國 立 清 華 大 學

碩 士 論 文

題目:口說英語重音辨識之初步研究
An Initial Study on Stress Detection for Spoken English

系別 資訊工程學系 組別

學號姓名 9662539 曾璟鈺 (Ching-Yu Tseng)

指導教授 張智星博士(Jyh-Shing Roger Jang)

中華民國九十七年六月

摘要

本篇論文的研究主旨是對於多音節之英語詞彙進行重音辨識。主要使用兩種辨識方法，「一維特徵參數的辨識方法」以及「兩步驟分類的辨識方法」。兩個方法所需要的輸入資料為單一個多音節詞彙的語音資料，經由強制對位產生子音與母音音素，再對各因素取得音高、音量向量以及持續時間。其中母音音素的資訊可以代表一個音節。最終需要得到的結果即一個詞彙只能有一個母音被標記為重音。

第一種方法是對一個詞彙的母音音素(亦代表音節)取出音高向量(pitch vector)和音量向量(volume vector)，再使用不同的計算方法，分別為中位數、平均值、最大值、導函數取最大值(maximum of derivative)、導函數取中位數、第一四分位數和第三四分位數，將音高及音量向量轉換成數值；因此，每種計算方法均會產生一個特徵參數值。最後，直接使用單一特徵值（如：各母音的音高中位數）來辨識一個詞彙的重音。

第二種方法分為兩個步驟，第一步驟是使用高斯混合模型(Gaussian Mixture Model, GMM)對各個母音音素(亦代表音節)的特徵參數做分類，分出重音與非重音兩類。因為一個詞彙只能有一個主要的重音音節，因此第二步驟主要對 n 個音節詞彙進行 n 類的分類問題來決定重音音節位於第幾個音節，例如 2 個音節詞彙可分為第一個音節為重音以及第二個音節為重音 2 類。而第二步的特徵是使用第一步驟所產生的各詞彙重音與非重音之 log likelihood。

在本論文的實驗中，第一種辨識方法進行了 8 組實驗，而最佳辨識率為 82.58%，使用的單一特徵為對音高向量取中位數。第二種辨識方法進行了 11 組實驗，第二步驟對 2、3、4 個音節詞彙進行分類，最佳辨識率分別為 90.36%、86.85%、85.65%。比第一種方法提高約 3~7%。本實驗結果顯示，我們提出的方法，可以有效地使用音高、音量和持續時間，成功辨識出口說英語重音。

Abstract

We introduce an initial study focusing on stress detection for multi-syllable English word. Two kinds of methods for stress detection are proposed. One is "the recognition method using one feature"; the other is "the recognition method using two-stage classification." The input data is a multi-syllable word, and we generate consonant phone and vowel phone using forced alignment. The information of vowel phone can represent a syllable. The final result must be a word that only select one vowel for stressed syllable.

The first method involves calculating median, average, maximum, maximum of derivative, mean of derivative, Q_1 and Q_2 values respectively from pitch and volume vectors extracted from vowel phones of a word. Each calculation can generate a feature; therefore, eight different features will be produced. The stress of an English word is detected by using one of eight features, such as the median of each vowel's pitch.

There are two stages in the second method. At stage 1, a two-class classifier using Gaussian Mixture Model is used to determine if a given syllable is stressed or not, based on a "bag of syllables" of the training set. Since a word can only have a major stressed syllable, we cast the task another classification problem at stage 2. At stage 2, we need select the stressed syllable from a given utterance of a word. This task can be cast into an n-class classification problem when determining a stressed syllable from an n-syllable word. The features for stage-2 classification can be selected from quantities computed by stage-1 classification.

We performed eight groups of experiments for the first method. The feature of median of the pitch vectors achieves the best recognition rate, 82.58%. For the second method, we performed 11 groups of experiments. The best recognition rates

for 2-, 3-, 4-syllable words are 90.36%, 86.85%, and 85.65%, respectively. Compared with the first method, the second method improves recognition rates of 7.78%, 4.27%, and 3.07%. The results show that our proposed approaches detect stress in spoken English word by using pitch, volume and duration.

目錄

圖片目錄

表格目錄

第1章 諸論

1.1 研究主題

英語，目前在世界各地中，是非常重要的溝通工具，為許多先進國家的主要母語，更是亞洲國家的第二語言，因此學習英語是一門十分重要的基本工作。然而英語詞彙中，重音扮演了極為重要的角色。

重音(stress)是一種語音現象，指的是在多音節的字詞中將某個音節念得特別明顯。這個被強調的音節會有音調較高、聲音較響亮、以及時間較長的現象。

所謂的明顯是相對的，因此只有一個音節的詞(如 eye, sun 等)不會有重音。兩個以上的音節會有主重音，三個以上的音節可能會有次重音。

英語是有重音的語言。假若某個詞彙的重音念錯，可能會因此聽不懂該詞彙是甚麼意思或是讓他人會錯意。有些英語詞彙的重音會因為它的詞性而有所改變。例如：increase 當動詞時重音在第二音節、當名詞時重音在第一音節。因此一個詞彙當重音位於不同的音節時，往往代表不同的意義。

本論文的研究主題是「英語重音辨識之初步研究」，意即要索取多音節英語詞彙中的主重音，在語音辨識方面，若能明確的辨識出重音音節的位置，就能增進辨識率[6]。此研究可以用於英語之語音評分或英語學習中評量之依據。

1.2 相關研究

目前有一些研究注意到在話語中突顯詞彙的不同(word prominence)與重音(stress)有極密切的關係。在Wang等人[1]研究了 13 個特徵，而其發現音節的時間(duration)與頻譜強度(spectrum intensity)是最可信賴的特徵。

在Tamburini [2]提出兩種類型的特徵來索取word prominence，包含了聲學的(acoustic)與韻律學的(prosodic)參數。前者包含了時間(duration)、基本頻率

(fundamental frequency)、能量(energy)、和spectral emphasis，而後者則是包含加重重音(stress accent)、音高重音(pitch accent)、和prominence-liked等特徵。使用多變量常態分布分prominent和non-prominent的模組。其中prominence探測器正確分類 80.73%的音節作為prominent或non-prominent。Jenkin等人[3]提出了三個對重音音節分類的替代技術，包含神經網絡(neural networks)，一階和二階馬可夫鏈(first and second order Markov chains)，和以規則為基底的方法。其中以神經網絡的方法最佳，最佳效能為 81-84%。在[4]中則是運用支援向量機(support vector machine)分類母音中重音與非重音，其效能達到 84.72%。

然而以上這些方法都是針對一個音節去做分類，分出該音節是否為重音。再多音節詞彙中只能有一個音節是被標記為主重音，因此我們需要後處理來達到一個詞彙中只有一個重音音節。

Tepperman and Narayanan [5]採用高斯混合模型(Gaussian Mixture Model, GMM)來建構重音(stressed)與非重音(unstressed)的模型，著重於每一個詞彙只會有一個重音音節，其方法是若一個詞彙出現兩個以上的重音，則比較GMM模組下“stressed”的log likelihood，取出有最大log likelihood的音節做為主重音。相反的，若一個詞彙中沒有標示出任何重音音節，則比較GMM模組下“unstressed”的log likelihood，取出有最小log likelihood的音節做為主重音。假若只有一個音節被標記為重音，則不需做任何更動。

1.3 本論文方法簡介與主要成果

本論文使用強制對位(Forced Alignment)的方法對各詞彙進行切音，產生各音素(phone)，對各音素取出三種特徵參數：音高向量(pitch vector)、音量向量(volume vector)以及持續時間(duration)。辨識方法分成一維特徵參數及多維特徵參數，多維特徵參數的主要方法是分為兩步驟分類法，首先使用高斯混合模型(Gaussian

Mixture Model, GMM)對各母音音素的特徵參數分重音及非重音，再將 GMM 分類後模組的 log likelihood，依照詞彙的音節總數分別使用 GMM 及支援向量機 (Support Vector Machine, SVM) 再次進行第二步驟的分類，目前只對有 2、3、4 個音節的詞彙做分類與辨識，最佳的辨識率分別為 90.36%、86.85%、85.65%。而以一維特徵參數(pitch 及 volume)來辨識的最佳辨識率分別為 82.58%、60.94%。

明顯可知使用兩步驟分類的方法比直接用一維參數來做辨識的辨識效果高出許多。

1.4 章節概要

本論文第二章介紹英語詞彙之重音的重要性，並說明本篇論文所探討之主題，隨後介紹英語詞彙重音辨識之系統架構與流程，並詳細說明本論文所用之研究方法。本論文採用 GMM 以及 SVM 進行訓練及進行相關實驗。

第三章介紹實驗的語音資料庫，記錄各種實驗之數據，進行分析錯誤結果。

第四章則是本論文的結論與未來展望，歸納並討論本論文的實驗結果及未來可進一步研究的方向。

第2章 英語詞彙重音辨識之研究

2.1 問題定義

在英語詞彙中，有兩個以上音節的詞彙稱為多音節詞彙。而一個詞彙中，某個音節念得特別明顯則稱為重音(stress)。這個被強調的音節會有音調較高、聲音較響亮、以及時間較長的現象。

音調較高即表示有較高的音高(pitch)；聲音較響亮則表示有較大的音量(volume)；而時間較長的現象表示重音音節的持續時間(duration)較長。因此本論文主要的特徵參數就以 pitch、volume 以及 duration 為主。

一個音節包含了一個或多個子音與一個母音，而母音代表了該音節的主要資訊。本論文主要探討的就是一個詞彙中各母音下的資訊，經由正規化以及不同的處理，將各個向量資料轉換成各個數值，產生各種不同的特徵參數。使用不同的特徵組合與比較利用分類器尋找出最佳的分類結果。而所謂不同的處理，其定義是由於一個音素下的音高及音量資訊是一串向量(vector)，因此可以使用 mean、median、max…等方法將一串向量的資訊轉換為一個數值。

2.2 系統架構及流程

本系統主要分為兩種方法，首先均使用 forced alignment 對每一個詞彙切音，產生 phone level。第一種方法是分別對同一個詞彙的各母音音素下的 pitch vector 與 volume vector 獨立處理，處理的方法有以下種：max、mean、median、Q_1(第一四分位數)、Q_3(第三四分位數)、max of derivation、mean of derivation，使得母音音素的 pitch vector 被計算成七種數值，volume vector 亦被計算成七種數值。再比較同一個詞彙中，各母音音素使用同一種方法，同一種特徵參數，取其有最大值的母音音素為重音音節。流程圖如<圖 1>，為一個詞彙下的辨識流程。

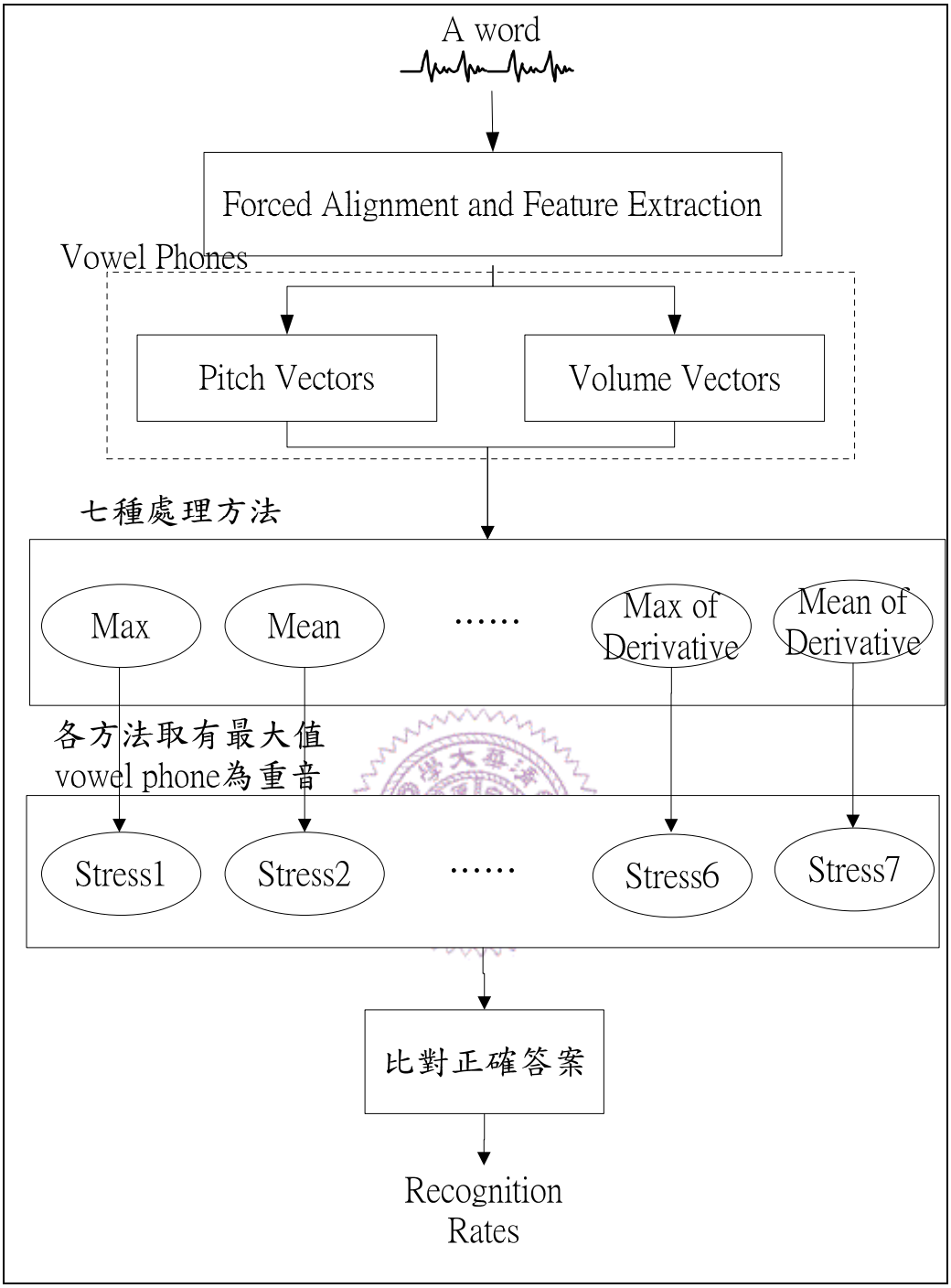

圖 1：一個詞彙下用 pitch 或 volume 單一特徵參數的辨識流程

<圖 2>為第二種方法的流程圖，主要是多維特徵參數的辨識方法，採用兩個步驟來索取重音音節。第一個步驟是對所有詞彙的各個母音音素做訓練，採用 GMM 的模組訓練重音(stressed)與非重音(unstressed)分類器。第二步驟則是以 n 個音節詞彙(n-syllable word)為單位，利用第一步驟的模組所產生的 log likelihood，即一個音節包含了重音(stressed)與非重音(unstressed)的 log likelihood，因此 n 個音節可以產生 2n 個特徵參數。利用 2n 個特徵參數以及已知該詞彙重音音節位置，再次使用 GMM 或 SVM 分類器來進行分類，得到最後的辨識結果。

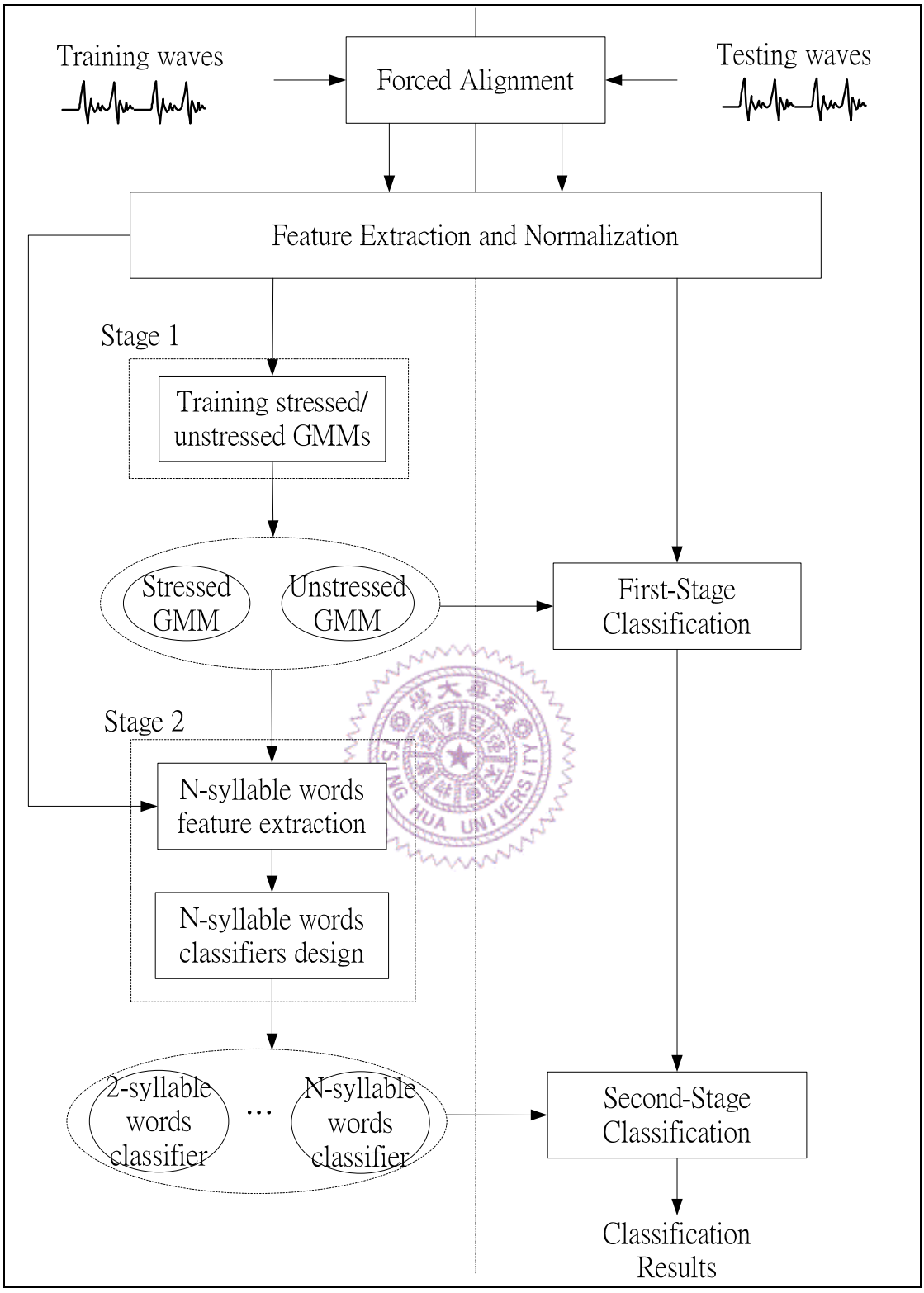

圖 2：多維特徵參數使用的 2-Stage 分類器流程圖

2.3 研究方法

2.3.1 各詞彙正確重音音節之前置處理

在進行各項實驗之前，首先要先取得語音資料庫中所有詞彙的正確重音位置，而取得的方式是根據 yahoo 網路字典，找出第一筆 KK 音標的重音符號(ˋ)位於第幾個母音，即代表該詞彙重音音節，此部分為初步正確答案。接著再人工檢查音檔，避免有錄音者念錯或該詞彙有多種念法，使得錄音者所錄製的重音音節與初步正確答案不同。最後再依據錄音者所錄製的重音位置來更改初步正確答案，此步驟後的正確答案將是多音節詞彙語音資料庫的正確重音音節。

