語者識別辨識率。在第四章的實驗中，將會詳細說明各別特徵參數設定的拒絕比例、門檻值，以及濾除不佳的測試語音後語者識別的實驗結果。

3.2.2. 高斯混合模型

第二種設定拒絕門檻值的方式為使用高斯混合模型，挑選出最佳的特徵參數組合，再配合最適當的高斯混合數，訓練一個可辨別測試語音好壞的模型，因此需要一個辨識率較高的訓練模型，藉由測試語音與此訓練模型的對數似然比例

(log likelihood rate)，設定拒絕的門檻值，以判斷輸入的測試語音是否需要拒絕。

首先簡單介紹高斯混合模型(Gaussian Mixture Model, GMM)的基本理論，高斯混合模型是由數個高斯機率密度函數的加權平均所組成，而單一高斯機率密度函數的數學定義如下式:

$$b(x) = g(x; \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma_i|}} \exp\left(-\frac{(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}{2}\right)$$

假設 x 在 d 維的特徵空間中，其中 μ_i 代表高斯機率密度函數的中心點，Σ_i 則表示高斯機率密度函數的共變異矩陣（Covariance Matrix）。將各個高斯機率密度函數加權且加總後，即是高斯混合模型。其數學表示式如下：

$$p(x|\lambda) = \sum_{i=1}^{M} w_i \, b(x)$$

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \quad , i = 1, \ldots, M$$

其中以 λ 表示語者識別辨識結果的高斯機率密度函數參數，M 則表示高斯混合數，而 w_i 表示加權值，且滿足下列條件:

$$\sum_{i=1}^{M} w_i = 1$$

在此論文中，利用高斯混合模型中的各個高斯機率密度函數，模擬語者識別辨識結果的密度分布。

為了訓練出辨識率較高的高斯混合模型，將測試語音音檔萃取出的四種特徵參數，包含平均音量、平均音高、平均清晰度和音框數，透過特徵選取的方式，挑選出最佳的特徵參數組合。在本論文中，使用特徵選取的方式為耗盡式搜尋法(Exhaustive Search)，且選用的特徵選取的分類方式為 K 個最近鄰居分類法(K-nearest Neighbor Rule, KNNR)，分類的類別即是測試語音經過理想人工標端點後語者識別的辨識結果，且使用 Leave-One-Out(LOO)的方式計算 KNNR 模型的

辨識率。因此將所有的測試語音中(共 3000 個音檔)，取一個測試語音作為 KNNR 的測試資料，其它 2999 個測試語音作為 KNNR 的訓練資料，每一個測試資料計算出 k 個距離最近的訓練資料點，且距離計算的方法使用最常用的歐基里德距離(Euclidean Distance)，然後將 k 個訓練資料點對應的類別進行投票，即代表此筆測試資料被分到的類別。在本實驗中，取 k 的個數為 1，則只須找到一個與測試語音最相近的訓練資料點，再將此測試語音分到訓練資料點的同一類別。由於在此使用的特徵參數有四種，因此總共有 15 種特徵參數組合的 KNNR 訓練模型，從 15 個 KNNR 模型中，挑選出辨識率最高的特徵參數組合，即為特徵選取後的最佳特徵參數組合。

最後利用最佳特徵參數組合訓練一個高斯混合模型，所使用的資料為所有的測試語音(共 3000 個音檔)經過理想人工標端點後，語者識別的辨識結果，取奇數筆資料為高斯混合模型的訓練資料，另外偶數筆資料為高斯混合模型的測試資料，總共各 1500 個音檔。接著調整高斯混合的個數，使得高斯混合模型的辨識率達到最高，將此高斯混合模型作為判斷測試語音好壞的濾除機制，且利用測試語音和此訓練模型的對數似然比例設定拒絕門檻值，拒絕掉對數似然比例較低的測試語音，當作此測試語音容易被辨識錯誤。在本實驗中，為了與各別特徵參數設定門檻值的實驗結果相比較，選取相同的拒絕比例下，所對應的對數似然比例作為拒絕門檻值。因此所有的測試語音經過高斯混合模型的拒絕門檻值後，保留下來的測試語音可得到語者識別的辨識率，即可降低語者識別的錯誤率。

3.3. 拒絕不完整的測試語音

本實驗所使用的資料庫當中，包含錄音不完整的測試語音，而在真實使用在門禁系統上，也會遇到相同的問題，因此若能在語者識別之前濾除不完整的測試

語音，即可降低語者識別的錯誤率。不完整測試語音的情形包含在語句的前、中、後段少字，在本論文中，主要針對語句前或後段少字的情形(如圖 3-4、圖 3-5)，提出事先濾除不完整測試語音的方法。所使用的方法為判斷第一個音框或最後一個音框是否有音量，若有音量則視為測試語音不完整，若無音量則視為測試語音完整，作為拒絕不完整測試語音的機制。

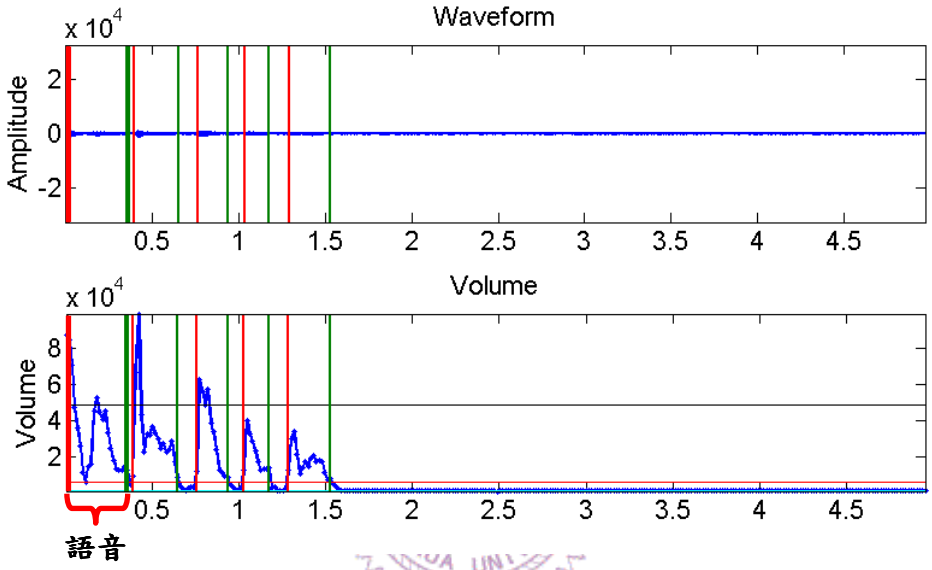


圖 3-4 測試語音前段少字的示意圖

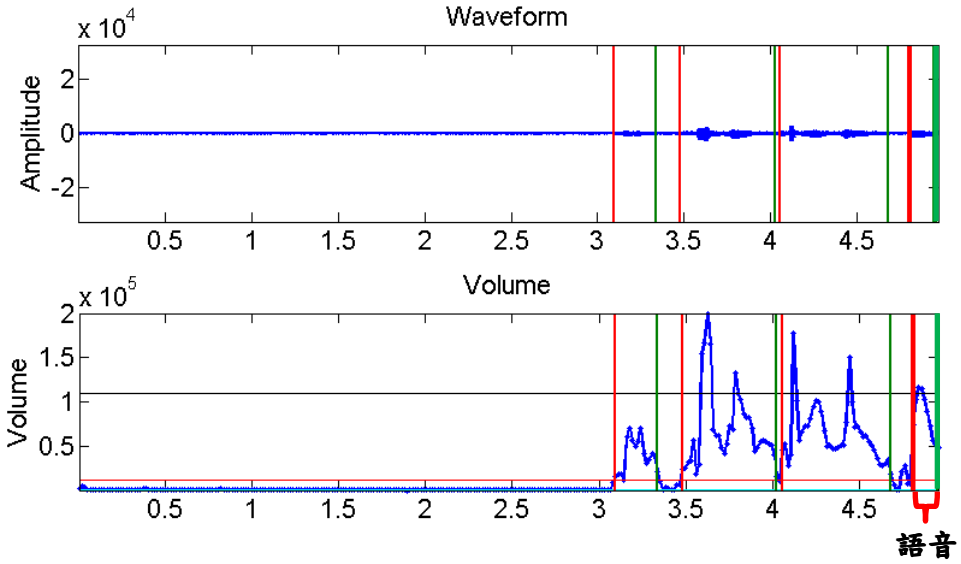


圖 3-5 測試語音後段少字的示意圖

此方法除了能夠濾除不完整的測試語音外，對於改善端點偵測錯誤後，在測試語音前後依然有雜訊的音檔(如圖 3-6)，或是測試語音的錄音內容為周圍環境的雜訊(如圖 3-7)，皆可藉由此方法濾除不佳的測試語音，避免發生語者識別辨識錯誤的情形。