2.3.2 特徵擷取與正規化

首先使用強制對位(Forced Alignment)的方法對各詞彙進行切音，產生各音素(phone)，對各音素取出三種特徵參數：音高向量(pitch vector)、音量向量(volume vector)以及持續時間(duration)。音高的擷取是採用 Unbroken Pitch Determination Using Dynamic Programming (UPDUDP)，而音量的擷取則是使用每個音框的絕對值總和。

對於音高向量與音量向量的正規化是採用 zero mean 和 unity variance 的方法，其中均以一個詞彙為單位來進行正規化。正規化後，母音音素的音高向量以及音量向量再各自取最大值(max)、平均值(mean)、中位數(median)、變異數(variance)，因此這部分就產生了 8 種特徵參數。

因為每個人說話的速度不同，所以對於持續時間的處理需先對說話速度(speech rate)做處理[4]。處理的方式如下：

(1) 計算每一種音素的平均時間，作為各個音素的預估時間

(2) 以一個詞彙為單位，各個音素的實際時間總和除以使用(1)所計算出的預估時間總和稱為 Ratio

(3) 各音素的實際時間乘以(2)計算出來的 Ratio 即為各音素的持續時間

(4) 使用 zero mean 和 unity variance 的方法正規化該詞彙下的所有音素的持續時間

(5) 取出母音音素部分的持續時間

(6) 重複(2)~(5)步驟取得所有詞彙下每個母音音素的持續時間

2.3.3分類方法

本論文主要使用了高斯混合模型(Gaussian Mixture Model, GMM)以及支援向量機(Support Vector Machine, SVM)兩種分類器，以下將依序介紹 GMM 與 SVM 的分類法則。

GMM

GMM 是單一高斯機率密度函數的延伸，由於 GMM 能夠平滑地近似任意形狀的密度分布，因此近年來常被用在語音與語者辨識，得到不錯的效果。

假設我們有一組在高維空間(維度為 d)的點$x_i, i = 1, \ldots, n$，若這些點的分佈近似橢球狀，則我們可用高斯密度函數$g(x, \mu, \Sigma)$ 來描述產生這些點的機率密度函：

$$g(x, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} exp\left[-\frac{1}{2}(x-\mu)^T \sum^{-1} (x-\mu)\right]$$

其中μ代表此密度函數的中心點，Σ則代表此密度函數的共變異矩陣(covariance matrix)，這些參數決定了此密度函數的特性，如函數形狀的中心點、寬窄及走向等。

如果我們的資料$X = \{x_1, \ldots, x_n\}$在 d 維空間中的分佈不是橢球狀，那麼就不適合以一個單一的高斯密度函數來描述這些資料點的機率密度函數。此時的變通方案，就是採用數個高斯函數的加權平均(Weighted average)來表示。若以 n 個高斯函數來

表示，則可表示成：

$$p(x) = \sum_{i=1}^{n} \alpha_i g(x, \mu_i, \Sigma_i)$$

而各個權重的值總和必須為 1，也就是

$$\sum \alpha_i = 1$$

以此種方式表示的機率密度函數，稱為「高斯混合密度函數」或是「高斯混合模型」(Gaussian Mixture Model)，簡稱 GMM。若要詳細的數學公式推導，可參考[9]。

在重音與非重音的訓練上，本研究中 GMM 採用了 2 維~9 維的特徵參數，此部分將會在 3.3 實驗二說明。

SVM

簡單來說，SVM 想要解決以下的問題：找出一個超平面(hyperplane)，也稱為決策函數(decision function)，使之將兩個不同的集合分開。為什麼使用超平面這個名詞，因為實際資料可能是屬於高維度的資料，而超平面意指在高維中的平面。

以二維空間上的分類做為範例，如<圖 3>

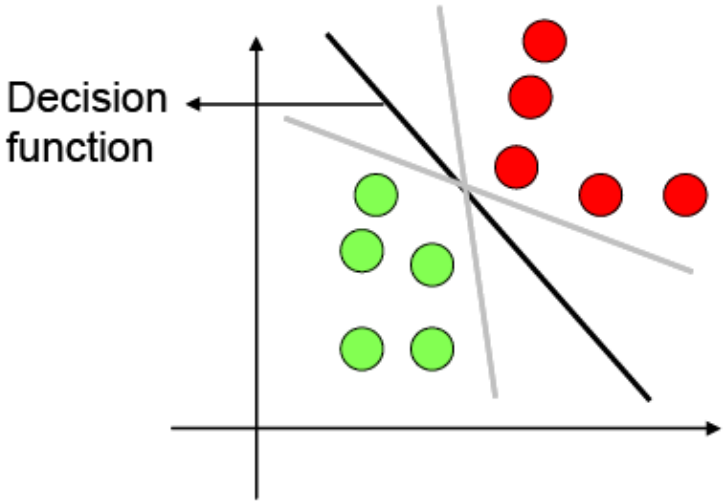

圖 3：SVM 分類示意圖

由觀察可得知，可以決定分類的超平面並非唯一，因此 SVM 的目標就是從中找出一最佳超平面。根據最佳超平面的定義是將 margin(每個類別資料點與其他類別資料點的最小距離)最大。將超平面定義如下：

$$M(w,b) = \frac{2}{|w|}$$

M 代表所要產生的 margin，b 代表 bias，依據圖 3 來看就是決策函數的垂直位移量。數學推導請參考[11]。經由數學的推導，最後的目標函數如下所示，其中ξ代表誤差項，即資料在邊界的地方重疊，因此需要誤差項來處理資料重疊於邊界的問題。然而誤差項希望越小越好，因此要給誤差項一個權重參數 C，給誤差的資料有一些懲罰，讓誤差資料多一些成本(cost)。

$$\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$\text{subject to } y_i(w^T x_i - b) \geq 1 - \xi_i \,, \xi_i \geq 0$$

以上討論的為線性可分割問題。倘若訓練資料為線性不可分割問題，如<圖 4>所示，變無法使用一般的線性分割方式。

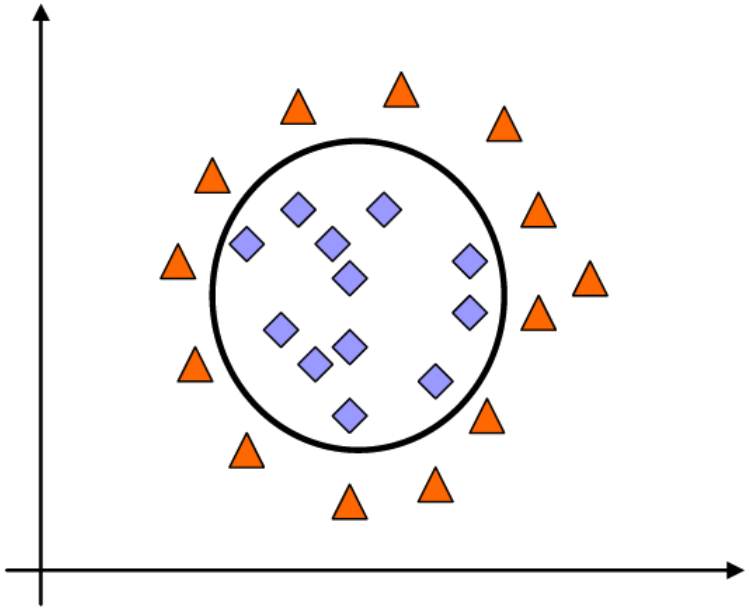

圖 4：SVM 線性不可分割示意圖

SVM 嘗試把線性不可分割的問題，映射至更高維度的特徵空間，使其成為

可線性分割的問題，並找出最佳超平面。以<圖 4>來說，使用多項式核心函數(polynomial kernel function)，將原來的二維空間映射至三維空間上，經由內積(inner product)運算其次方為 2，假設原來的座標為(x, y)，新的座標為(x', y', z')，那麼轉換的定義如下：

$$(x', y', z') = (x, \sqrt{2}xy, y)$$

在本論文中使用 LIBSVM [10]進行實驗，核心函數採用放射型核心函數(radial basis function, RBF)，函數如下：

$$K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2), \gamma > 0$$

因 RBF 處理速度上比其他核心函數來的快，且只需要調整參數為γ，故採用此方法。

2.3.4 辨識方法

主要的辨識方法分成一維特徵參數辨識方法以及多維特徵參數的辨識方法，多維的辨識方法主要是採用 GMM 以及 SVM 來分類取得辨識結果，系統架構及流程可參考 2.2 章節。

一維特徵值的辨識方法

對同一個詞彙下各個母音音素的音高向量以及音量向量分別取最大值(max)、平均值(mean)、中位數(median)、第一四分位數(Q_1)、第三四分位數(Q_3)、對向量取導函數後再取最大值(max of derivation)以及取導函數後再取平均值(mean of derivation)，使得母音音素的音高向量變成五種數值，音量向量亦變成七種數值。因此每一個母音音素都會產生以下表格的資訊：

Vowel phone	**max**	**mean**	**median**	**Q_1**	**Q_3**	**max of derivation**	**mean of derivation**
Pitch	P1	P2	P3	P4	P5	P6	P7
Volume	V1	V2	V3	V4	V5	V6	V7

表 1：各母音音素 1 維特徵參數處理後產生之內容架構

其中四分位數是將資料由小而大排序，再將之分為四等分，則會產生三個分割點，從小的分割點算起，第一個分割點稱為第一四分位數，以 Q_1 表示；第二個分割點稱為第二四分位數即中位數；第三個分割點稱為第三四分位數，以 Q_3 表示。

同一個詞彙的每一個母音音素均使用以上的方法產生十種數值後，比較同一個詞彙中各母音使用同一種方法以及同一個特徵參數所產生的數值，取得有最大值的母音音素作為重音音節。如<表 2>所示，表示同一個詞彙中有 n 個母音音素，而 X_i 代表使用音高向量取平均值的方法，Y_i 代表使用音高向量取中位數的方法，由此範例可得知 max{ X_1,…, X_n}為音高使用平均值取得的重音(Stress using mean of pitch)，max{ Y_1,…, Y_n}為音高使用中位數取得的重音(Stress using median of pitch)。依此類推，一個詞彙可以得到十個辨識後的重音音節。而後再與正確答案的重音音節比較，即可取得辨識率。

word		
Vowel phone 1	….	Vowel phone N
X_1=mean(pitch)	….	X_n=mean(pitch)
Y_1=median(pitch)	….	Y_n=median(pitch)

表 2：同一詞彙中使用同一種方法、同一個特徵參數的示意圖

多維特徵值選取類型及兩步驟分類的辨識方法

多維的特徵參數，主要於 2.3.2 章介紹過，經過正規化處理後，各母音音素的音高向量(pitch vector)以及音量向量(volume vector)分別取最大值(max)、平均值(mean)、中位數(median)、變異數(variance)，產生 8 種特徵值，再加上經過處理後的持續時間(duration)，總共有 9 種特徵值，而本論文分別採用為 2 維特徵參數、3 維特徵參數、9 維特徵參數以及利用 sequential forward selection(SFS)選出有最佳辨識率之特徵參數組合等實驗，來取得最佳辨識結果，此部分將於 3.3 節實驗二中做說明。

至於辨識方法主要分為兩個步驟，第一個步驟是對所有詞彙的各個母音音素做訓練，使用高斯混合模型(Gaussian Mixture Model, GMM) 訓練重音(stressed)與非重音(unstressed)的模組。第二步驟要選出一個詞彙中重音的音節，以 n 個音節詞彙(n-syllable word)為單位，利用第一步驟的模組所產生的 log likelihood，即一個音節包含了重音(stressed)與非重音(unstressed)的 log likelihood，因此 n 個音節可以產生 2n 個特徵參數。使用 2n 個特徵參數以及已知該詞彙重音音節位置，再次使用 GMM 或 SVM 分類器來進行 n 類的分類，得到最後的辨識結果。

以<表 3>為例，2 個音節的詞彙經過第一步驟的 GMM 模組後，第一個母音音素會產生兩個 log likelihood (A1,A2)，分別代表重音與非重音的 log likelihood，第二個母音音素亦產生兩個 log likelihood (B1,B2)。因此在第二個步驟的特徵參數就是{A1,A2,B1,B2}4 個參數。

2-syllable word	**Vowel phone 1**	**Vowel phone2**
Stressed log likelihood	A1	B1
Unstressed log likelihood	A2	B2

表 3：2 個音節詞在第一步驟後的 log likelihood

第3章 實驗結果與討論分析

3.1 資料庫說明

本論文使用的語音資料庫名稱為「多音節英語詞彙 2008」，由 10 位男性、12 位女性共 22 人，且均為非以英文為母語(non-native speaker)所錄製，詞彙總數為 3387 個詞，取樣頻率為 16kHz，解析度 16bits，各詞彙錄音時間為 3 秒。每個人錄製的詞彙均不重覆。重音音素與非重音音素的比例為 3:7。

下列<表 4>代表本實驗之資料庫各個音節詞彙重音位置位於第幾個音節的總數，也顯示出各音節詞彙的總數，由該表格可以看出 2、3、4 個音節詞彙的數量較多且也較足夠做為分類。而本實驗中使用 GMM 的部分均有採用資料分布的情況來做調整。

Numbers of syl. / Stress Position s	n=2	n=3	n=4	n=5	n=6	n=7
s=1	429	774	158	17	8	0
s=2	110	639	485	70	9	1
s=3		66	295	168	6	0
s=4			4	94	35	0
s=5				0	11	5
s=6					0	0
s=7						0
Total	539	1479	942	349	69	6

表 4：各音節詞彙之重音位置分析

3.2 實驗一：使用單一特徵值辨識方法

使用一維特徵值的辨識結果如下<表 5>，音高辨識結果以中位數(median)處理音高向量的方法最佳，而音量辨識結果以平均值(mean)以及第一四分位數(Q_1)處理音量向量的方法最佳。以整體而言，音高及音量採用 mean 以及 median 的辨識率相差不多，Q_1 的音量辨識與 mean 的辨識相同但於 pitch 均比 mean/median 的方法差了 1.5%左右，Q_3 的音高級音量辨識均比前三個方法差，但是比使用最大值(max)的辨識結果來的好，而取導函數後的最大值(max of derivative)以及取導函數後的平均值(mean of derivative)的方法辨識效果遠不如其他三種方法。

方法	pitch 辨識率(%)	volume 辨識率(%)
Mean	83.1199	61.5064
Median	83.6261	60.1369
Q_1	81.8398	61.5064
Q_3	82.3162	58.4102
Max	77.3742	56.5644
Max of derivative	52.0095	42.9889
Mean of derivative	38.7019	26.7043

表 5：一維特徵值辨識結果

3.3 實驗二：使用多維特徵值各組合進行 2 步驟分類

本實驗依人數來分訓練與測試資料，訓練與測試的人數比例為 15:7。為避免資料可能分布不均等問題，分成三組訓練測試資料，而訓練與測試的人交錯組合，分別進行分類器的訓練與測試，而後在求其平均，才視為最後辨識結果。

而本實驗中，分別使用了 2 維、3 維、6 維以及 9 維的特徵參數來分類，因

語音資料庫在 2、3、4 個音節詞彙(2,3,4-syllable word)的資料量較足夠做分類(參考<表 4>)，故在此部分的第二步驟分類時只實作這三種音節詞彙。

由於執行一組特徵組合的時間需要花費約 5~6 小時，因此對於選取特徵的組合並非使用暴力法對於所有可能的組合作實驗；對於二、三維特徵組合的選取是以音高向量及音量向量只選取一種方法，因此音高或音量最多只會出現一次，不會有兩個以上音高或音量的處理方法同時存在，例如: mean(pitch)和 median(pitch)不會同時出現。

3.3.1 2 維特徵參數組合

2 維特徵參數主要有音量取平均值(簡稱 mean Volume)、音量取中位數(簡稱 median Volume)、音高取平均值(簡稱 mean Pitch)、音高取中位數(簡稱 median Pitch)、持續時間(稱 Duration)，選擇前四種特徵是參考 3.2 節中使用一維辨識方法後音高及音量最佳的兩種特徵參數，這些特徵參數均於 2.3.2 節說明如何取得及正規化的方法，而組合成以下 5 種做分類實驗：

1. mean Volume + Duration
2. median Pitch + median Volume
3. mean Pitch + mean Volume
4. median Pitch + mean Volume
5. median Pitch + Duration