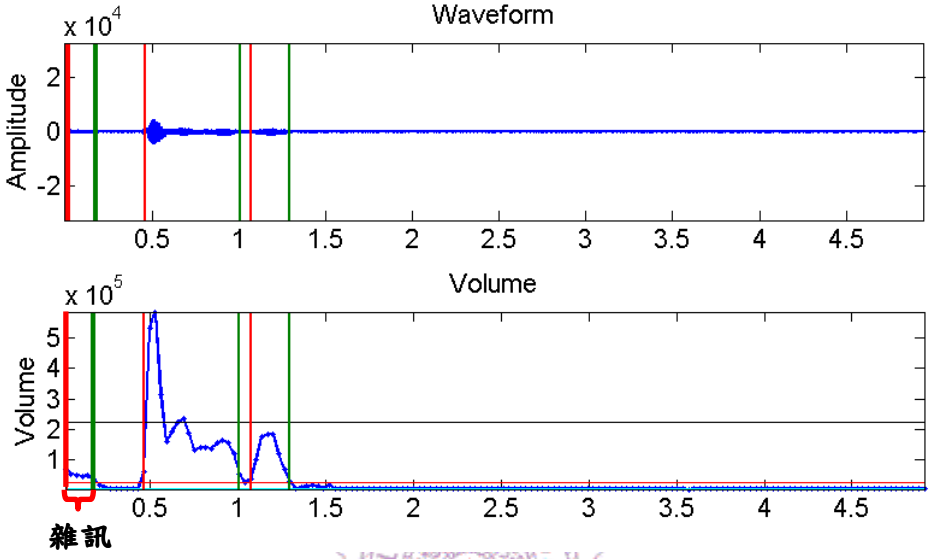


圖 3-6 改善端點偵測錯誤後，測試語音前仍有雜訊的示意圖

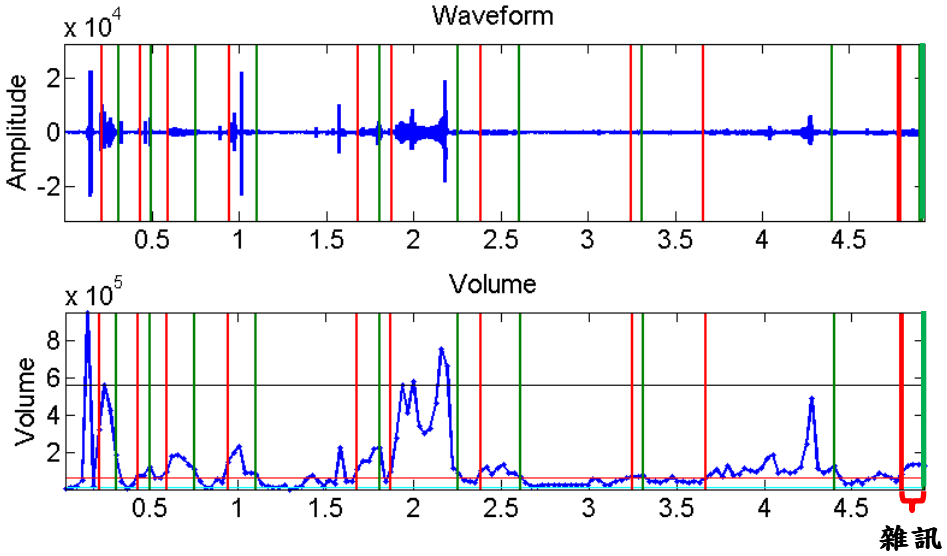


圖 3-7 測試語音內容為環境雜訊的示意圖

第4章　實驗結果與分析

4.1.　初始語者識別的實驗結果與分析

本實驗所使用的資料庫為 PDA 錄音檔，測試語音和參考語音的資料分配與錄音條件如下表 4-1 所示。

	錄音檔	錄音人數	每句錄音時間	每人錄音句子	每句重複次數	錄音檔總數
測試語音	第一次	100 人	5 秒	10 句	3 次	3000 個
參考語音	第二次	100 人	5 秒	10 句	3 次	3000 個

表 4-1 資料庫

資料經過特徵擷取後，分別取用兩種不同資料型態的特徵值，一種為整數型態的特徵值，為了日後能夠應用在 PDA 上，先在電腦上模擬測試；另一種為浮點數型態，主要使用在電腦上。測試語音經過特徵擷取後，使用梅爾倒頻譜參數作動態時間扭曲的特徵比對，得到初始語者識別的辨識結果。初始辨識結果如下：整數型態的辨識率為 93.10%，浮點數型態的辨識率為 96.53%。由於整數型態的精準度不如浮點數型態，所以在計算距離的部份會有誤差，以致於整數型態的辨識率較低。

以浮點數型態的辨識結果作錯誤分析，辨識錯誤的音檔總共有 103 個，包含端點偵測錯誤的音檔有 70 個，音量太小的音檔有 21 個，錄音內容不完整的音檔有 7 個，沒錄到正確句子的音檔有 2 個，以及正常的音檔有 3 個。由此得知大多數辨識錯誤的原因為端點偵測錯誤，因此為了改善端點偵測的錯誤，作了改善端

點偵測錯誤接受和拒絕的實驗。另外，若能在語者識別之前，濾除音量太小或是錄音內容不完整的音檔，便可降低語者識別的錯誤率，因此根據第三章所提出的濾除方法作了以下實驗。

4.2. 改善端點偵測錯誤的實驗結果

首先分別將兩種改善端點偵測錯誤的方法，初步計算出改善後端點偵測的辨識率，得知套用在資料庫上是否能夠改善端點偵測的辨識率，再使用此方法作語者識別的實驗。

4.2.1. 改善端點偵測錯誤接受的實驗結果與分析

將所有的測試語音和參考語音(共 6000 個音檔)，經過自動端點偵測過後，以及改善端點偵測錯誤接受後，與理想人工標端點的結果相比對，得到端點偵測的辨識率(如下表 4-2 所示)。

原始端點偵測	改善端點偵測錯誤接受	錯誤降低率
91.05%	**92.18%**	12.63%

表 4-2 端點偵測的實驗結果-改善端點偵測錯誤接受

由表可知，經過改善端點偵測錯誤接受之後，端點偵測的錯誤降低率達 12.63%，以端點偵測而言，此方法確實可改善端點偵測的辨識率。接著利用改善端點偵測錯誤接受的方法作語者識別的實驗，實驗結果如表 4-3 所示，不論是整

數型態或是浮點數型態的特徵參數，所得到的語者識別辨識率皆有提升，語者識別的錯誤降低率分別可達 15.51%和 47.26%，並且與理想人工標端點的語者識別結果相比較，發現改善端點偵測錯誤接受後的辨識率已相當接近理想值。由於整數型態的特徵參數精準度不如浮點數的特徵參數，所以語者識別的錯誤降低率也相對較低。

	初始結果	理想人工標端點	改善端點偵測錯誤接受	錯誤降低率
整數型態	93.10%	95.33%	**94.17%**	15.51%
浮點數型態	96.53%	98.83%	**98.17%**	47.26%

表 4-3 語者識別的實驗結果-改善端點偵測錯誤接受

以浮點數型態的實驗結果而言，經過改善端點偵測錯誤接受後，仍辨識錯誤的音檔有 55 個，其中端點偵測錯誤的音檔有 14 個，雜訊緊鄰在語音片段前後的音檔有 8 個(如圖 4-1、圖 4-2)，另外，將語音片段切掉而保留雜訊片段的音檔有 2 個(如圖 4-3)，除此之外，語音片段過短而被切除的音檔有 4 個(如圖 4-4)。

由於本實驗中改善端點偵測錯誤接受的方法，以設定兩音訊片段之間的時間間隔和較小音訊片段的時間為判斷條件。雖然可以將大部分音檔的雜訊去除，但仍有一些音檔的雜訊是緊鄰語音片段前後(如圖 4-1、圖 4-2)，因此會被誤判為語音片段而保留下來。另外，有些音檔的雜訊片段比語音片段的時間長，所以可能會判斷錯誤，反而將語音片段去除，把雜訊片段保留下來 (如圖 4-3)。除此之外，在此階段尚未改善端點偵測錯誤拒絕的部分，所以仍有語音片段過短而被切除的音檔(如圖 4-4)。

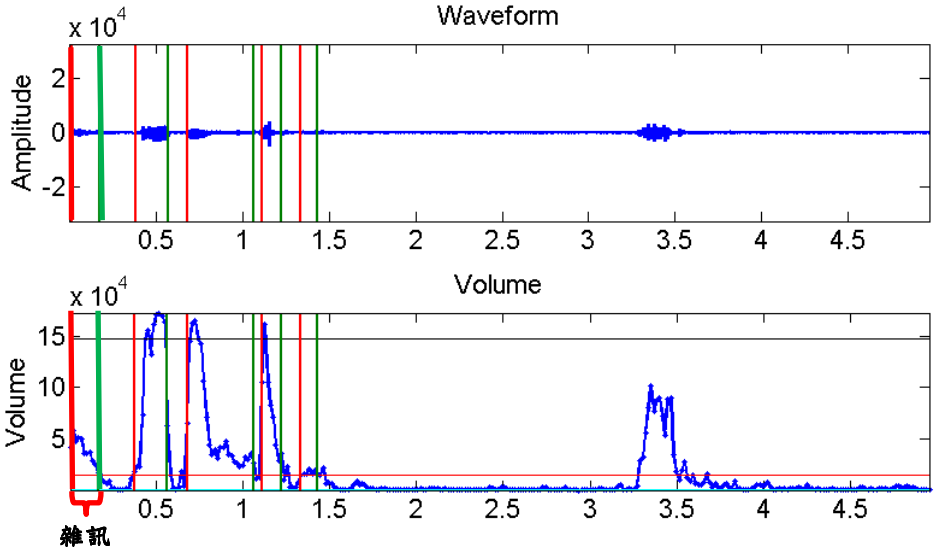


圖 4-1 語音片段前緊鄰雜訊的示意圖

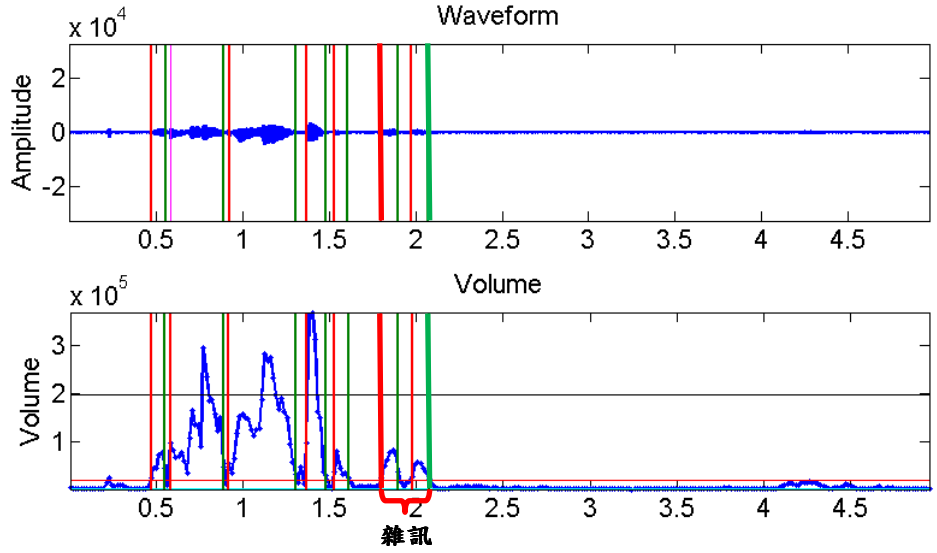


圖 4-2 語音片段後緊鄰雜訊片段的示意圖

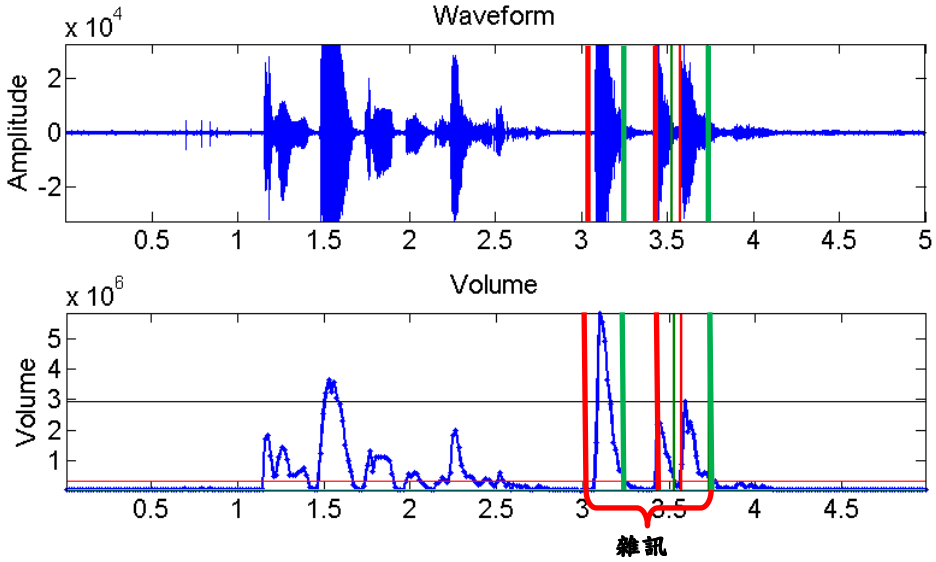


圖 4-3 切錯成雜訊片段的示意圖

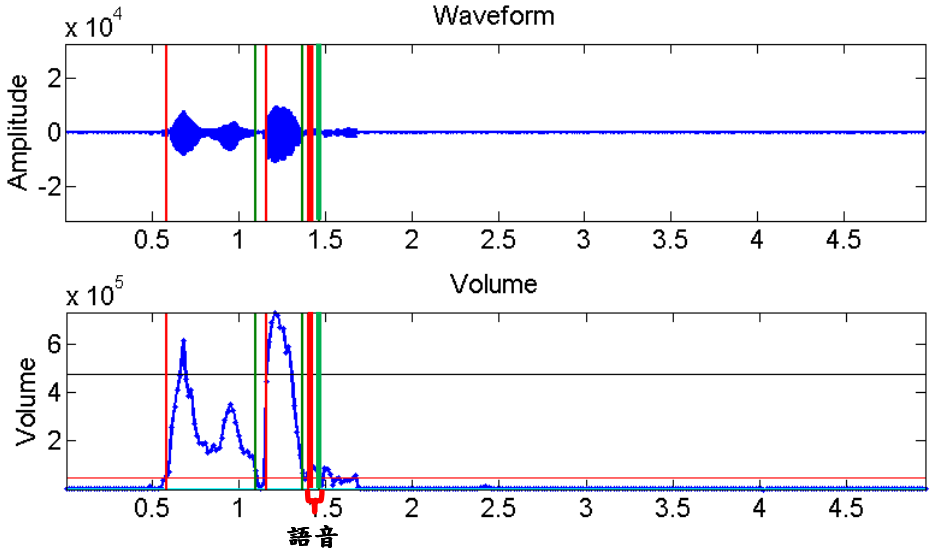


圖 4-4 最後一個語音片段過短的示意圖

4.2.2. 改善端點偵測錯誤拒絕的實驗結果與分析

同樣地，將所有的測試語音和參考語音(共 6000 個音檔)，使用改善端點偵測錯拒絕的方法後，與理想人工標端點的結果相比較，初步求得端點偵測的辨識率，實驗結果如下表 4-4 所示。