(1) mean Volume + Duration

第一步驟：

使用 GMM 分類，其最佳辨識率發生在 Gaussian number 為 4，而辨識率有 76.58%，如<圖 5>所示。

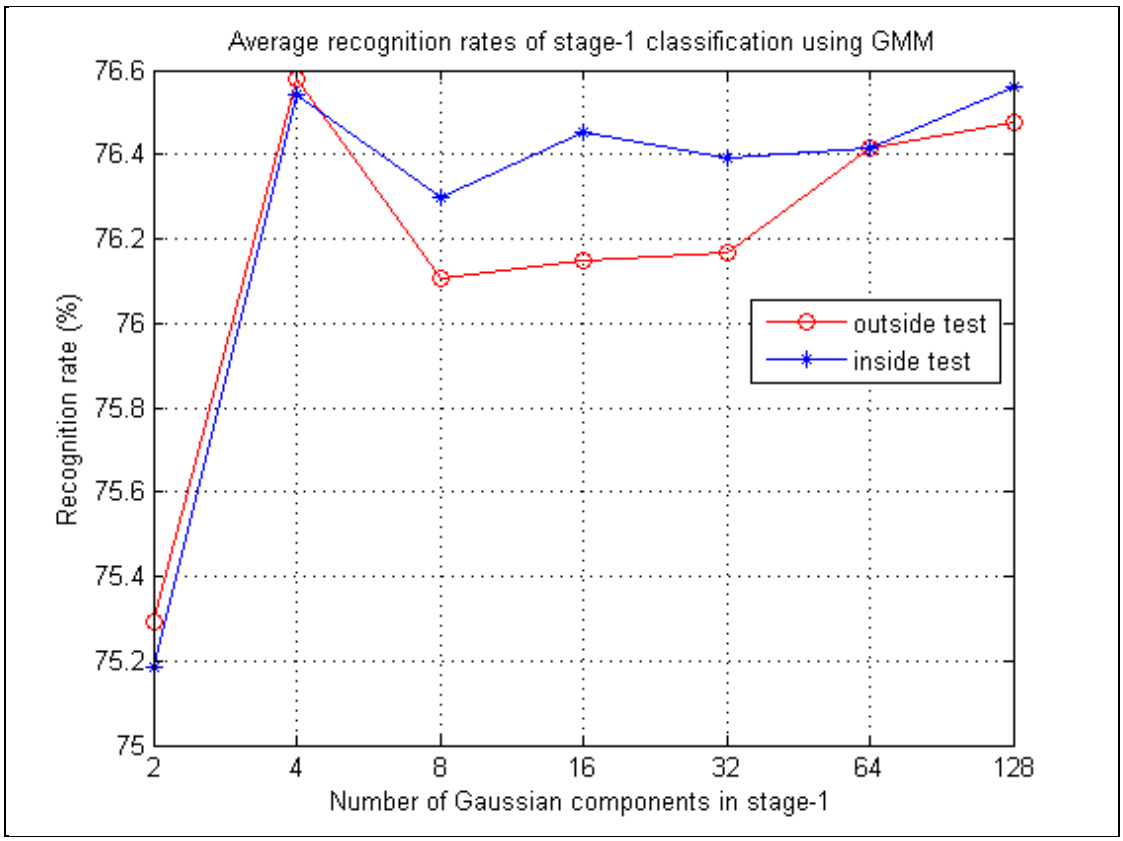

圖 5：第一步驟使用 GMM 分類器，3 組訓練/測試資料之平均辨識率

第二步驟：

分別使用 GMM 與 SVM 分類器的最佳辨識結果如下<表 6>。

Classifier / N-syllable word	GMM (%)	SVM (%)
2-syllable word	83.8753	85.2865
3-syllable word	74.1048	76.0136
4-syllable word	69.9087	70.5108

表 6：第二步驟使用 GMM 與 SVM 於 2,3,4 音節詞彙的最佳辨識率

在 GMM 的方法中，2、3、4 音節詞彙有最佳辨識率發生在 Gaussian Number 為 1 的情況，如<圖 6>所示。在 SVM 的方法中，2、3、4 音節詞彙各自使用的參數如<表 7>所示。

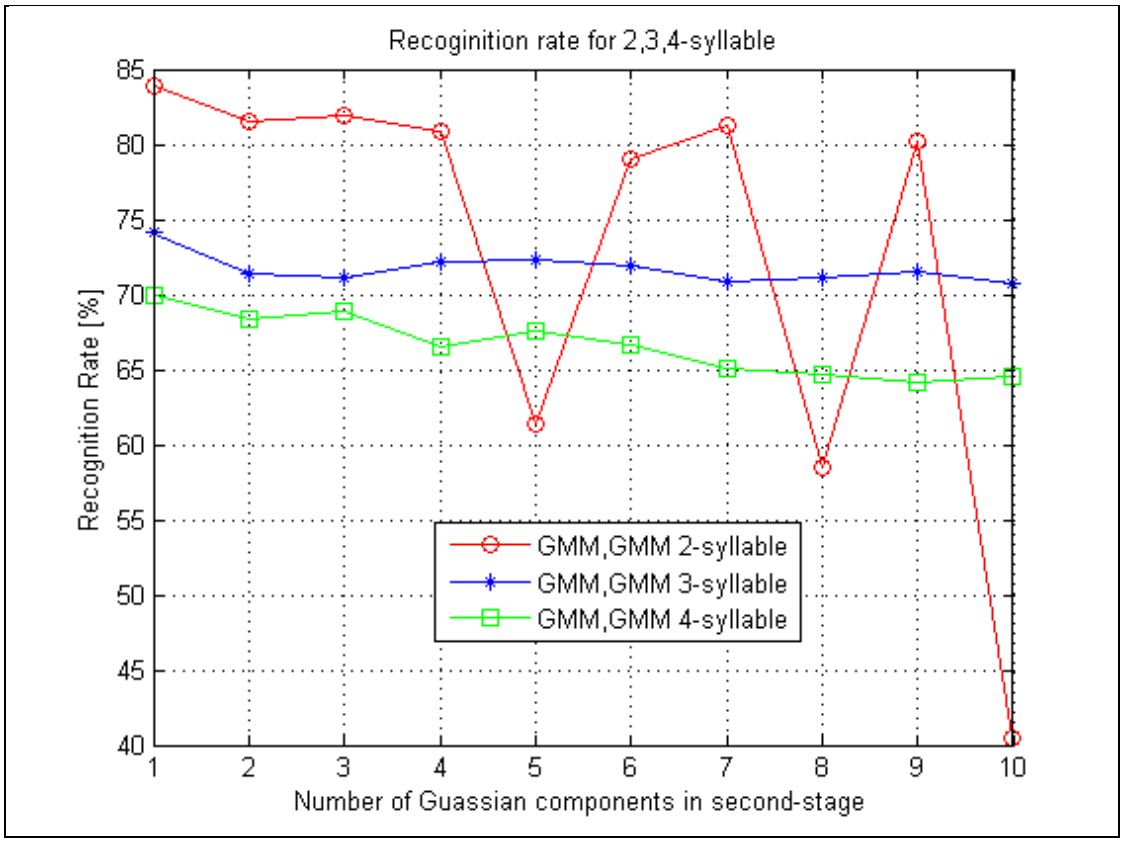

圖 6：2,3,4 音節詞彙使用 GMM 的各自辨識率

	Gamma (γ)	**Cost (C)**
2-syllable word	2^{-9}	2^{4}
3-syllable word	2^{-3}	2^{0}
4-syllable word	2^{-10}	2^{4}

表 7：2,3,4 音節詞彙於 SVM 中各自使用之參數

(2) median Pitch + median Volume

第一步驟：

使用 GMM 分類，其最佳辨識率發生在 Gaussian number 為 2，而辨識率有 83.39%，如<圖 7>所示。

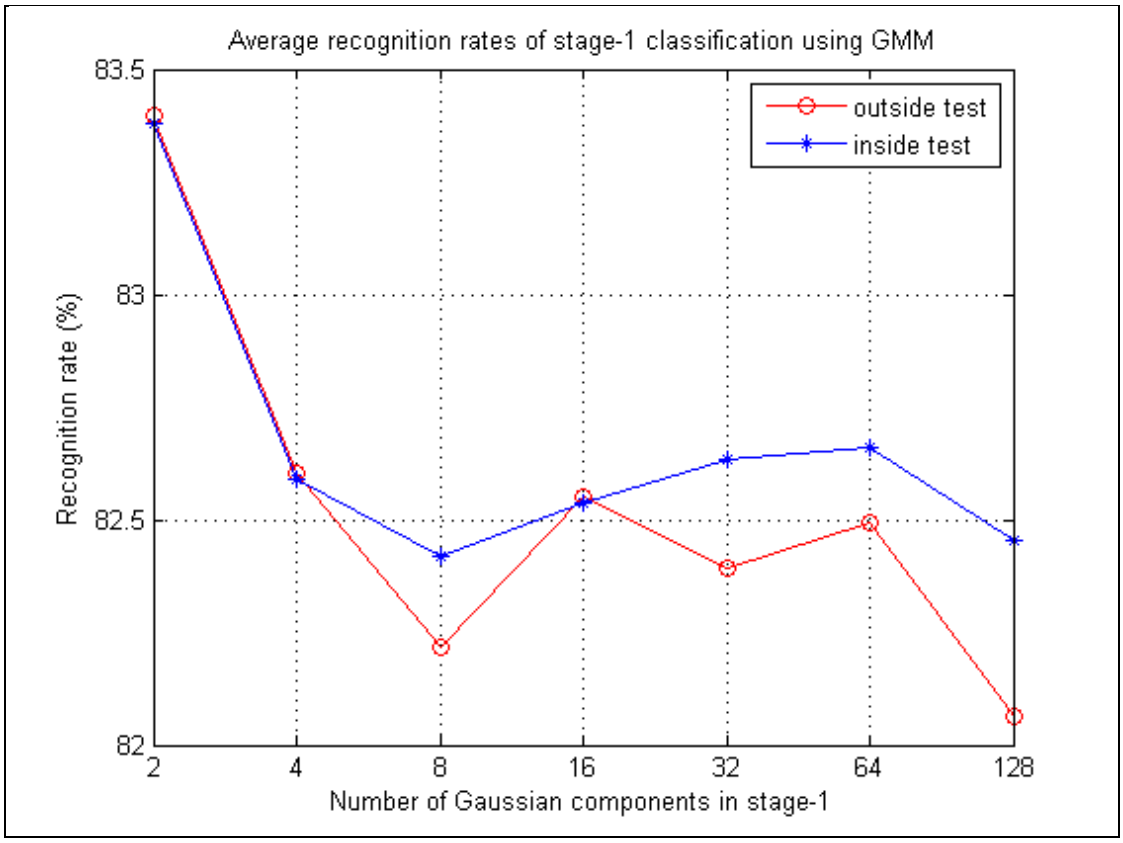

圖 7：第一步驟使用 GMM 分類器，3 組訓練/測試資料之平均辨識率

第二步驟：

分別使用 GMM 與 SVM 分類器的最佳辨識結果如下<表 8>。

Classifier / N-syllable word	GMM (%)	SVM (%)
2-syllable word	88.0871	90.1239
3-syllable word	81.7133	86.2399
4-syllable word	83.7974	82.2988

表 8：第二步驟使用 GMM 與 SVM 於 2,3,4 音節詞彙的最佳辨識率

在 GMM 的方法中，2、3、4 音節詞彙有最佳辨識率各自發生在 Gaussian Number 為 6、1、5 的情況，如<圖 8>所示。在 SVM 的方法中，2、3、4 音節詞彙各自使用的參數如<表 9>所示。

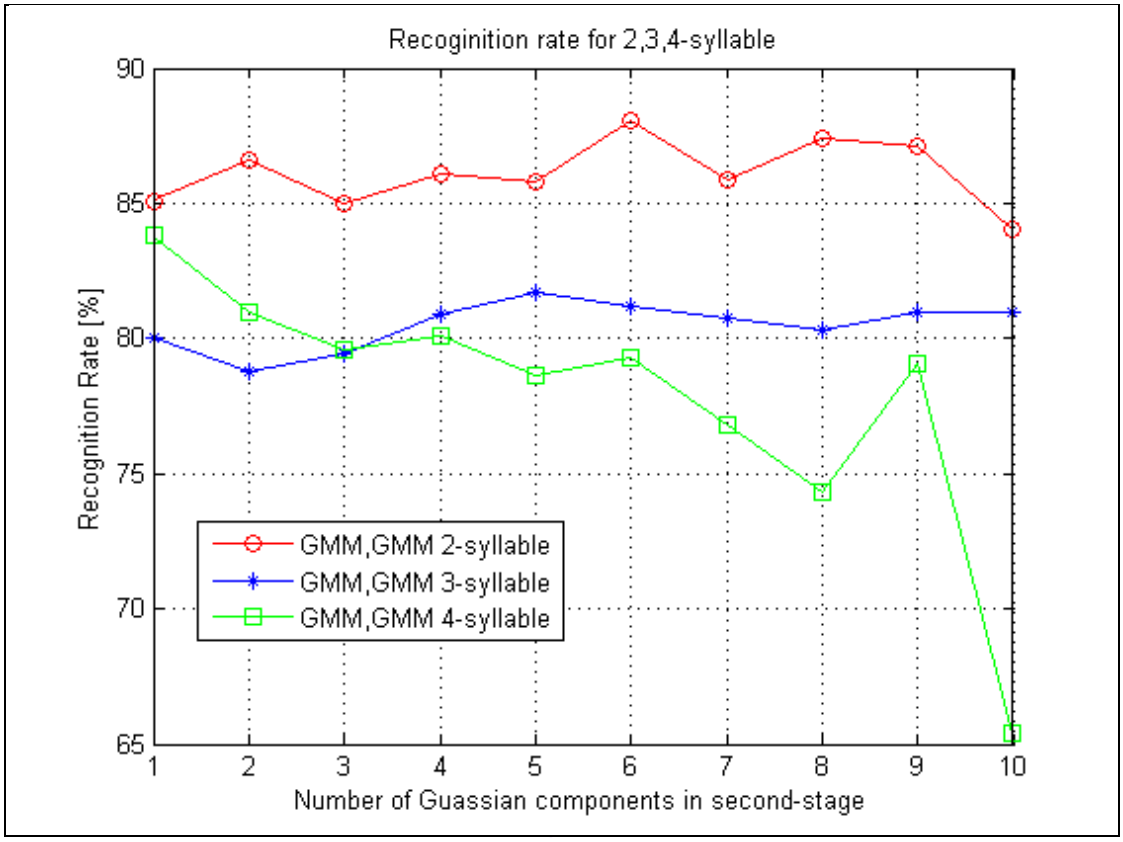

圖 8：2,3,4 音節詞彙使用 GMM 的各自辨識率

	Gamma (γ)	**Cost (C)**
2-syllable word	2^{-5}	2^{3}
3-syllable word	2^{-3}	2^{1}
4-syllable word	2^{-6}	2^{0}

表 9：2,3,4 音節詞彙於 SVM 中各自使用之參數

(3) mean Pitch + mean Volume

第一步驟：

使用 GMM 分類，其最佳辨識率發生在 Gaussian number 為 2，而辨識率有 83.29%，如<圖 9>所示。

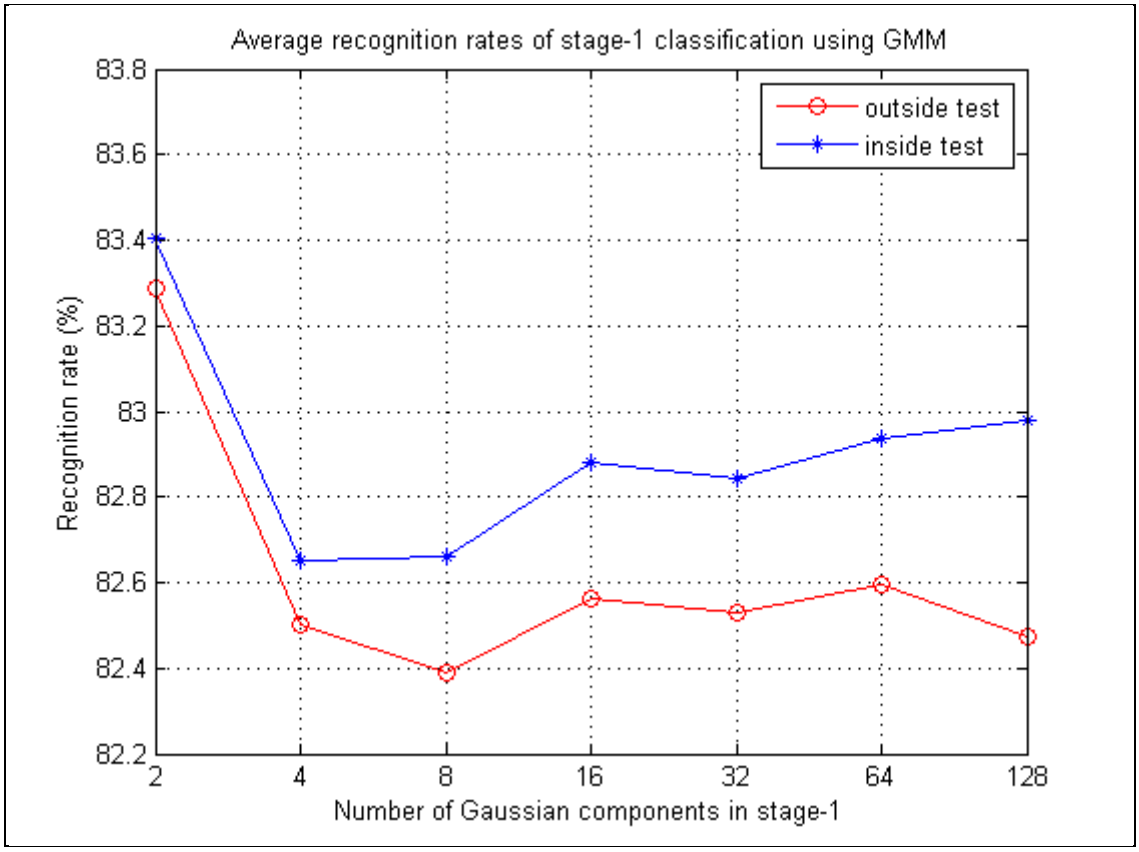

圖 9：第一步驟使用 GMM 分類器，3 組訓練/測試資料之平均辨識率

第二步驟：

分別使用 GMM 與 SVM 分類器的最佳辨識結果如下<表 10>。

Classifier / N-syllable word	GMM (%)	SVM (%)
2-syllable word	86.6295	88.9746
3-syllable word	81.6632	85.5325
4-syllable word	83.3157	82.2575