原始端點偵測	改善端點偵測錯誤接受	改善端點偵測錯誤拒絕	錯誤降低率
91.05%	92.18%	**89.48%**	-17.54%

表 4-4 端點偵測的實驗結果-改善端點偵測錯誤拒絕

由實驗數據發現，經過改善端點偵測錯誤拒絕的方法後，反而比原始端點偵測的錯誤率還高，從端點辨識錯誤的音檔中作分析，大多數過短的音訊片段為偶發性的雜訊(如圖 4-5)，而只有少數過短的音訊片段為語音。因此使用此方法後，反而將大部分音檔的雜訊保留下來，使得端點偵測的辨識率下降。若以此修改後的端點偵測作語者識別，得到的實驗結果如表 4-5，不論是整數型態或是浮點數型的特徵參數，語者識別的錯誤率皆提高，比語者識別的初始辨識率還低，因此不考慮改善端點偵測錯誤拒絕的方法，只採用改善端點偵測錯誤接受的方法，接續以下的實驗。

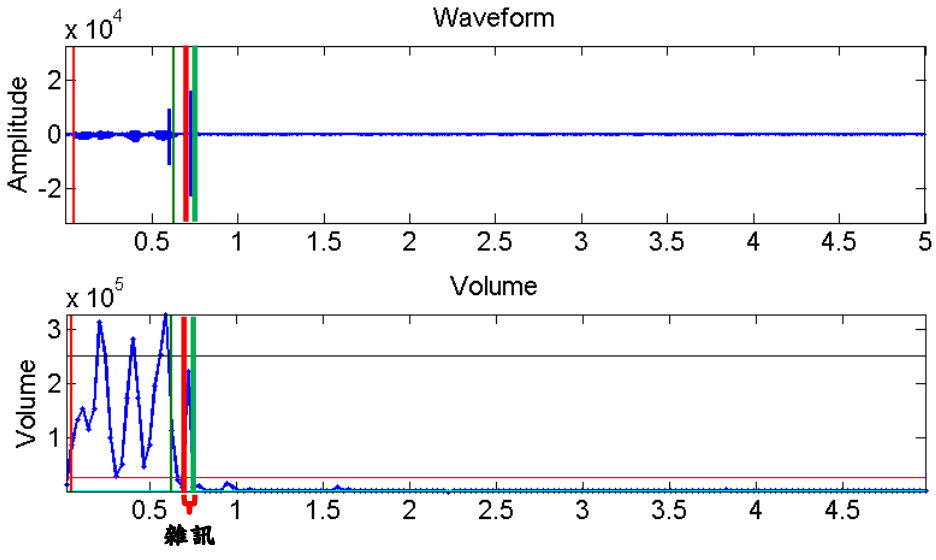


圖 4-5 緊鄰在語音片段後的小雜訊片段示意圖

	初始結果	理想人工標端點	改善端點偵測錯誤接受	改善端點偵測錯誤拒絕	錯誤降低率
整數型態	93.10%	95.33%	94.17%	**91.97%**	-16.38%
浮點數型態	96.53%	98.83%	98.17%	**95.80%**	-21.04%

表 4-5 語者識別的實驗結果-改善端點偵測錯誤拒絕

4.3. 設定拒絕門檻值的實驗結果

在本實驗中，以浮點數型態的特徵參數，且經過理想人工標端點後的語者識別結果，作為設定門檻值的依據，再分別利用各別特徵參數設定門檻值，以及高斯混合模型設定門檻值的方式，濾除不佳的測試語音，然後保留下來的測試語音作語者識別的實驗。

4.3.1. 各別特徵參數設定門檻值的實驗結果

測試語音在拒絕率 0%至 30%之下，且拒絕率以 1%為單位，分別使用四種特徵參數設定門檻值，所得到的語者識別結果如圖 4-6。由圖可知，只有使用平均音量作為門檻時，隨著拒絕率的提高，語者識別的辨識率有明顯的提升，得知平均音量對於語者識別的辨識率影響甚大，因此首先考量平均音量的門檻值設定。除了平均音高作門檻時，不論測試語音的拒絕率為何，所得到語者識別的辨識率皆比 0%時的辨識率低，因此不考慮平均音高對於語者識別辨識率的影響。另外，音框數和平均清晰度作為門檻時，測試語音在拒絕率 10%以內，語者識別的辨識率有小幅度的提升，所以接續考慮音框數和平均清晰度兩種的門檻值設定。

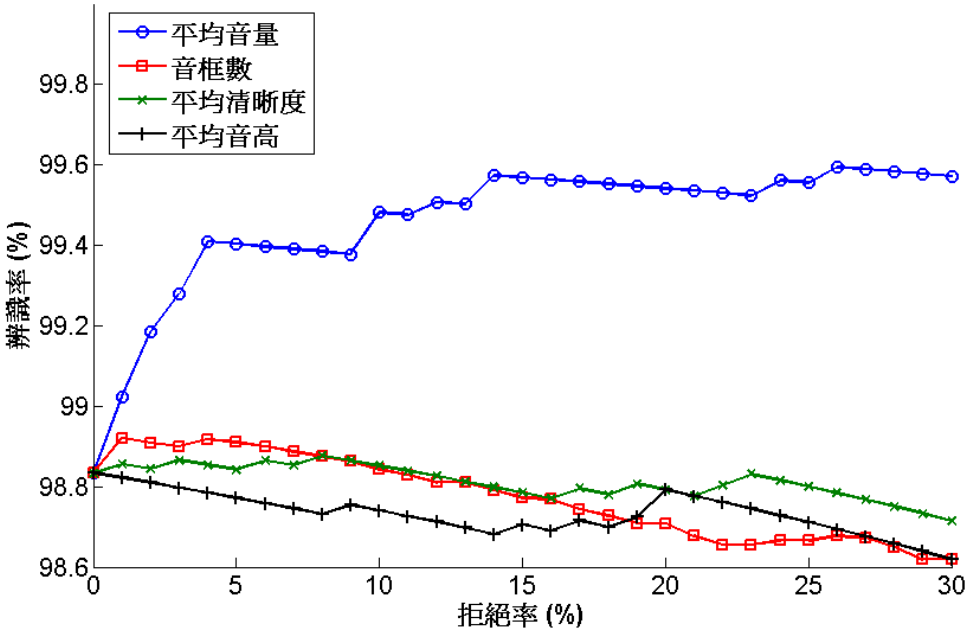


圖 4-6 拒絕率與語者識別的辨識率關係圖

由圖 4-6 可知，以平均音量作門檻時，雖然測試語音的拒絕率在 26%至 30%的辨識率接近 100%，但是若拒絕率太高，對於實際測試者的可用性不高，因此挑選拒絕率 10%以內語者識別辨識率最高者。實驗結果顯示，拒絕率 4%時可達到最佳的辨識率，為了增加精確度，再將拒絕率縮小範圍在 3%至 5%，且以 0.1%為單位，得到的語者識別結果如圖 4-7。由圖 4-7 可知，在拒絕率 3.9%可達到最佳的辨識率(99.4103%)，所以平均音量的門檻值選在拒絕率 3.9%時，得到平均音量的門檻值為 90213.37。以理想人工標端點後的 3000 個測試語音而言，低於平均音量門檻值的音檔有 117 個，因此總共拒絕 117 個音檔。

觀察圖 4-6 得知，以音框數作門檻時的辨識提升率比平均作門檻時高，因此接著考量音框數的門檻值設定。同樣地，挑選拒絕率 10%以內語者識別辨識率最高者，當拒絕率 1%時，辨識率最高可達 98.9211%，再將拒絕率縮小範圍在 0%至 2%，且以 0.1%為單位，得到語者識別結果如圖 4-8。由圖 4-8 可知，在拒絕率 0.6%時，可達到最高的辨識率(98.9265%)，所以音框數的門檻值設定在拒絕