表 10：第二步驟使用 GMM 與 SVM 於 2,3,4 音節詞彙的最佳辨識率

在 GMM 的方法中，2、3、4 音節詞彙有最佳辨識率各自發生在 Gaussian Number 為 9、10、1 的情況，如<圖 10>所示。在 SVM 的方法中，2、3、4 音節詞彙各自使用的參數如<表 11>所示。

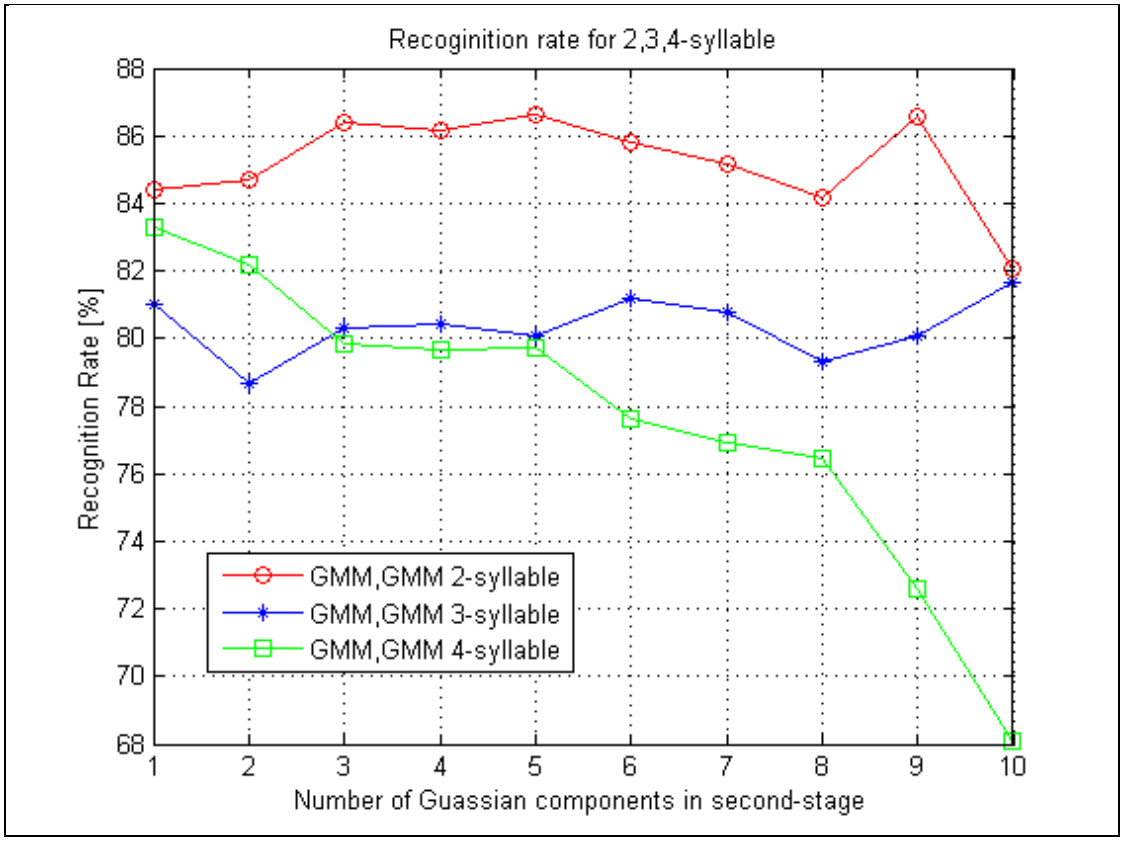

圖 10：2,3,4 音節詞彙使用 GMM 的各自辨識率

	Gamma (γ)	**Cost (C)**
2-syllable word	2^{-6}	2^{3}
3-syllable word	2^{-4}	2^{2}
4-syllable word	2^{-10}	2^{5}

表 11：2,3,4 音節詞彙於 SVM 中各自使用之參數

(4) median Pitch + mean Volume

第一步驟：

使用 GMM 分類，其最佳辨識率發生在 Gaussian number 為 2，而辨識率有 83.05%，如<圖 11>所示。

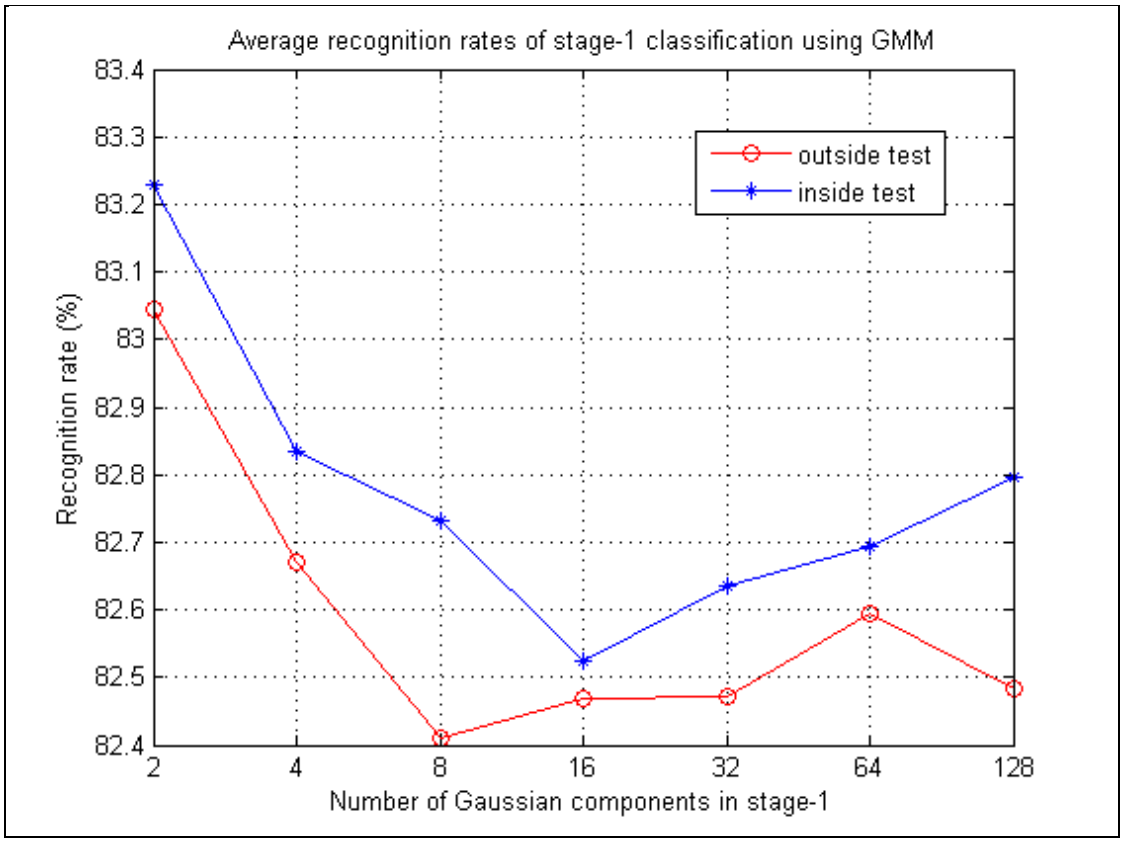

圖 11：第一步驟使用 GMM 分類器，3 組訓練/測試資料之平均辨識率

第二步驟：

分別使用 GMM 與 SVM 分類器的最佳辨識結果如下<表 12>。

Classifier / N-syllable word	GMM (%)	SVM (%)
2-syllable word	86.0032	90.1444
3-syllable word	80.8901	85.5612
4-syllable word	83.4411	81.8871

表 12：第二步驟使用 GMM 與 SVM 於 2,3,4 音節詞彙的最佳辨識率

在 GMM 的方法中，2、3、4 音節詞彙有最佳辨識率各自發生在 Gaussian Number 為 7、4、1 的情況，如<圖 12>所示。在 SVM 的方法中，2、3、4 音節詞彙各自使用的參數如<表 13>所示。

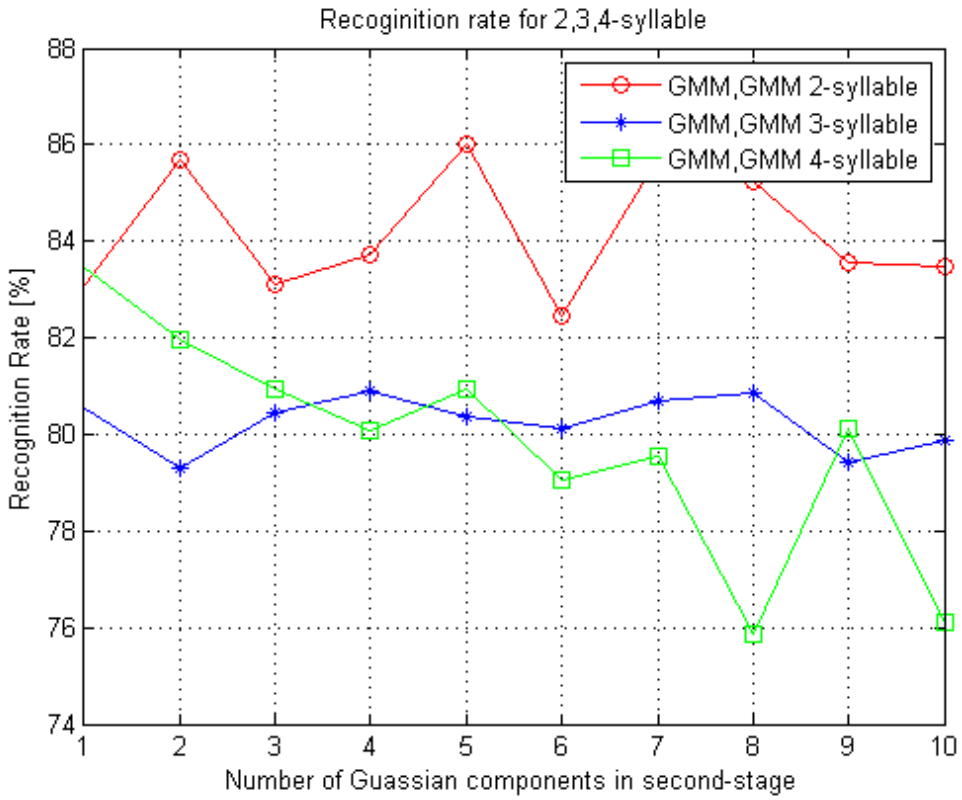

圖 12：2,3,4 音節詞彙使用 GMM 的各自辨識率

	Gamma (γ)	**Cost (C)**
2-syllable word	2^{-8}	2^{5}
3-syllable word	2^{-7}	2^{4}
4-syllable word	2^{-10}	2^{5}

表 13：2,3,4 音節詞彙於 SVM 中各自使用之參數

(5) median Pitch + Duration

第一步驟：

使用 GMM 分類，其最佳辨識率發生在 Gaussian number 為 64，而辨識率有 82.05%，如<圖 13>所示。

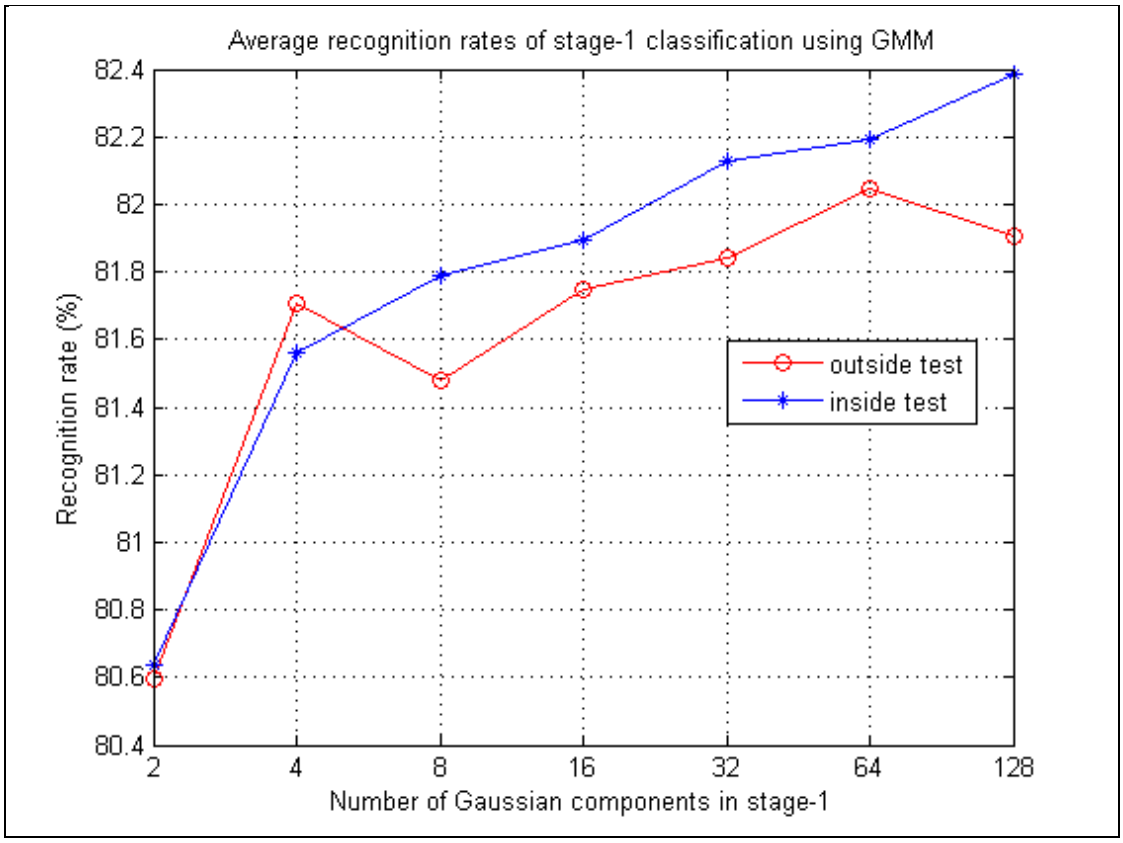

圖 13：第一步驟使用 GMM 分類器，3 組訓練/測試資料之平均辨識率

第二步驟：

分別使用 GMM 與 SVM 分類器的最佳辨識結果如下<表 14>。

Classifier / N-syllable word	GMM (%)	SVM (%)
2-syllable word	87.9912	88.223
3-syllable word	84.0523	84.5335
4-syllable word	85.3817	85.6469

表 14：第二步驟使用 GMM 與 SVM 於 2,3,4 音節詞彙的最佳辨識率

在 GMM 的方法中，2、3、4 音節詞彙有最佳辨識率各自發生在 Gaussian Number 為 4、1、2 的情況，如<圖 14>所示。在 SVM 的方法中，2、3、4 音節詞彙各自使用的參數如<表 15>所示。

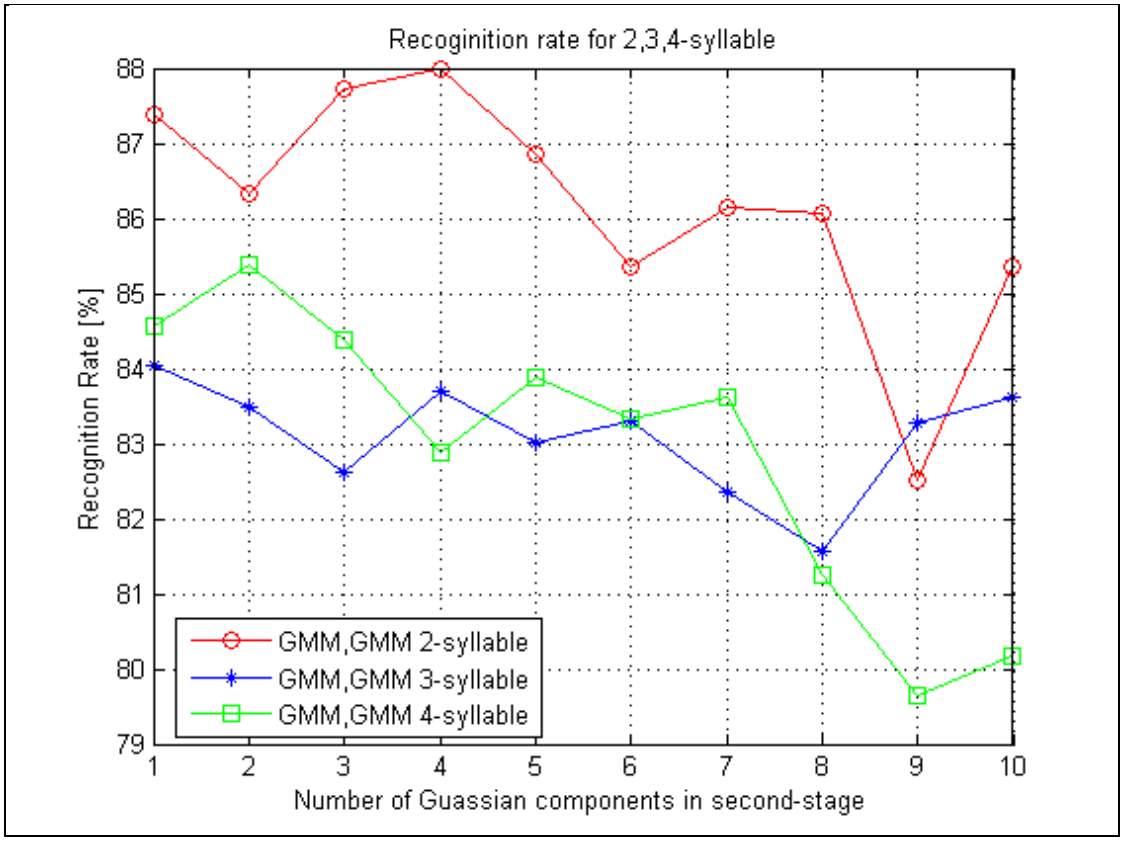

圖 14：2,3,4 音節詞彙使用 GMM 的各自辨識率

	Gamma (γ)	**Cost (C)**
2-syllable word	2^{-4}	2^{3}
3-syllable word	2^{-1}	2^{1}
4-syllable word	2^{-10}	2^{5}