率 0.6%，得到音框數的門檻值為 75。若以理想人工標端點後的 3000 個測試語音而言，低於音框量門檻值的音檔有 18 個，因此總共拒絕 18 個音檔。

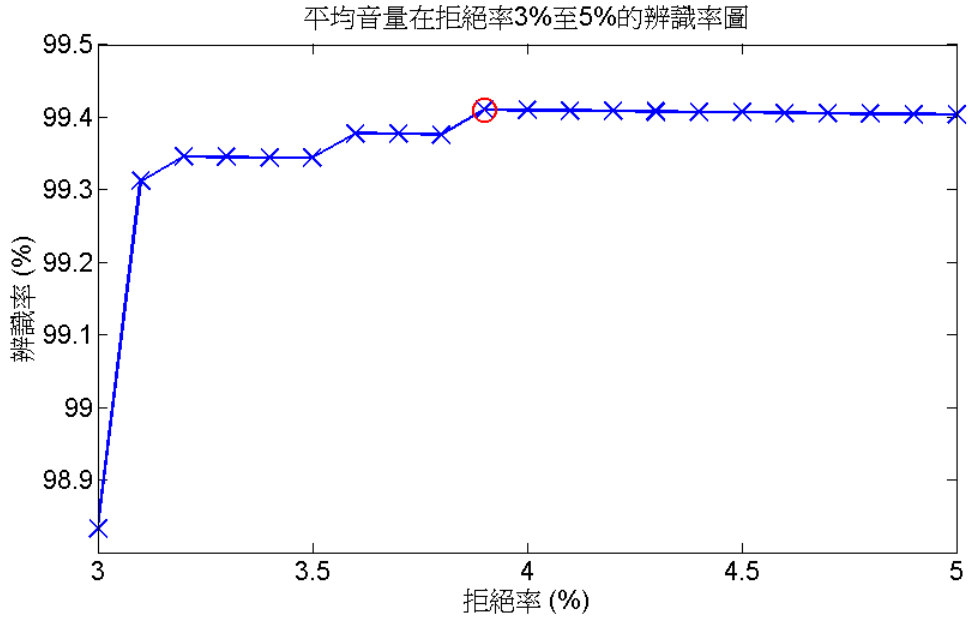


圖 4-7 平均音量-在拒絕率 3%至 5%的辨識率圖

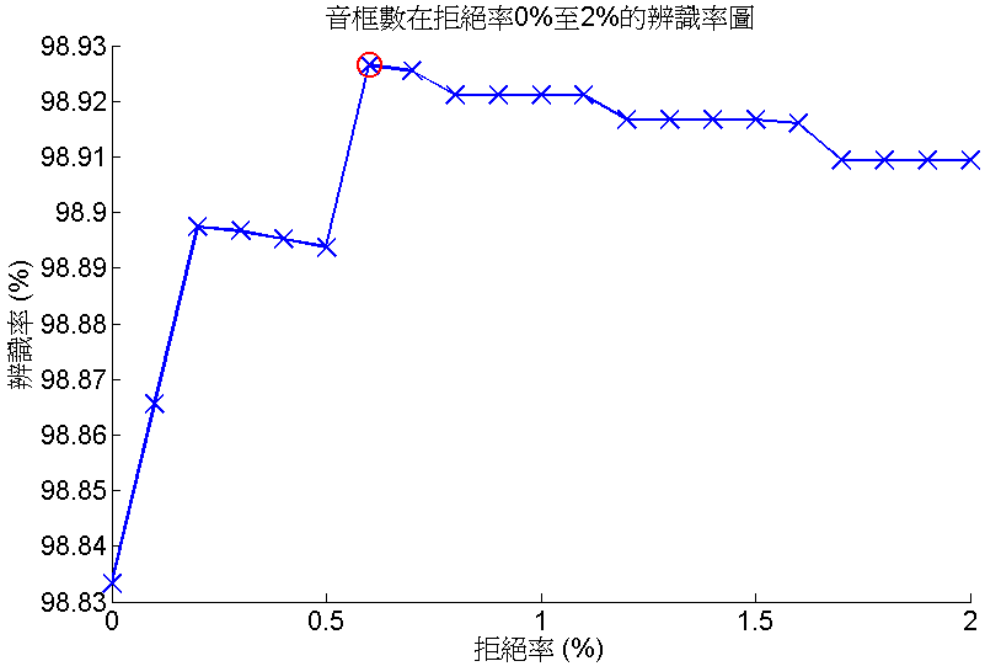


圖 4-8 音框數-在拒絕率 0%至 2%的辨識率圖

另外，由圖 4-6 可知，以平均清晰度作門檻時，在拒絕率 10%以內的辨識提升率僅在 0%至 0.02%左右，因此先以平均音量和音框數作門檻，得到語者識別的實驗結果(見表 4-6、4-7)。以理想人工標端點後的 3000 個測試語音而言，在兩道門檻值設定下，全部拒絕率佔 4.53%，總共拒絕 136 個音檔，浮點數型態的辨識率提升至 99.51%，而錯誤降低率可達 85.88%。而使用改善端點偵測錯誤接受的測試語音，全部拒絕率佔 5.77%，總共拒絕 173 個音檔，浮點數型態的辨識率提升至 99.19%，而錯誤降低率可達 76.66%，相當接近理想人工標端點的實驗結果。

	初始結果	(1)理想人工標端點	(1)+(2)兩道門檻值設定	拒絕比例	錯誤降低率
整數型態	93.10%	95.33%	**97.31%**	4.53%	61.01%
浮點數型態	96.53%	98.83%	**99.51%**	4.53%	85.88%

表 4-6 理想人工標端點後且兩道門檻值設定下的實驗結果

	初始結果	(1)改善端點偵測錯誤接受	(1)+(2)兩道門檻值設定	拒絕比例	錯誤降低率
整數型態	93.10%	94.17%	**96.64%**	5.77%	51.30%
浮點數型態	96.53%	98.17%	**99.19%**	5.77%	76.66%

表 4-7 改善端點偵測錯誤後且兩道門檻值設定下的實驗結果

若前面兩道門檻值設定後，再加上平均清晰度作門檻，且拒絕率設定在 10%以內，得到語者識別的辨識結果如圖 4-9。由圖 4-9 得知，平均清晰度作門檻時，拒絕率 6%可達到最高的總體辨識率。若縮小範圍在拒絕率 5%至 7%之下，且以 0.1%為單位，語者識別的辨識結果如圖 4-10。由圖 4-10 可知，依然在拒絕率 6%

時，可達最高的總體辨識率(99.5902%)，所以平均清晰度的門檻設定在拒絕率 6%時，所對應到平均清晰度的門檻值為 39.27。若理想人工標端點後的 3000 個測試語音而言，總共拒絕 180 個平均清晰度較低的音檔。

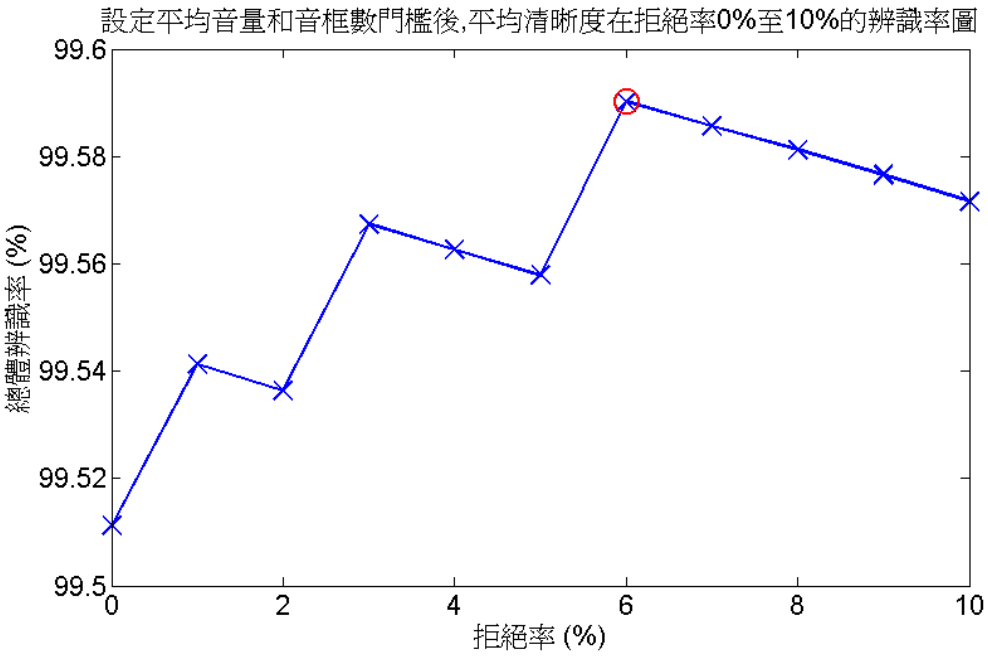


圖 4-9 前兩道門檻設定後，平均清晰度在拒絕率 0%至 10%的辨識率圖

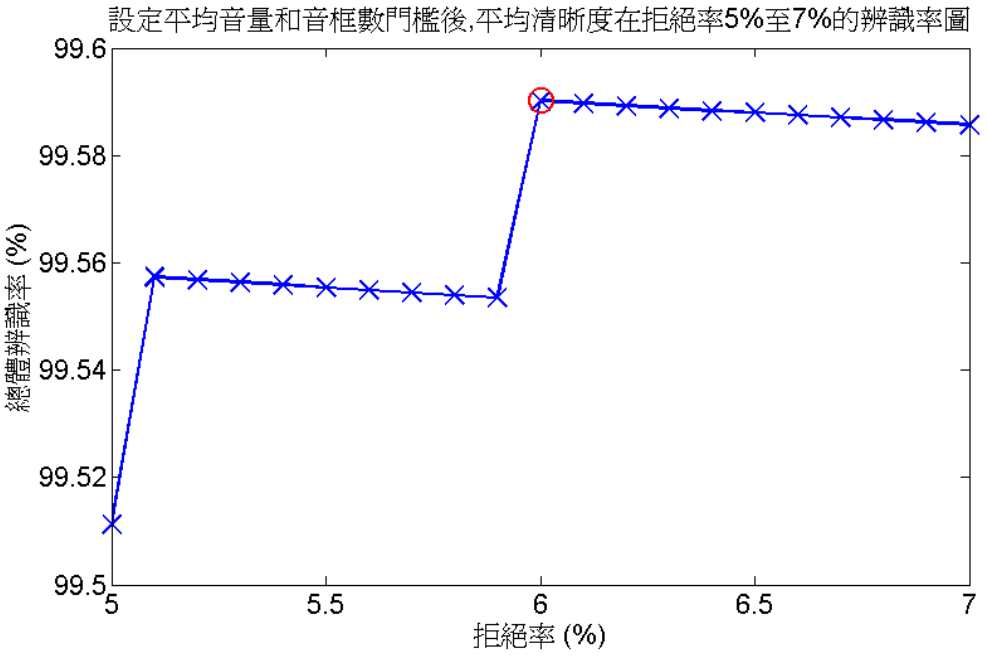


圖 4-10 前兩道門檻設定後，平均清晰度在拒絕率 5%至 7%的辨識率圖

由於此三種特徵參數所設定的門檻值，測試語音只要低於其中一種門檻即拒絕，所以所有拒絕的音檔皆取聯集，而在前兩道門檻設定之下，加上平均清晰度作門檻，且拒絕率取 0%至 10%，得到總體辨識率和全部拒絕率，實驗結果如圖 4-11。實驗結果顯示，平均音量的門檻設定在拒絕率 4%，且音框數的門檻設定在拒絕率 0.6%，加上平均清晰度的門檻設定在拒絕率 6%之下，此時語者識別的辨識率最高(99.5902%)。而全部拒絕率佔 10.53%，以測試資料 3000 個音檔而言，總共拒絕 316 個音檔。若利用此三道門檻值拒絕測試語音，語者識別的實驗結果分別見表 4-8、4-9。

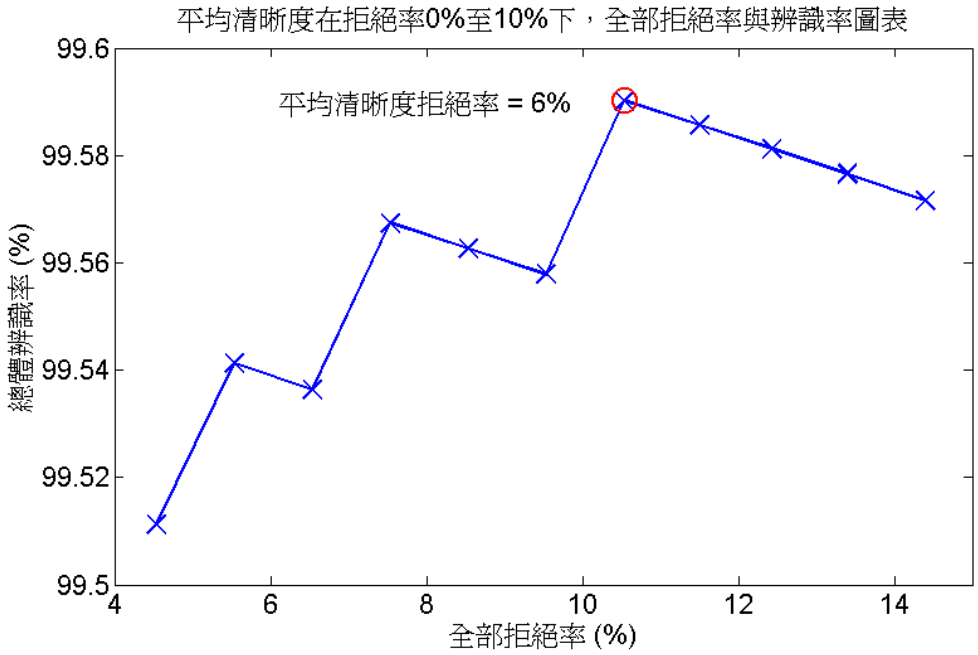


圖 4-11 平均清晰度在拒絕率 0%至 10%下，全部拒絕率與辨識率圖

	初始結果	(1)理想人工標端點	(1)+(2)兩道門檻值設定	(1)+(3)三道門檻值設定	拒絕比例	錯誤降低率
整數型態	93.10%	95.33%	97.31%	**97.28%**	10.53%	60.58%
浮點數型態	96.53%	98.83%	99.51%	**99.59%**	10.53%	88.18%

表 4-8 理想人工標端點後且三道門檻值設定下的實驗結果

	初始結果	(1)改善端點偵測錯誤接受	(1)+(2)兩道門檻值設定	(1)+(3)三道門檻值設定	拒絕比例	錯誤降低率
整數型態	93.10%	94.17%	96.64%	**96.57%**	11.67%	50.29%
浮點數型態	96.53%	98.17%	99.19%	**99.17%**	11.67%	76.08%

表 4-9 改善端點偵測錯誤接受後且三道門檻值設定下的實驗結果

由表 4-8、4-9 的實驗結果得知，只有理想人工標端點後的測試語音，在浮點數型態的特徵參數下，加上平均清晰度的門檻值設定辨識率有提升外，其餘三種實驗結果辨識率皆下降。因此若以改善端點偵測錯誤接受的測試語音為主，則只使用平均音量和音框數設定門檻值的辨識結果較佳，所以選用兩道門檻值設定作為各別特徵參數設定門檻值的方法，接續以下實驗。