表 15：2,3,4 音節詞彙於 SVM 中各自使用之參數

3.3.2 3 維特徵參數組合

3 維特徵參數主要有音量取平均值(簡稱 mean Volume)、音量取中位數(簡稱 median Volume)、音高取平均值(簡 mean Pitch)、音高取中位數(簡稱 median Pitch)、持續時間(Duration)，這些特徵參數均於 2.3.2 節說明如何取得及正規化的方法，而組合成以下 4 種做分類實驗：

(1) median Pitch + median Volume + Duration

第一步驟：

使用 GMM 分類，其最佳辨識率發生在 Gaussian number 為 16，而辨識率有 84.90%，如<圖 15>所示。

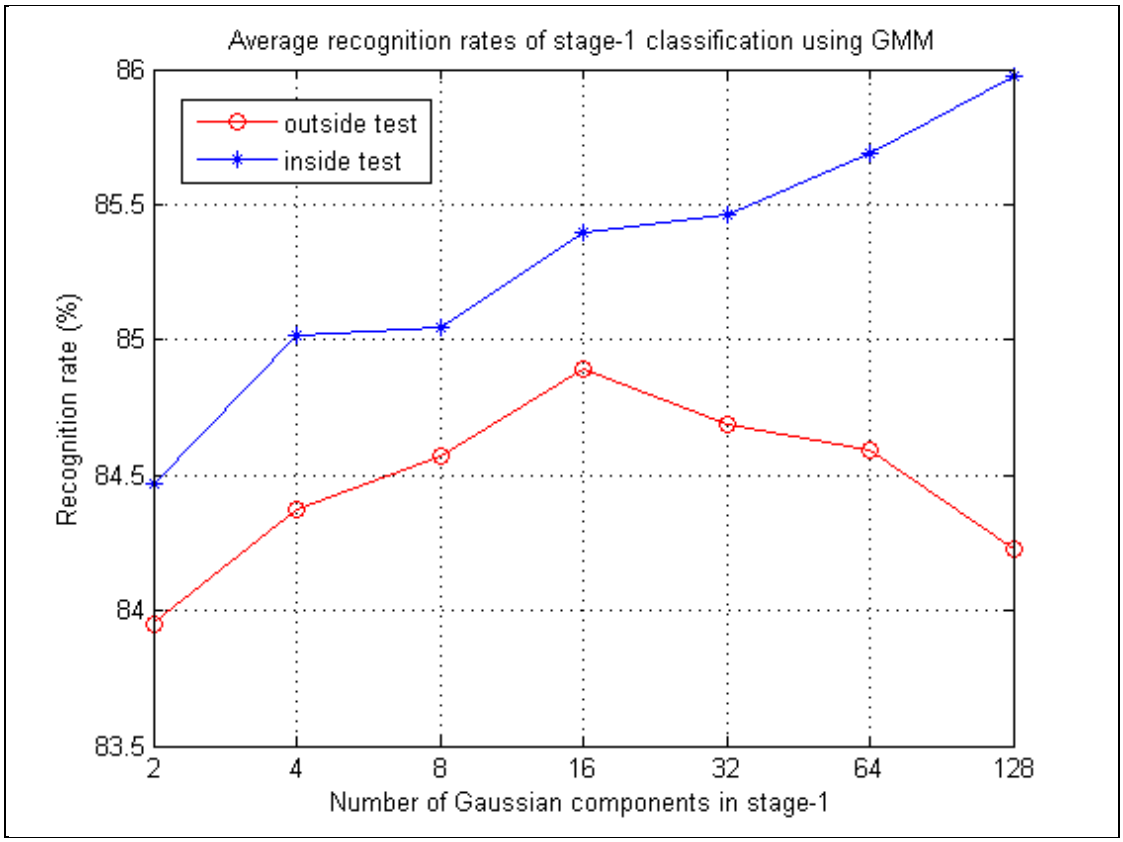

圖 15：第一步驟使用 GMM 分類器，3 組訓練/測試資料之平均辨識率

第二步驟：

分別使用 GMM 與 SVM 分類器的最佳辨識結果如下<表 16>。

Classifier / N-syllable word	GMM (%)	SVM (%)
2-syllable word	88.2648	89.29
3-syllable word	84.7116	86.3189
4-syllable word	83.9835	84.8915

表 16：第二步驟使用 GMM 與 SVM 於 2,3,4 音節詞彙的最佳辨識率

在 GMM 的方法中，2、3、4 音節詞彙有最佳辨識率各自發生在 Gaussian Number 為 6、6、2 的情況，如<圖 16>所示。在 SVM 的方法中，2、3、4 音節詞彙各自使用的參數如<表 17>所示。

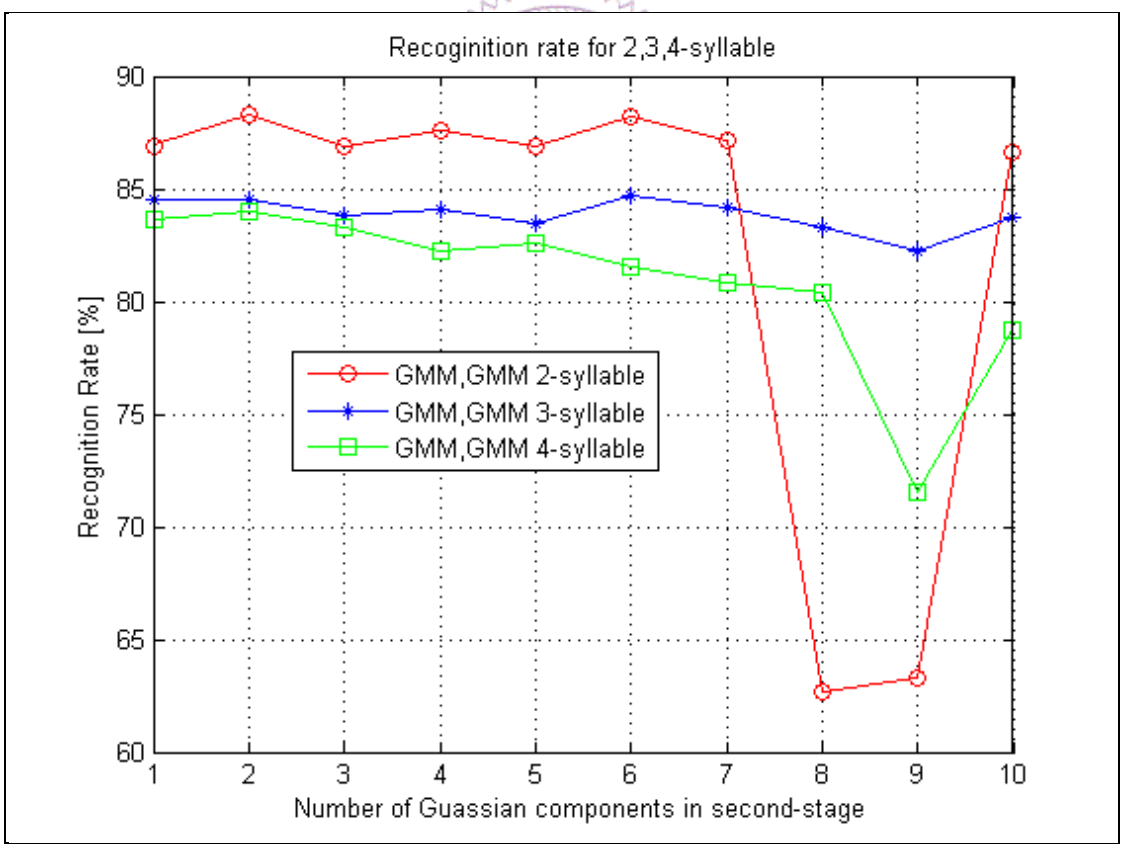

圖 16：2,3,4 音節詞彙使用 GMM 的各自辨識率

	Gamma (γ)	Cost (C)
2-syllable word	2^{-10}	2^{3}
3-syllable word	2^{-10}	2^{5}
4-syllable word	2^{-6}	2^{0}

表 17：2,3,4 音節詞彙於 SVM 中各自使用之參數

(2) mean Pitch + mean Volume + Duration

第一步驟：

使用 GMM 分類，其最佳辨識率發生在 Gaussian number 為 32，而辨識率有 85.40%，如<圖 17>所示。

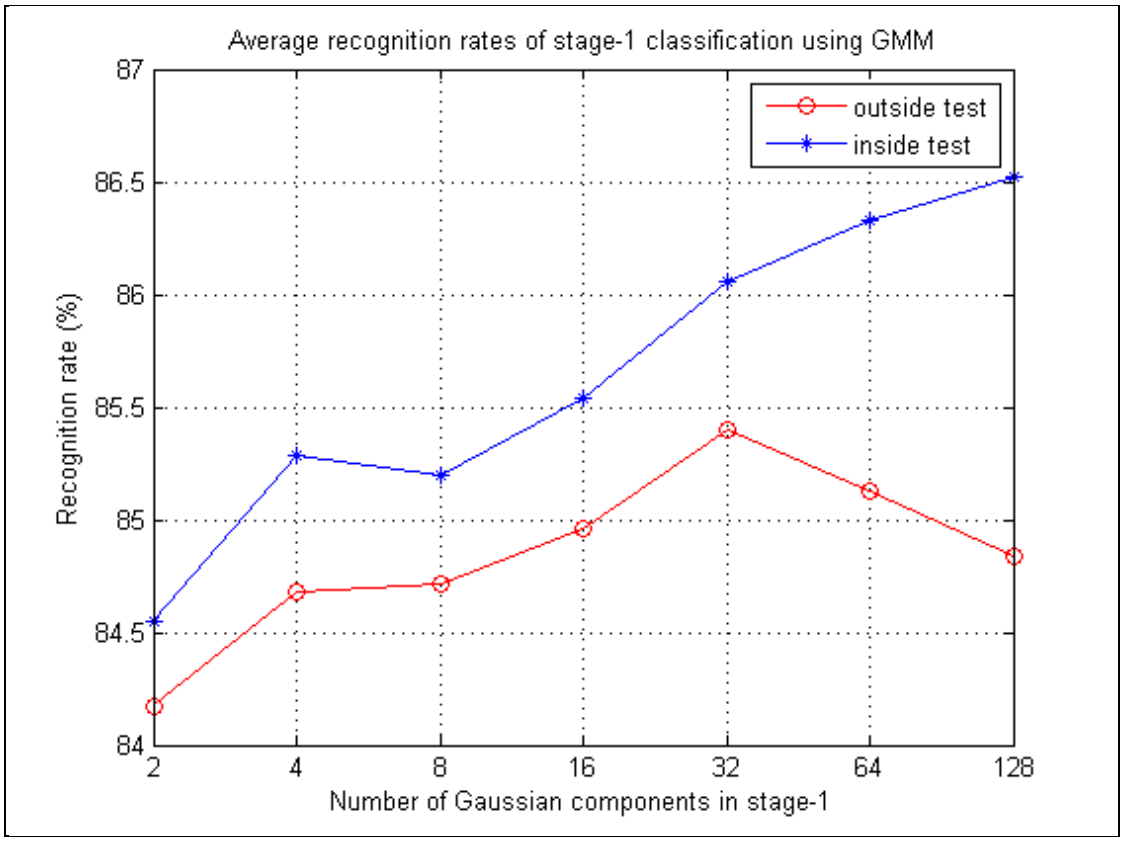

圖 17：第一步驟使用 GMM 分類器，3 組訓練/測試資料之平均辨識率

第二步驟：

分別使用 GMM 與 SVM 分類器的最佳辨識結果如下<表 18>。

Classifier / N-syllable word	GMM (%)	SVM (%)
2-syllable word	88.4638	89.7378
3-syllable word	85.6087	86.8486
4-syllable word	84.2909	84.613

表 18：第二步驟使用 GMM 與 SVM 於 2,3,4 音節詞彙的最佳辨識率

在 GMM 的方法中，2、3、4 音節詞彙有最佳辨識率各自發生在 Gaussian Number 為 1、2、1 的情況，如<圖 18>所示。在 SVM 的方法中，2、3、4 音節詞彙各自使用的參數如<表 19>所示。

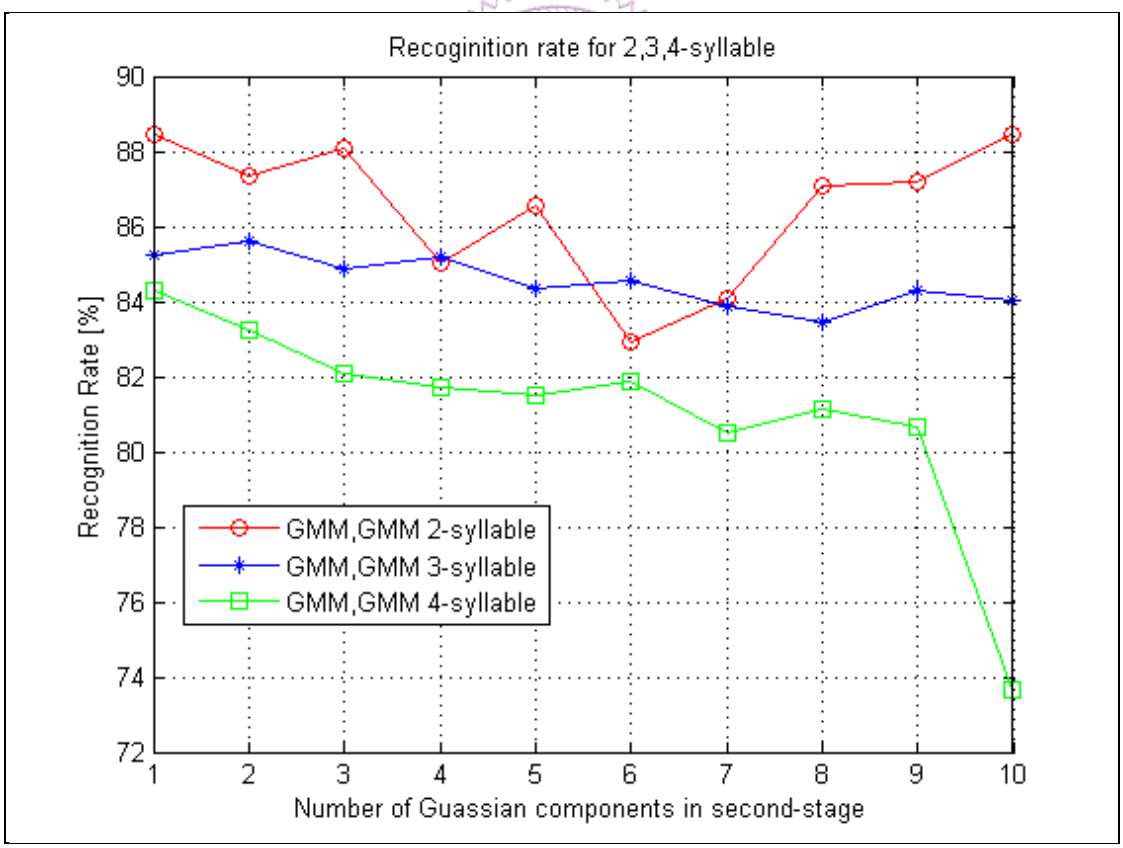

圖 18：2,3,4 音節詞彙使用 GMM 的各自辨識率

	Gamma (γ)	Cost (C)
2-syllable word	2^{-6}	2^{2}
3-syllable word	2^{-9}	2^{4}
4-syllable word	2^{-9}	2^{2}