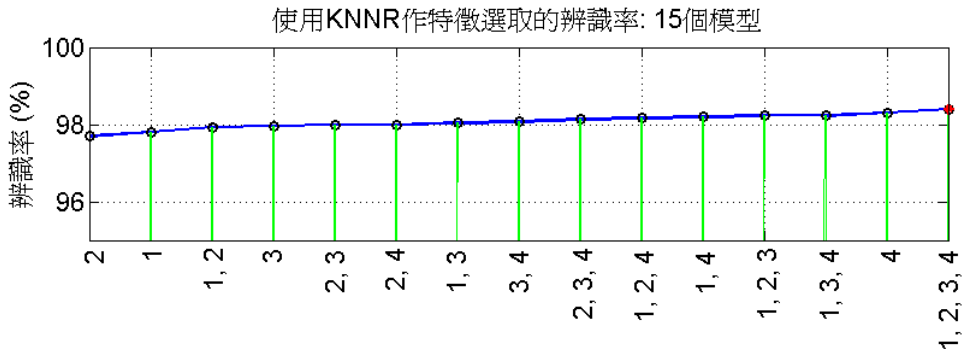
4.3.2. 高斯混合模型設定門檻值的實驗結果

將測試語音經過理想人工標端點後的辨識結果，當作高斯混合模型的訓練和測試資料，資料分配如下表 4-10 所示。取用所有測試語音的基數筆資料作為高斯混合模型的訓練資料，而偶數筆資料作為高斯混合模型的測試資料。

	錄音檔	錄音檔總數	每句錄音時間
訓練資料	第一次(奇數筆)	1500 個	5 秒
測試資料	第一次(偶數筆)	1500 個	5 秒

表 4-10 高斯混合模型的資料分配

首先將所有的測試語音作特徵選取，經過竭盡式搜尋以及使用 KNNR 分類後，得到 15 種特徵參數組合的 KNNR 模型，KNNR 模型的辨識率如圖 4-12 所示。實驗結果顯示，同時選取 4 個特徵參數時，KNNR 模型的辨識率(98.4%)最高，因此在本實驗中選取 4 個特徵參數作為最佳的特徵參數組合。



1:平均音量, 2: 平均清晰度, 3: 平均音高, 4: 音框數

圖 4-12 使用 KNNR 作特徵選取辨識率圖

經過特徵選取過後，同時使用 4 個特徵參數訓練高斯混合模型，並且調整高斯混合數，得到一個辨識率較高的高斯混合模型，依此作為判斷測試語音好壞的機制。高斯混合模型的訓練資料和測試資料中包含兩個類別，分別為辨識正確和錯誤兩種，而訓練資料的類別比為 1483:17，測試資料的類別比為 1482:18，由於高斯混合數必須小於類別資料數，因此取 16 作為高斯混合數的上限。設定高斯混合數的範圍從 1 至 16，在不同的高斯混合數下，得到高斯混合模型的內部測試和外部測試辨識率如圖 4-13。由圖 4-13 得知，當高斯混合數為 16 時，高斯混合模型的內部測試辨識率(99.53%)和外部測試辨識率(98.87%)最高，因此使用高斯混合數為 16 的高斯混合模型，作為設定拒絕門檻值的依據。

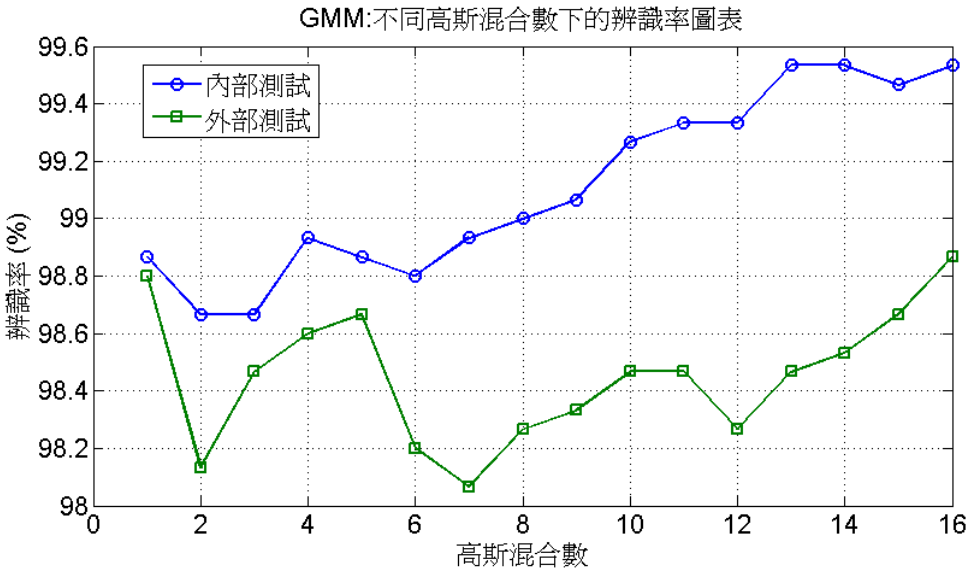


圖 4-13 高斯混合模型: 不同高斯混合數下的辨識率圖表

為了與各別特徵參數設定門檻值的方法作比較，因此選取相同的拒絕比例，將所有測始語音與高斯混合模型的對數似然比例由小到大排列，根據相同拒絕比例，設定對數似然比例的門檻值，拒絕對數似然比例較低的測試語音。以理想人工標端點的測試語音而言，拒絕率取 4.53%，所對應的門檻值為-7.6984，語者識別的實驗結果見表 4-11；而經過改善端點偵測錯誤接受的測試語音而言，拒絕率取 5.77%，所對應的門檻值為-7.3932，語者識別的實驗結果見表 4-12。實驗結果顯示，不論是理想人工標端點後，或是改善端點偵測錯誤後的測試語音，經過高斯混合模型設定拒絕門檻，語者識別的辨識率皆有提升，但是和表 4-6、4-7 的實驗結果相比，使用兩道門檻值設定的方法錯誤降低率較高。

	初始結果	**(1)理想人工標端點**	**(1)+(2)GMM 設定門檻值**	**拒絕比例**	**錯誤降低率**
整數型態	93.10%	95.33%	**95.71%**	4.53%	37.83%
浮點數型態	96.53%	98.83%	**99.02%**	4.53%	71.76%

表 4-11 理想人工標端點後且 GMM 門檻值設定下的實驗結果

	初始結果	(1)改善端點偵測錯誤接受	(1)+(2)GMM設定門檻值	拒絕比例	錯誤降低率
整數型態	93.10%	94.17%	**94.84%**	5.77%	25.22%
浮點數型態	96.53%	98.17%	**98.80%**	5.77%	65.42%

表 4-12 改善端點偵測錯誤接受後且 GMM 門檻值設定下的實驗結果

4.4. 拒絕不完整測試語音的實驗結果

將所有測試語音 3000 個音檔，判斷第一個和最後一個音框是否有音量，當作濾除不完整的測試語音機制，其中有 19 個音檔第一個或是最後一個音框有音量，拒絕比例佔所有音檔的 0.63%。而在拒絕的句子當中，包含真正錄音不完整的測試語音有 6 個音檔，而測試語音前後有雜訊的有 10 個音檔，錄音內容不正確的有 2 個音檔，和錄音內容完整的有 1 個音檔。

在理想人工標端點後、改善端點偵測錯誤接受後兩種情況下，且設定拒絕門檻分別為各別特徵參數設定門檻值、高斯混合模型設定門檻值後的實驗結果，加上拒絕不完整測試語音作語者識別的實驗，總共有四種實驗結果(表 4-13、表 4-14、表 4-15、表 4-16)。

	初始結果	(1)理想人工標端點	(1)+(2)兩道門檻值設定	(1)+(2)+(3)拒絕不完整的測試語音	拒絕比例	錯誤降低率
整數型態	93.10%	95.33%	97.31%	**97.44%**	4.93%	62.90%
浮點數型態	96.53%	98.83%	99.51%	**99.58%**	4.93%	87.90%

表 4-13 拒絕不完整測試語音的實驗結果一

	初始結果	(1)改善端點偵測錯誤接受	(1)+(2)兩道門檻值設定	(1)+(2)+(3)拒絕不完整的測試語音	拒絕比例	錯誤降低率
整數型態	93.10%	94.17%	96.64%	**96.94%**	6.20%	55.65%
浮點數型態	96.53%	98.17%	99.19%	**99.40%**	6.20%	82.71%

表 4-14 拒絕不完整測試語音的實驗結果二

	初始結果	(1)理想人工標端點	(1)+(2)GMM 設定門檻值	(1)+(2)+(3)拒絕不完整的測試語音	拒絕比例	錯誤降低率
整數型態	93.10%	95.33%	95.71%	**95.79%**	5.03%	38.99%
浮點數型態	96.53%	98.83%	99.02%	**99.09%**	5.03%	73.78%

表 4-15 拒絕不完整測試語音的實驗結果三

	初始結果	(1)改善端點偵測錯誤接受	(1)+(2)GMM 設定門檻值	(1)+(2)+(3)拒絕不完整測試語音	拒絕比例	錯誤降低率
整數型態	93.10%	94.17%	94.84%	**95.06%**	6.30%	28.41%
浮點數型態	96.53%	98.17%	98.80%	**98.97%**	6.30%	70.32%

表 4-16 拒絕不完整測試語音的實驗結果四

四種實驗結果得知，使用各別特徵參數設定拒絕門檻的方法，比使用高斯混合模型設定拒絕門檻的方法，加上拒絕不完整測試語音後錯誤降低率較高，且經過拒絕不完整測試語音後，浮點數型態的辨識率相當接近 100%。

4.5. 錯誤分析

以理想人工標端點而言，取浮點數型態的特徵參數，且經過各別特徵參數設定拒絕門檻值，加上拒絕不完整的測試語音後，仍辨識錯誤的音檔總共有 12 個。由於設定平均音量和音框數的拒絕比例，無法將所有音量小或是音框數少而辨識錯誤的情形濾除，所以依然有音量太小、音框數太少造成辨識錯誤的情形。

若以改善端點偵測錯誤後，取浮點數型態的特徵參數，且經過各別特徵參數設定拒絕門檻值，加上拒絕不完整的測試語音後，仍辨識錯誤的音檔總共有 17 個。其中包含語音片段前後緊鄰雜訊(如圖 4-1、4-2)，以及雜訊片段比語音片段長，經過改善端點偵測錯誤接受後，反而保留雜訊片段切除語音片段(如圖 4-3)，由於測試語音含雜訊片段使得辨識錯誤。另外，由於改善端點偵測錯誤拒絕後，錯誤降低率反而下降，所以只改善端點偵測錯誤接受，語音片段過小被切除的音檔仍會辨識錯誤。除此之外，和理想人工標端點的情形相同，平均音量、音框數設定的門檻值無法排除所有辨識錯誤的音檔。

另外，在相同拒絕比例下，高斯混合模型設定門檻值的錯誤降低率比各別特徵參數設定拒絕門檻值的錯誤降低率低。從圖 4-6 得知，這四種特徵參數中，平均音量對於語者識別的辨識結果影響最大，因此針對平均音量作拒絕門檻值，即可將錯誤降低率大幅提升。但高斯混合模型使用四種特徵參數訓練模型，並且設定拒絕門檻值，相對地平均音量的拒絕比例較小，因此對於降低錯誤率的效果也有限。

第5章　結論與展望

在本實驗中，對於文本相關的語者識別，利用改善端點偵測錯誤以及濾除不佳測試語音的方式，確實能夠提升語者識別的辨識率和錯誤降低率。各個階段的實驗結果，整理如下表 5-1，在改善端點偵測錯誤拒絕的部分，由於錯誤降低率比改善端點偵測錯誤接受低，所以不採用改善端點偵測錯誤拒絕的方法，接續以下各個階段的實驗。而在相同拒絕比例下，各別參數設定門檻值與高斯混合模型設定門檻值的方式相比較，各別參數設定拒絕門檻的錯誤降低率較高，因此選用各別特徵參數設定拒絕門檻值，加上拒絕不完整測試語音為最佳的實驗結果，且不論整數型態或浮點數型態的特徵參數，其拒絕比例佔全部測試語音的 6.20%，以測試語音 3000 個音檔而言，總共拒絕 186 個音檔，錯誤降低率分別可達 55.65% 和 82.71%。

	初始結果	(1)改善端點偵測錯誤接受	(2)改善端點偵測錯誤拒絕	(1)+(3)各別特徵參數設定門檻值	(1)+(4) GMM 設定門檻值	(1)+(3)+(5)拒絕不完整的測試語音	拒絕比例	錯誤降低率
整數型態	93.10%	94.17%	91.97%	96.64%	94.84%	**96.94%**	6.20%	55.65%
浮點數型態	96.53%	98.17%	95.80%	99.19%	98.80%	**99.40%**	6.20%	82.71%

表 5-1 各階段的實驗結果

將理想人工標端點和改善端點偵測錯誤接受後，經過各別特徵參數設定門檻值，加上拒絕不完整測試語音的方法，兩種辨識結果相比較(見表 5-2)，由實驗結果得知，改善端點偵測錯誤接受下的最終結果，辨識率逼近理想人工標端點下

的最終結果。由於理想人工標示的端點準確度較高，而且拒絕門檻值的設定以理想人工標端點的辨識結果為標準，所以理想人工標端點的辨識率較高，而拒絕比例也較低，錯誤降低率也較高。

	初始結果	理想人工標端點下的最終結果	拒絕比例	錯誤降低率	改善端點偵測錯誤下的最終結果	拒絕比例	錯誤降低率
整數型態	93.10%	**97.44%**	4.93%	62.90%	**96.94%**	6.20%	55.65%
浮點數型態	96.53%	**99.58%**	4.93%	87.90%	**99.40%**	6.20%	82.71%

表 5-2 理想人工標端點和改善端點偵測錯誤下的最終結果比較表

實驗結果顯示，浮點數型態的特徵參數作語者識別，辨識率最高可達 99.40%，相當逼近 100%。但整數型態的特徵參數由於精確度比較低，.辨識率最高可達 96.94%，若實際應用在門禁系統上，辨識率需要再提升，使得錯誤率愈低愈好。

改善整數型態的特徵參數辨識率，未來可朝下列方法研究改進，其一改變語者識別的方法，可嘗試使用從任意處比對的動態時間扭曲，即可改善測試語音不完整而辨識錯誤的情形。其二改變梅爾倒頻譜參數的維度，在此論文中使用 13 維的梅爾倒頻譜參數，未來可嘗試使用 26 維、39 維的梅爾倒頻譜參數，但相對地維度較高的梅爾倒頻譜參數在計算時間上較耗時。

參考文獻

【1】 Joseph P. Campbell, Jr., *Speaker recognition: A tutorial*, Proceedings of the IEEE Volume 85, Issue 9, Sep 1997.

【2】 Wutiwiwatchai, C. Achariyakulporn, V. Tanprasert, C. , *Text-dependent speaker identification using LPC and DTW for Thailanguage*, TENCON 99. Proceedings of the IEEE Region 10 Conference.

【3】陳江村, 張智星, 李俊毅, 吳銘鈞, "結合 HMM 和 DTW 的兩階段式門禁系統", Proceedings of the Seventh Conference on Artificial Intelligence and Applications (第七屆人工智慧與應用研討會), Tai-Chung, Taiwan, Nov 2002.

【4】 吳銘鈞, "以音節為基礎之語者識別", 清華大學碩士論文, 民國 92 年

【5】 林青慧, "強韌式語者辨識系統：從麥克風、市話到手機", 清華大學碩士論文, 民國 93 年

【6】 趙怡翔, "鑑別式訓練法於語者驗證之研究", 交通大學博士論文, 民國 98 年

【7】 許世俊, "用於高斯混合模型語者辨認之區別式訓練方法", 清華大學碩士論文, 民國 85 年

【8】 吳金池, "語者辨識系統之研究", 中央大學碩士論文, 民國 90 年

【9】 陳俊傑, "類神經模糊與軟式計算在語者辨識上的應用", 清華大學碩士論文, 民國 86 年

【10】Philip McLeod, Geoff Wyvill, *A smarter way to find pitch*, University of Otago. Department of Computer Science, 2005.

【11】Jyh-Shing Roger Jang, 線上中文教材: 音訊處理與辨識 http://neural.cs.nthu.edu.tw/jang/books/audioSignalProcessing/

【12】 張文杰，陳鼎允，陳子和，曾志仁，廖元甫，莊堯棠，“結合韻律與聲學訊息之強健性漢語語者驗證系統”，Chinese computational linguistics (ROCLING), 2006.

【13】 李俊毅，“語音評分”，清華大學碩士論文，民國91年

【14】 楊壁如，“語者與歌者識別”，清華大學碩士論文，民國89年