表 19：2,3,4 音節詞彙於 SVM 中各自使用之參數

(3) <u>mean Pitch + median Volume + Duration</u>

第一步驟：

使用 GMM 分類，其最佳辨識率發生在 Gaussian number 為 8，而辨識率有 84.70%，如<圖 19>所示。

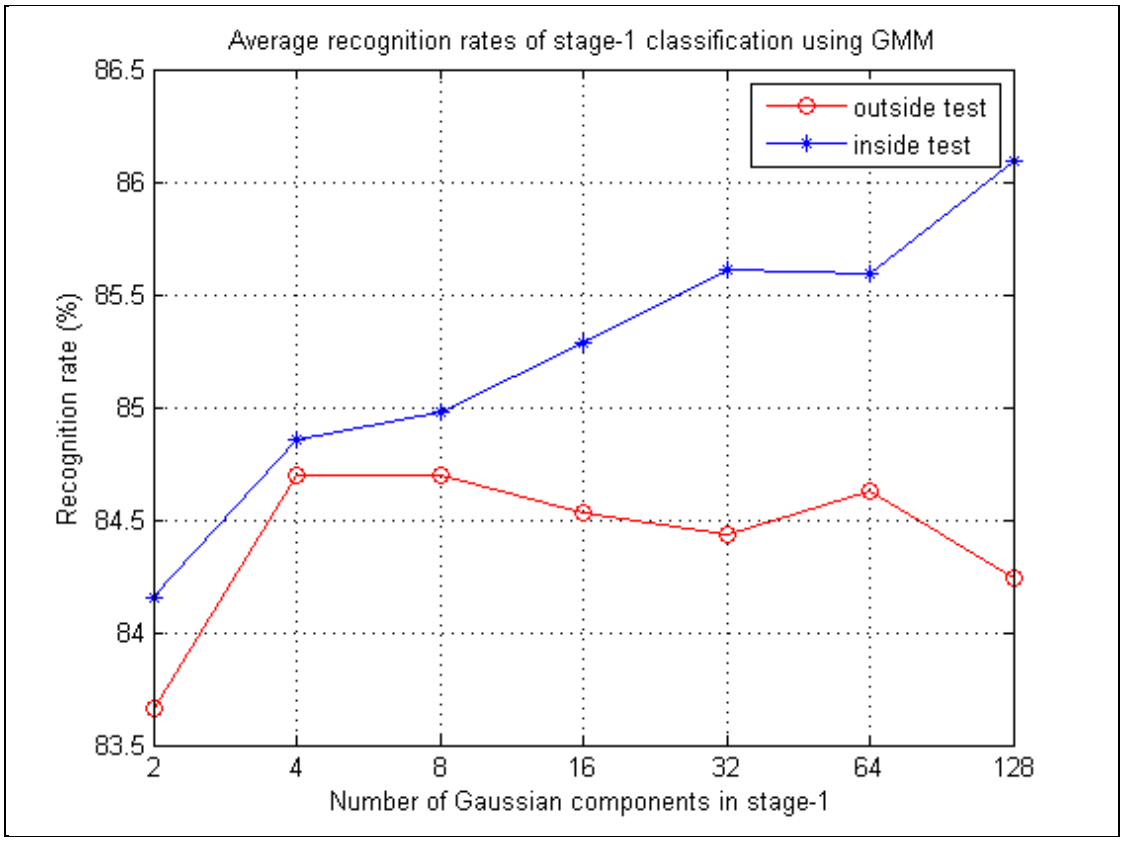

圖 19：第一步驟使用 GMM 分類器，3 組訓練/測試資料之平均辨識率

第二步驟：

分別使用 GMM 與 SVM 分類器的最佳辨識結果如下<表 20>。

Classifier / N-syllable word	GMM (%)	SVM (%)
2-syllable word	89.7784	90.3563
3-syllable word	85.3564	86.2877
4-syllable word	84.4411	84.9514

表 20：第二步驟使用 GMM 與 SVM 於 2,3,4 音節詞彙的最佳辨識率

在 GMM 的方法中，2、3、4 音節詞彙有最佳辨識率各自發生在 Gaussian Number 為 4、4、1 的情況，如<圖 20>所示。在 SVM 的方法中，2、3、4 音節詞彙各自使用的參數如<表 21>所示。

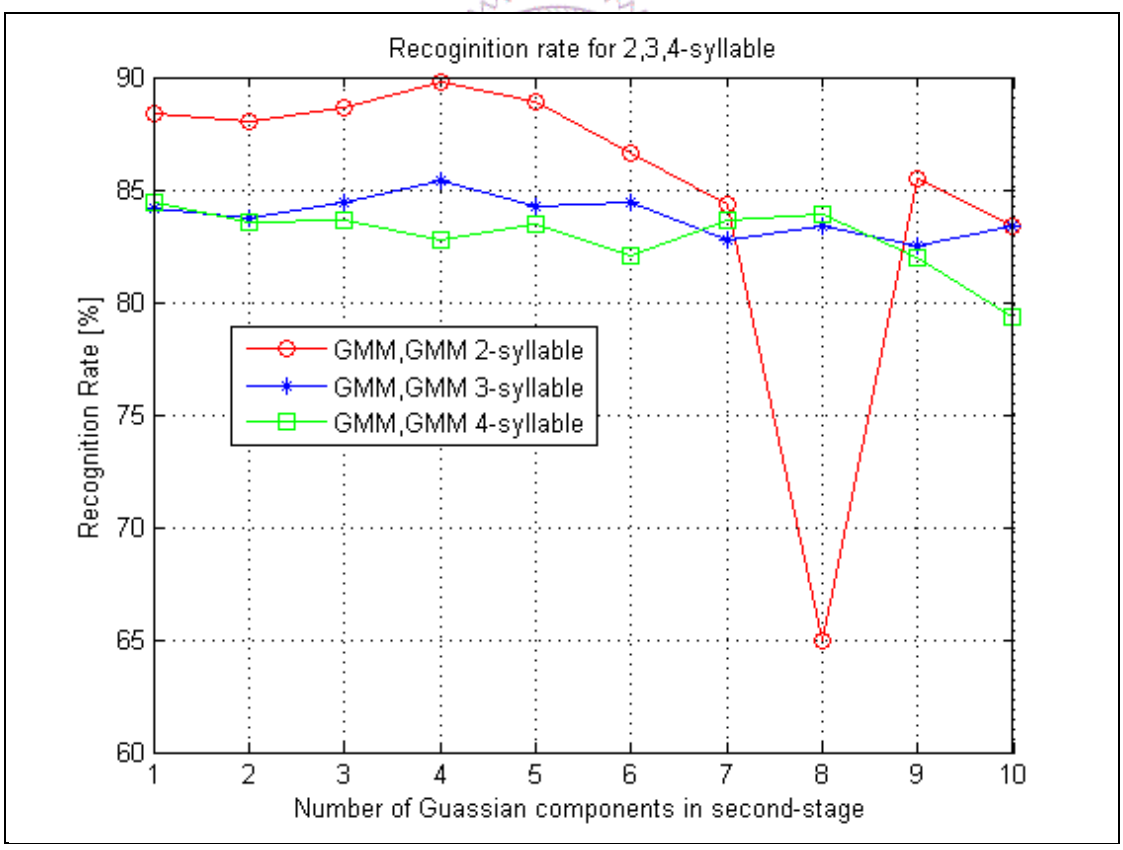

圖 20：2,3,4 音節詞彙使用 GMM 的各自辨識率

	Gamma (γ)	Cost (C)
2-syllable word	2^{-6}	2^{4}
3-syllable word	2^{-5}	2^{1}
4-syllable word	2^{-6}	2^{1}

表 21：2,3,4 音節詞彙於 SVM 中各自使用之參數

(4) median Pitch + mean Volume + Duration

第一步驟：

使用 GMM 分類，其最佳辨識率發生在 Gaussian number 為 32，而辨識率有 85.30%，如<圖 21>所示。

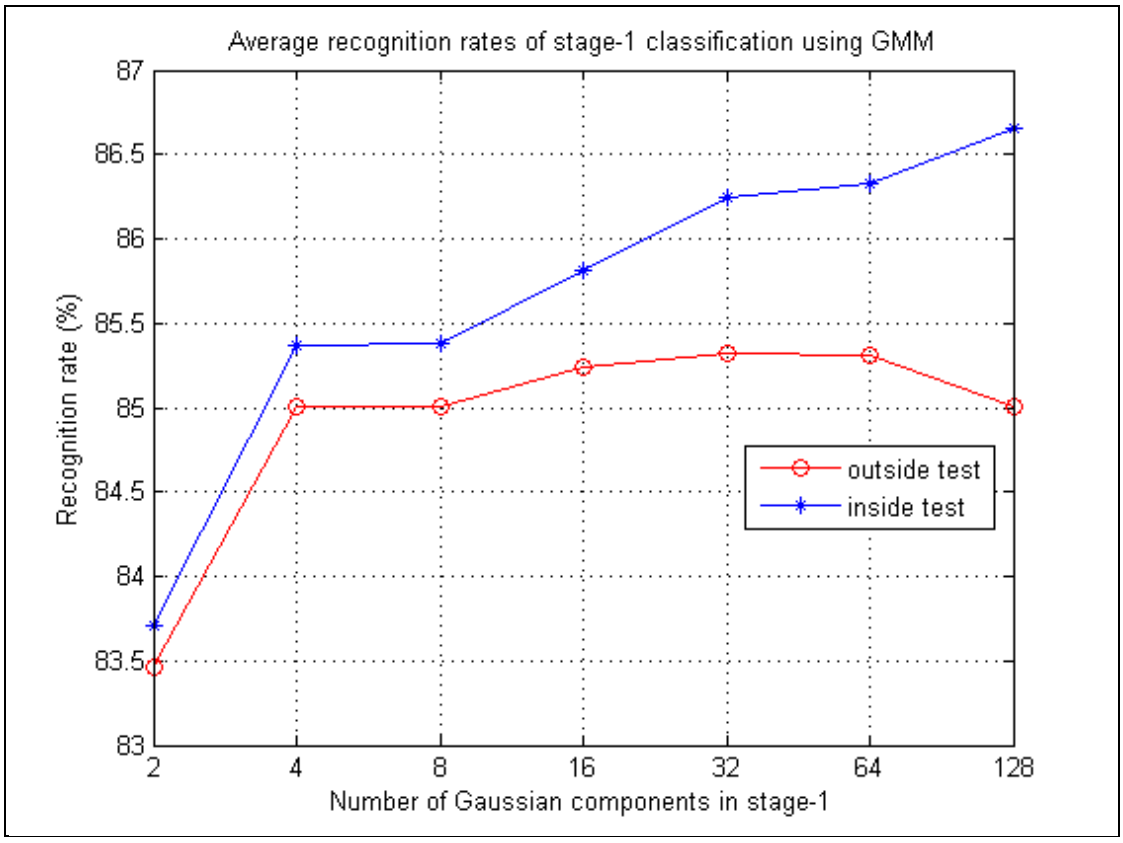

圖 21：第一步驟使用 GMM 分類器，3 組訓練/測試資料之平均辨識率

第二步驟：

分別使用 GMM 與 SVM 分類器的最佳辨識結果如下<表 22>。

Classifier / N-syllable word	GMM (%)	SVM (%)
2-syllable word	88.4137	89.4394
3-syllable word	85.5105	86.4994
4-syllable word	84.5928	84.161

表 22：第二步驟使用 GMM 與 SVM 於 2,3,4 音節詞彙的最佳辨識率

在 GMM 的方法中，2、3、4 音節詞彙有最佳辨識率各自發生在 Gaussian Number 為 3、4、1 的情況，如<圖 20>所示。在 SVM 的方法中，2、3、4 音節詞彙各自使用的參數如<表 23>所示。

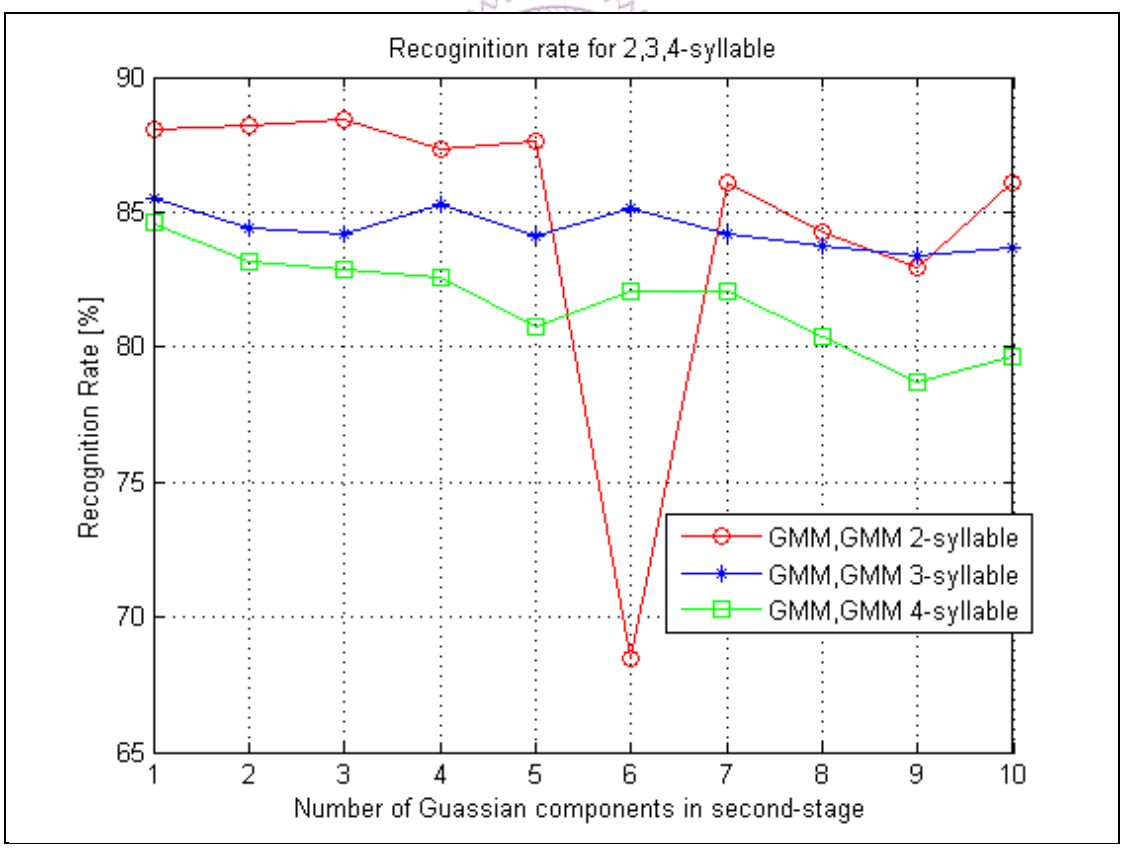

圖 22：2,3,4 音節詞彙使用 GMM 的各自辨識率

	Gamma (γ)	Cost (C)
2-syllable word	2^{-7}	2^{2}
3-syllable word	2^{-7}	2^{2}
4-syllable word	2^{-8}	2^{3}

表 23：2,3,4 音節詞彙於 SVM 中各自使用之參數

3.3.3 6 維特徵參數組合

6 維特徵參數的選取是使用 sequential forward selection(SFS)由 Whitney 在 1971 年提出的方法，主要使用 k 個最近鄰居分類法(KNNR)和一次挑一個(level-one-out, LOO)辨識率預估法，而其步驟主要為(1)第一個挑選的特徵必定是辨識率最高的特徵。(2)下一個挑選的特徵必定是和原本已選取的特徵合併後，辨識率最高的一個。重覆步驟(2)直到找到最佳辨識率的特徵組合。而主要的特徵參數為以下 9 個：對音高向量及音量向量取最大值(max)、平均值(mean)、中位數(median)、變異數(variance)以及持續時間(Duration)。結果如<圖 23>所示，由左至右分別為最初始的單一特徵的辨識，慢慢的增加一個最佳的辨識結果，最右邊則為最後 9 個特徵，可以得知最佳辨識率為 83.7%，而其特徵參數的組合為對音高向量取最大值(簡稱 maxP)、取中位數(medianP)、取平均值(meanP)，對音量向量取中位數(medianV)、取變異數(varV)以及持續時間(duration)。而圖中可以看到維度越高，在 SFS 的辨識率則會慢慢的增長。

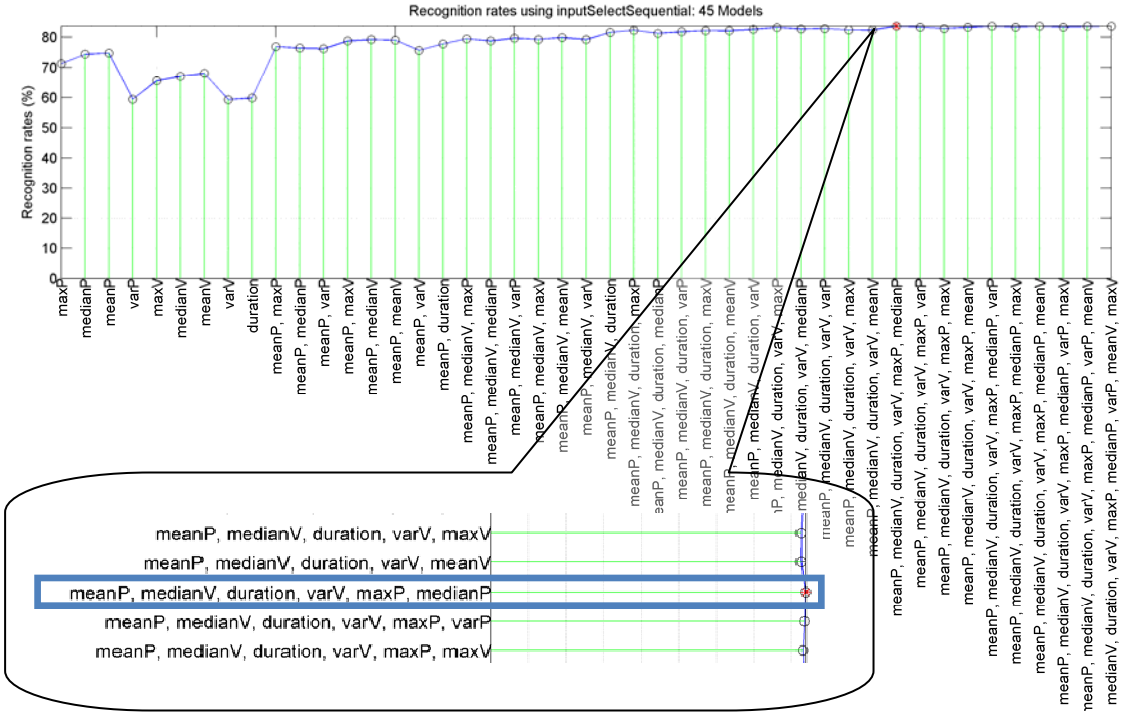

圖 23：使用 SFS 選取最佳特徵組合

maxP + medianP + meanP + medianV + varV + Duration

第一步驟：

使用 GMM 分類，其最佳辨識率發生在 Gaussian number 為 64，而辨識率有 83.37%，如<圖 24>所示。

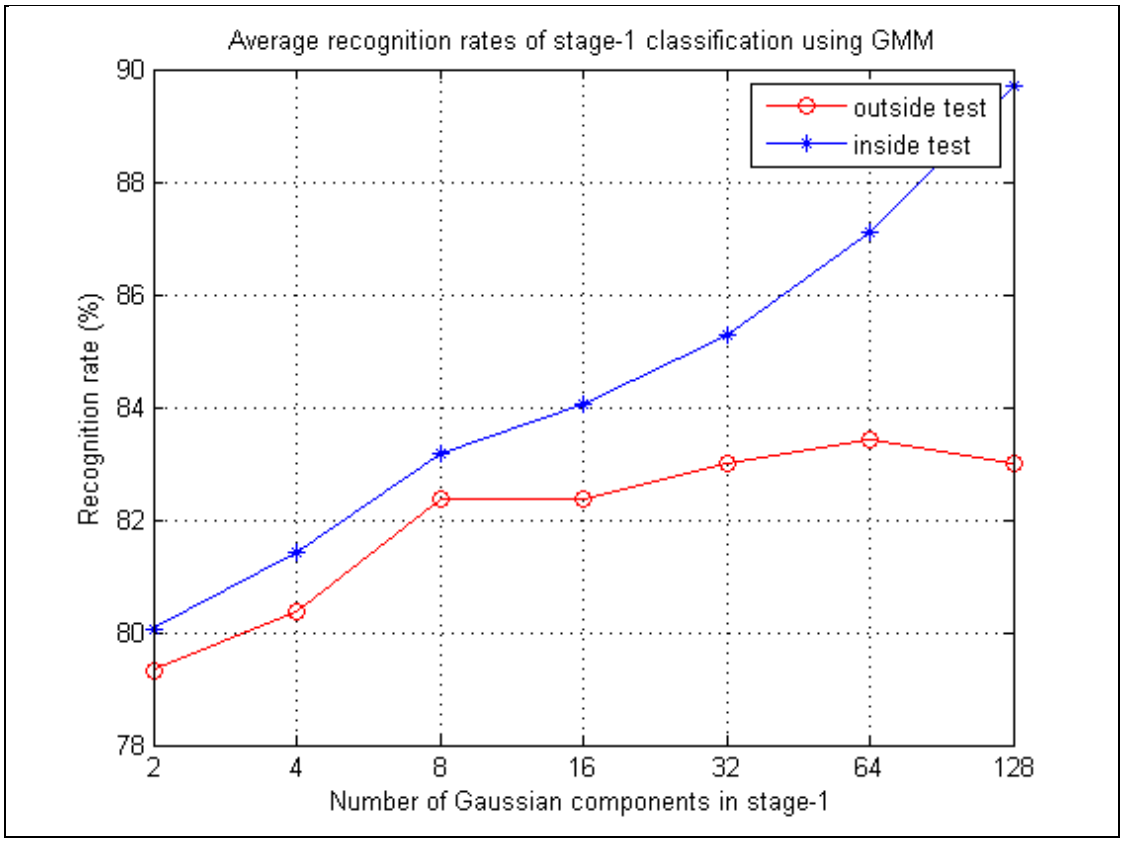

圖 24：第一步驟使用 GMM 分類器，3 組訓練/測試資料之平均辨識率

第二步驟：

分別使用 GMM 與 SVM 分類器的最佳辨識結果如下<表 24>。

Classifier / N-syllable word	GMM (%)	SVM (%)
2-syllable word	85.9409	87.6245
3-syllable word	81.4523	83.6749
4-syllable word	83.2222	82.6387

表 24：第二步驟使用 GMM 與 SVM 於 2,3,4 音節詞彙的最佳辨識率

在 GMM 的方法中，2、3、4 音節詞彙有最佳辨識率各自發生在 Gaussian Number 為 2、2、6 的情況，如<圖 25>所示。在 SVM 的方法中，2、3、4 音節詞彙各自使用的參數如<表 25>所示。

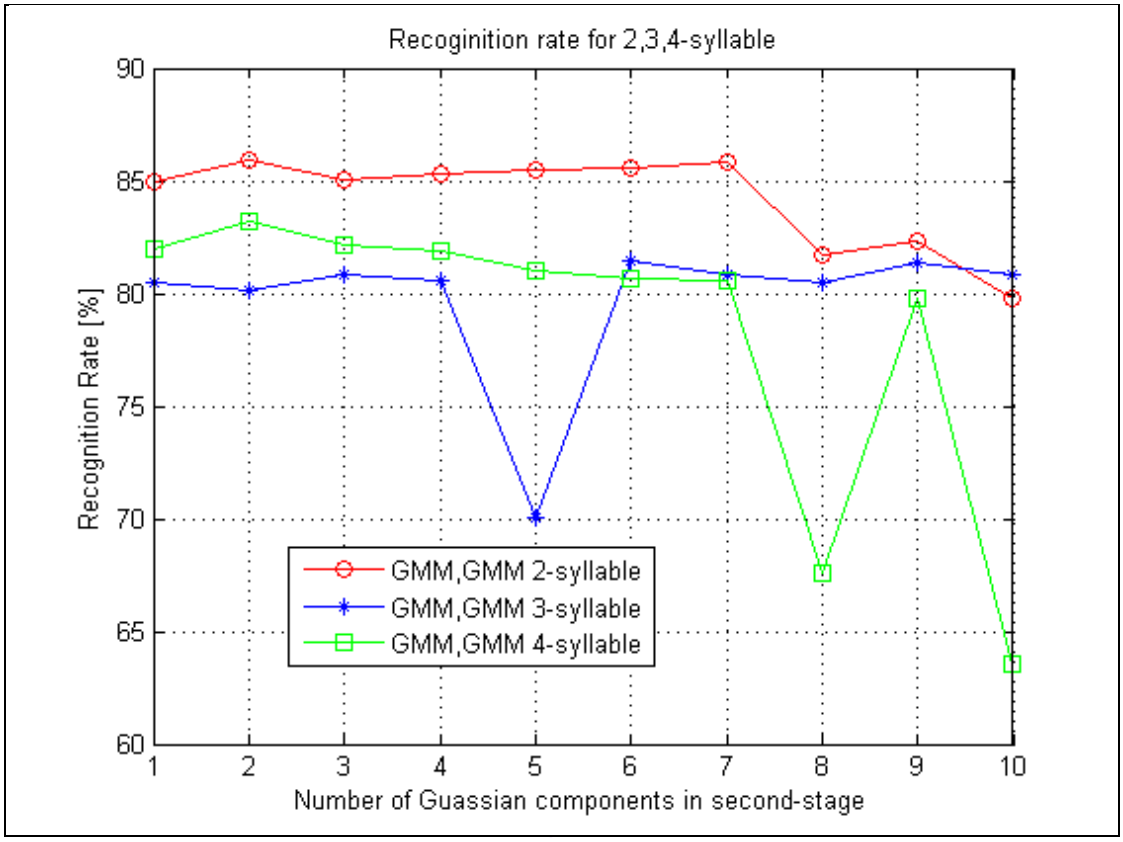

圖 25：2,3,4 音節詞彙使用 GMM 的各自辨識率

	Gamma (γ)	**Cost (C)**
2-syllable word	2^{-10}	2^{2}
3-syllable word	2^{-8}	2^{-1}
4-syllable word	2^{-10}	2^{0}

表 25：,3,4 音節詞彙於 SVM 中各自使用之參數

3.3.4 9 維特徵參數組合

9 維特徵參數分別為對音高向量及音量向量取最大值(max)、平均值(mean)、中位數(median)、變異數(variance)以及持續時間(Duration)。

第一步驟

使用 GMM 分類，其最佳辨識率發生在 Gaussian number 為 8，而辨識率有 82.90%，如<圖 26>所示。

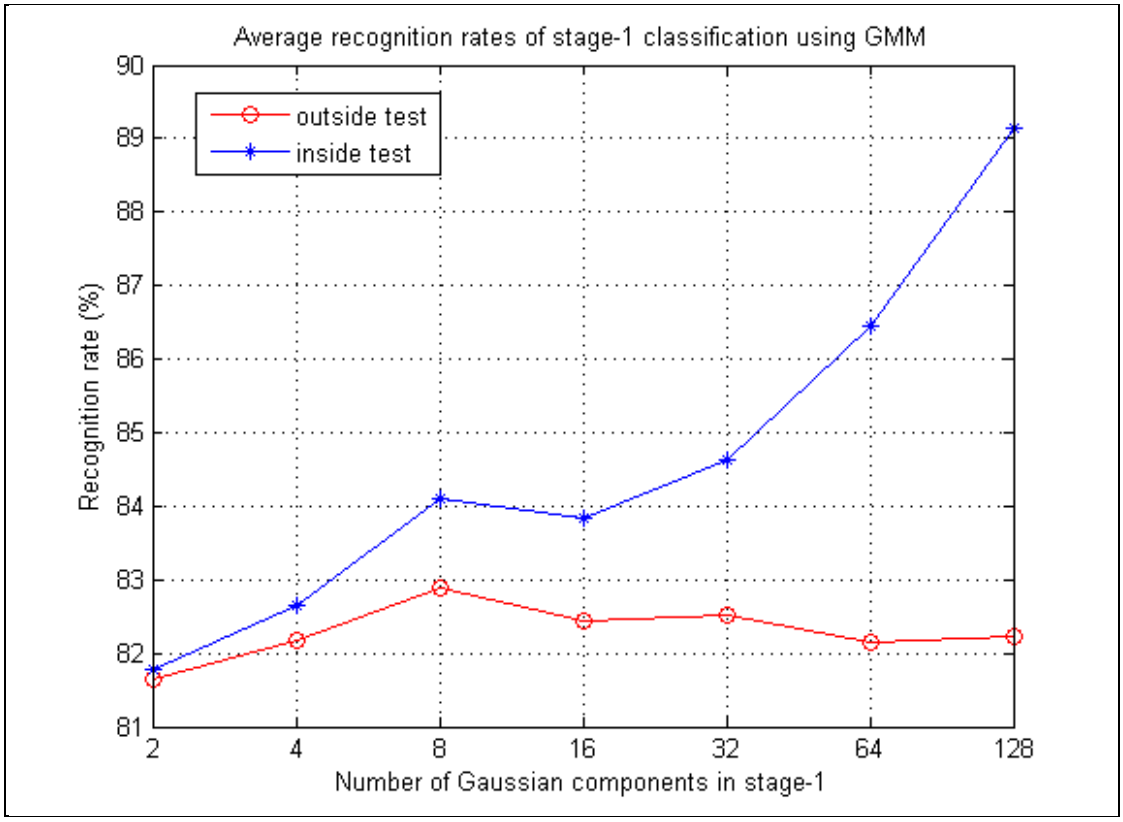

圖 26：第一步驟使用 GMM 分類器，3 組訓練/測試資料之平均辨識率

第二步驟

分別使用 GMM 與 SVM 分類器的最佳辨識結果如下<表 26>。

Classifier / N-syllable word	GMM (%)	SVM (%)
2-syllable word	86.8625	88.1854
3-syllable word	80.4849	82.8172
4-syllable word	79.9091	81.3705

表 26：第二步驟使用 GMM 與 SVM 於 2,3,4 音節詞彙的最佳辨識率

在 GMM 的方法中，2、3、4 音節詞彙有最佳辨識率各自發生在 Gaussian Number 為 2 的情況，如<圖 27>所示。在 SVM 的方法中，2、3、4 音節詞彙各自使用的參數如<表 27>所示。

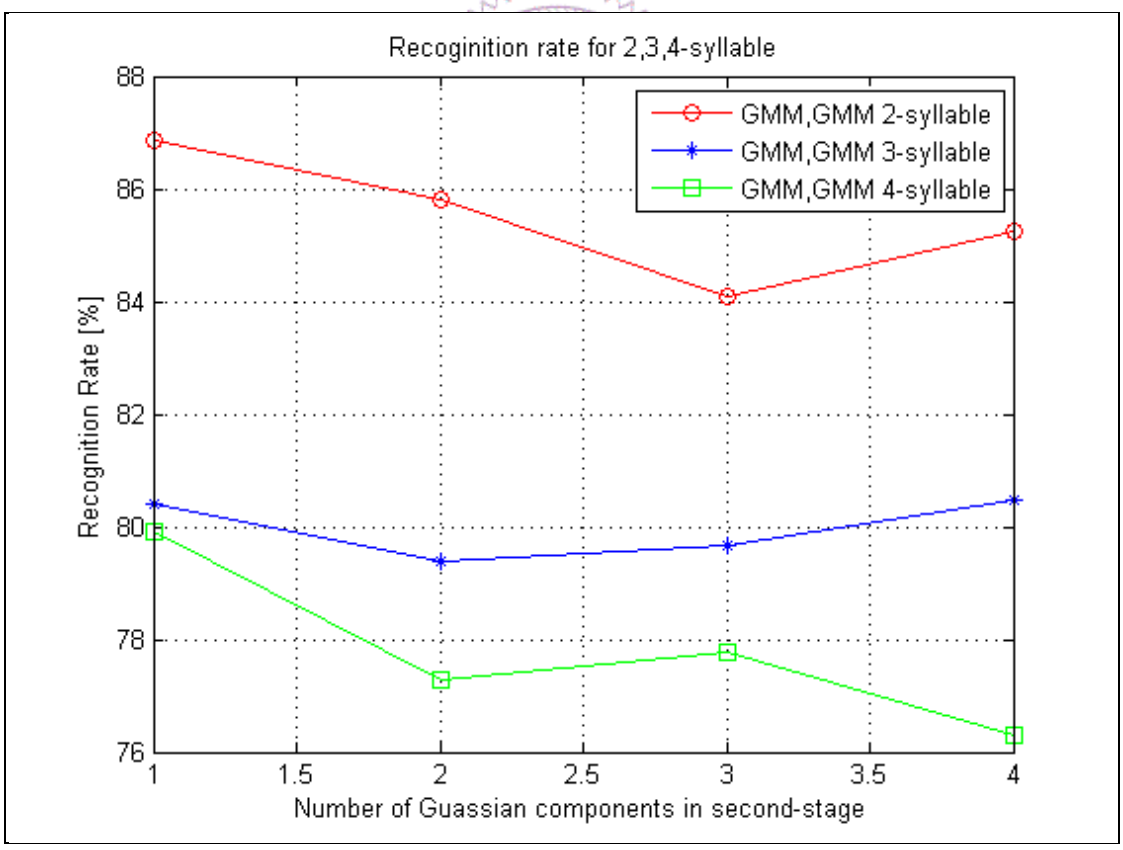

圖 27：2,3,4 音節詞彙使用 GMM 的各自辨識率

	Gamma (γ)	Cost (C)
2-syllable word	2^{-8}	2^{0}
3-syllable word	2^{-9}	2^{1}
4-syllable word	2^{-10}	2^{1}

表 27：2,3,4 音節詞彙於 SVM 中各自使用之參數

3.4 實驗分析

在一維辨識方法中(3.2 節)，有最佳的辨識結果為使用音高向量取中位數(簡稱 median Pitch)的方法，其辨識率達到 82.5768%，其次為使用音高向量取平均值(簡稱 mean Pitch)的方法，其辨識率為 82.0204%。由此可知，只使用音高做為辨識特徵參數已經有粗略的辨識效果。

下列<圖 28>為第二步驟多維辨識方法各組合的辨識率比較圖表，縱座標為辨識率，橫座標為各組合，分別從 2 維、3 維、6 維以及 9 維的順序排列，此部分可參考<圖 29>，圖中依序的表示各組合位於<圖 28>橫座標各點之值。

於<圖 28>中，劃出 2、3、4 個音節詞彙的所有辨識率，圓圈(●)代表 2 個音節詞彙(2-syllable word)，三角形(▲)代表三個音節詞彙(3-syllable word)，正方形(■)代表四個音節詞彙(4-syllable word)，而虛線與直線分別代表使用 GMM 以及 SVM 分類器。

整體而言，使用 SVM 的分類器比使用 GMM 的分類器的辨識結果還要好，最佳可提高 1%；2 個音節詞彙的辨識率均比 3、4 個音節詞彙的辨識率佳，即表示音節數越多，辨識率會受影響而下降；在 2 個音節詞彙中，有最佳辨識率的特徵參數為{mean Pitch, median Volume, Duration}，使用 SVM 分類器的辨識率為 90.36%，使用 GMM 分類器的辨識率為 89.78%；在 3 個音節詞彙中，有最佳辨識率的特徵參數為{mean Pitch, mean Volume, Duration}，使用 SVM 分類器的辨

識率為 86.85%，使用 GMM 分類器的辨識率為 85.61%；在 4 個音節詞彙中，有最佳辨識率的特徵參數為{median Pitch, Duration}，使用 SVM 分類器的辨識率為 85.65%，使用 GMM 分類器的辨識率為 85.38%。<表 28>為各音節詞彙使用 GMM 以及 SVM 的最佳辨識率以及所使用的特徵組合

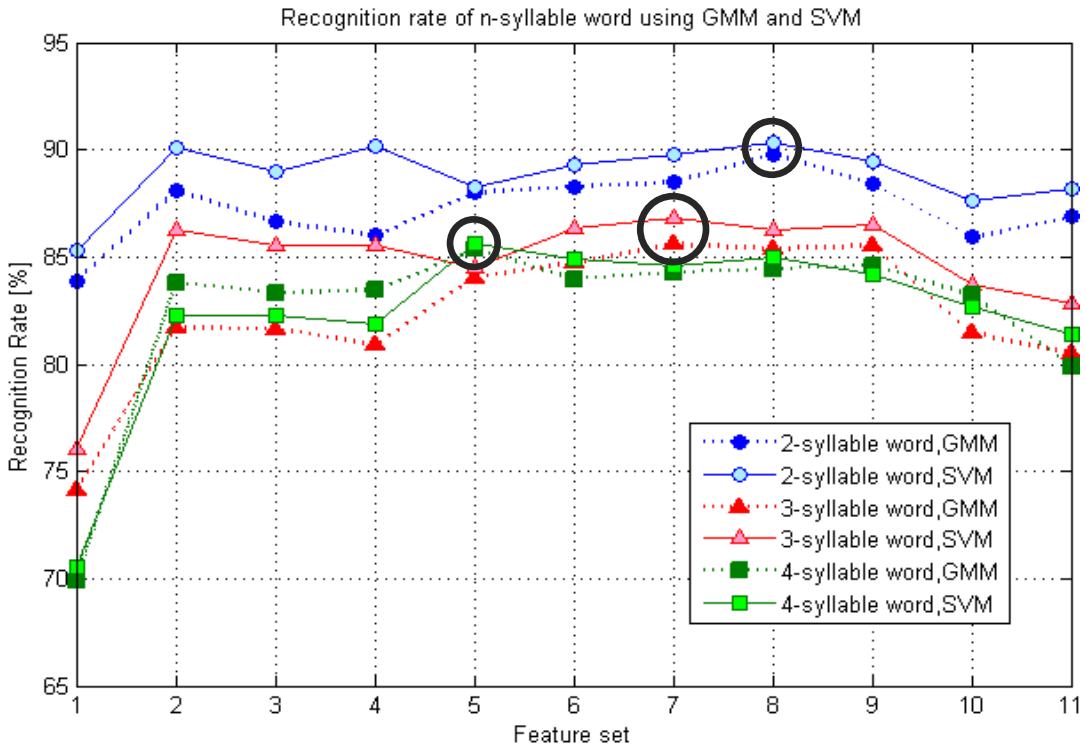

圖 28：於第二步驟，2,3,4 音節詞彙
使用 GMM 及 SVM 各種特徵參數組合之辨識率

X Label
(1) mean V + D
(2) median P + median V
(3) mean P + mean V
(4) median P + mean V
(5) median P + D
(6) median P + median V + D
(7) mean P + mean V + D
(8) mean P + median V + D
(9) median P + mean V + D
(10) max,median,mean P + median,var V + D
(11) max,median,mean,variance P +
max,median,mean,variance V + D

V:Volume P:Pitch D:Duration

圖 29：橫座標各特徵參數組合

Classifier / N-syllable words	GMM(%)	SVM(%)
2-syllable word {meanP,medianV,D}	89.78%	90.36%
3-syllable word {meanP,meanV,D}	85.61%	86.85%
4-syllable word {medianP,D}	85.38%	85.65%

表 28：2,3,4 音節詞彙使用 GMM 及 SVM 之最佳辨識率

觀察{median Pitch, Duration}這兩個特徵參數的組合，發現第一步驟分類母音是否為重音的辨識率只有 82.05%，遠不如其他特徵參數組合的辨識率，原因如<圖 30>中所示，左圖為第一步驟的示意圖，重音與非重音資料混淆的部分(即交集的部分)較多，但是右圖第二步驟中 4 個音節詞彙的重音與非重音的資料分布，重音與非重音混淆的情況較少，因此辨識率提高了。因此各個音節詞彙需要使用不同的特徵組合才能達到各自的最佳辨識結果。

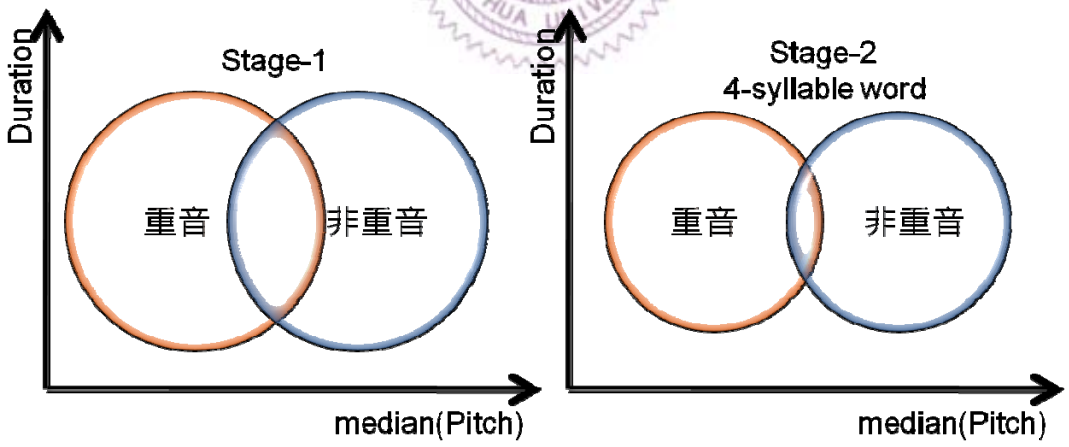

圖 30：重音與非重音分布示意圖

使用有最佳辨識結果的特徵組合{mean Pitch, median Volume, Duration}，與簡單條件式判斷重音的方法做比較，max(stressed)是指使用一個詞彙各母音辨識為重音的 log likelihood 取有最大值的音節為重音；min(unstressed)是指使用一個詞彙各母音辨識為非重音的 log likelihood 取有最小值的音節為重音；

max(stressed-unstressed)是指一個詞彙各母音辨識為重音與非重音的 log likelihood 相減後取有最大值的音節為重音；KNNR (K=1,23)就是使用 1,2,3 個最近鄰居分類法後再取這三個辨識結果的平均；JT method[5]則是一個詞彙各母音先辨識是否為重音，若有兩個以上的母音為重音，則取重音的機 log likelihood 最大值的音節為重音，若沒有一個母音辨識為重音，則取非重音的 log likelihood 有最小值的音節為重音。

如<表 29>，簡單條件式判斷重音的方法中最佳辨識結果是使用重音與非重音 log likelihood 相減後取最大值的方法，辨識率達到 85.38%，但是使用 SVM 的方法最佳辨識率則有 90.36%，提升了約 5%的辨識率。

Simple-Rule method	
Max(stressed prob)	75.56%
Min(unstressed prob)	69.28%
Max(stressed-unstressed)	**85.38%**
Knnr , k=1,2,3	81.62%
JT method	72.5%
Parametric method	
GMM	**90.36%**
SVM	89.78%

表 29：Simple-Rule method 與 Parametric method 的比較

3.5 錯誤分析

依據 3.4 實驗分析可知，使用一維特徵參數辨識方法最佳辨識率為 82.5768%，而使用多維特徵參數的辨識2、3、4 個音節詞彙的最佳辨識率為 90.36%、86.85%、85.65%，因此本章節要來探討有最佳辨識率下的錯誤資料，討論錯誤發生的情況有哪些情形。

根據在 2 個音節詞彙的辨識方法中，有最佳辨識率的特徵參數{ mean Pitch, median Volume, Duration }，分析其第一步驟用 GMM 分類母音音素(vowel phone)後的錯誤資料，找出母音音素之錯誤排名，分為重音辨識成非重音以及非重音辨識成重音兩類型。下列表格為 Outside test 錯誤資料的分析，可以看出本來是重音的母音辨識錯誤佔全部錯誤資料的 18.38%是發生在 eh 的母音，17.11%發生在 ih 的母音，14.89%發生在 ae 的母音，而非重音的母音辨識錯誤的部分佔全部錯誤資料的 29.24%為 ah 的母音，18.52%為 ih 的音，17.35%為 er 的母音。

重音辨識成非重音 Stressed→Unstressed		非重音辨識成重音 Unstressed→Stressed	
eh(ɛ)	18.38%	ah(ʌ,ə)	29.24%
ih(ɪ)	17.11%	ih(ɪ)	18.52%
ae(æ)	14.89%	er(ɚ, ɝ)	17.35%

表 30：第一步驟辨識錯誤之母音排名

針對以上前三名辨識錯誤的重音與非重音母音，畫出其中三個特徵的分布，並與辨識正確母音的特徵做比較。<圖 31>為母音 eh 為重音的資料，• 點代表辨識正確，x 點代表辨識錯誤(意思是重音但辨識成非重音)，左邊的圖為 3 維特徵的圖形，右邊的圖為兩兩特徵為一組的分布圖，由此可看出 • 的區塊大致分布在音量、音高及持續時間較大的部分，而辨識錯誤的 x 區塊則有三個特徵較小趨勢。<圖 35>為母音 ih 為重音的資料，<圖 36>為母音 ae 為重音的資料，均有相似的分布情況。<圖 32>至<圖 34>分別代表了母音 eh 中音高、音量、時間的分布圖形，主要使用接收者操作特徵曲線(Receiver operating characteristic curve,簡稱 ROC)來觀察資料分布情況，而這三張圖中可以看到錯誤的資料(◯)與正常的資料(□)都為常態分布的圖形，然而正常資料的最高點均遠大於錯誤資料的最高點。

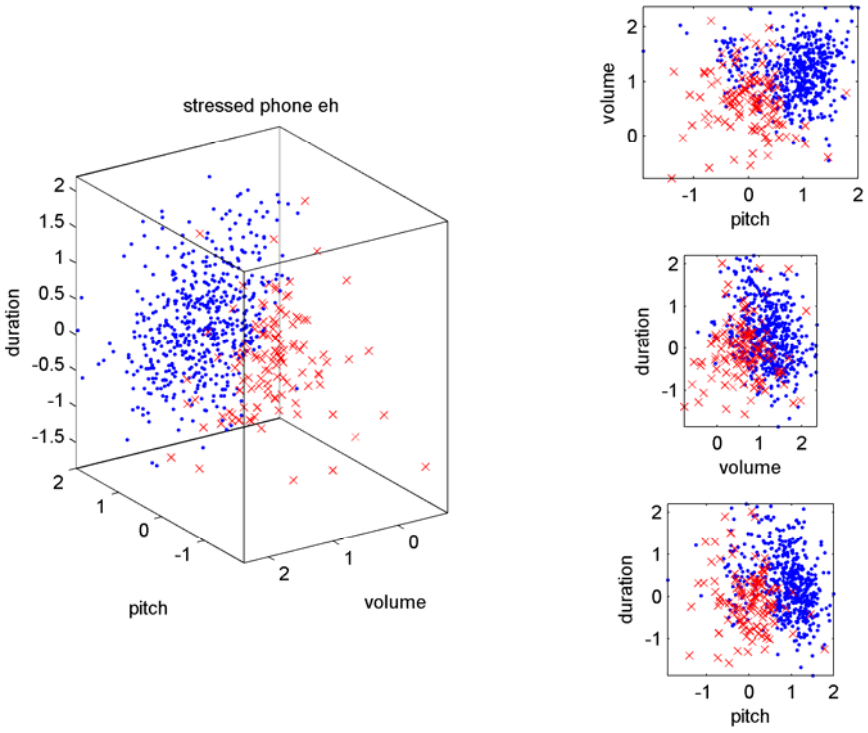

圖 31：eh 為重音的分布圖；x 點代表辨識錯誤，• 點代表辨識正確

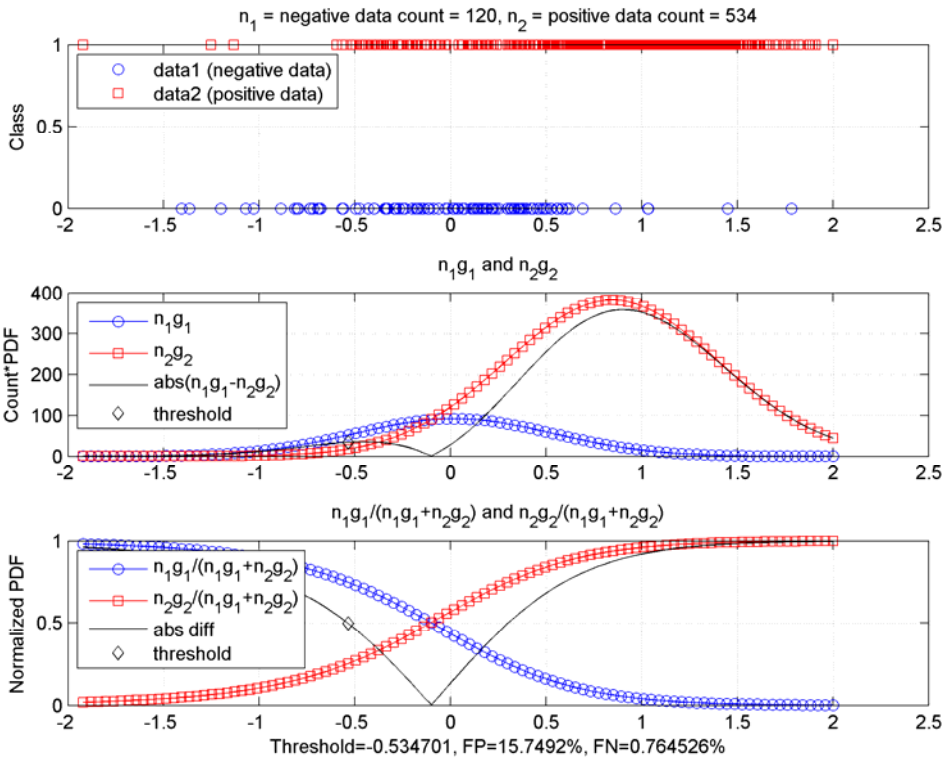

圖 32：eh 的音高特徵之分布

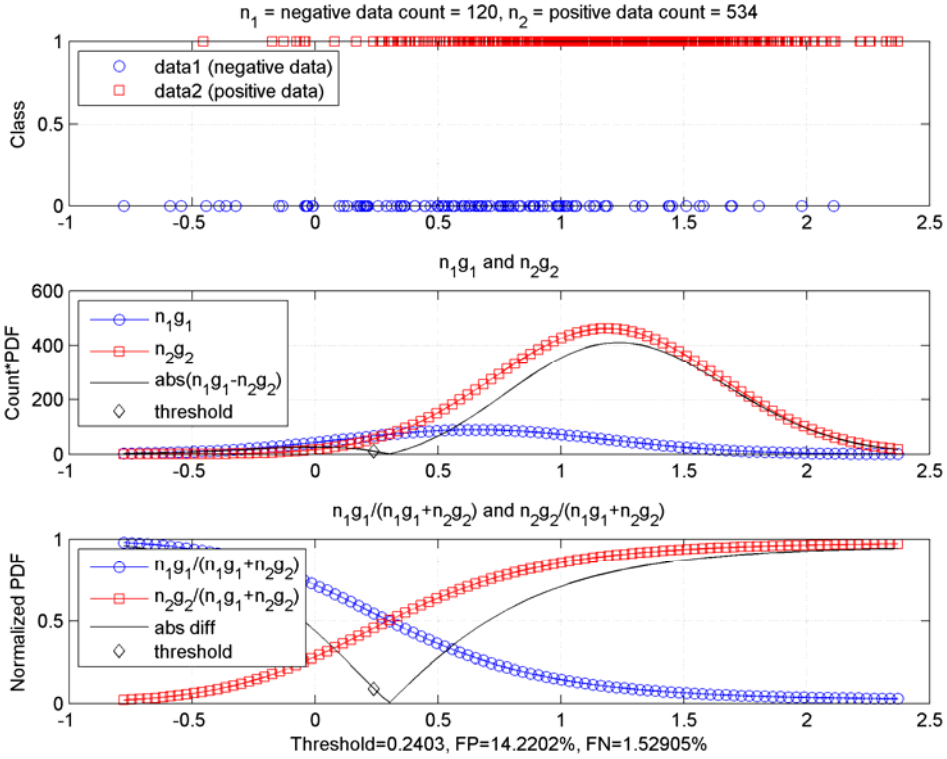

圖 33：eh 的音量特徵之分布

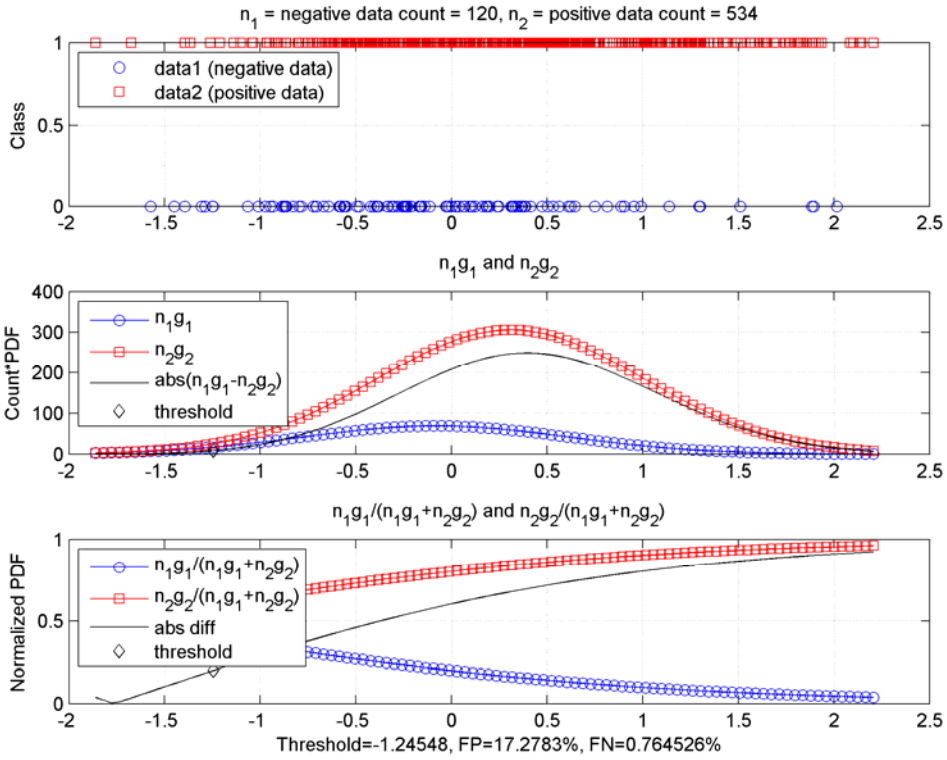

圖 34：eh 的時間特徵之分布

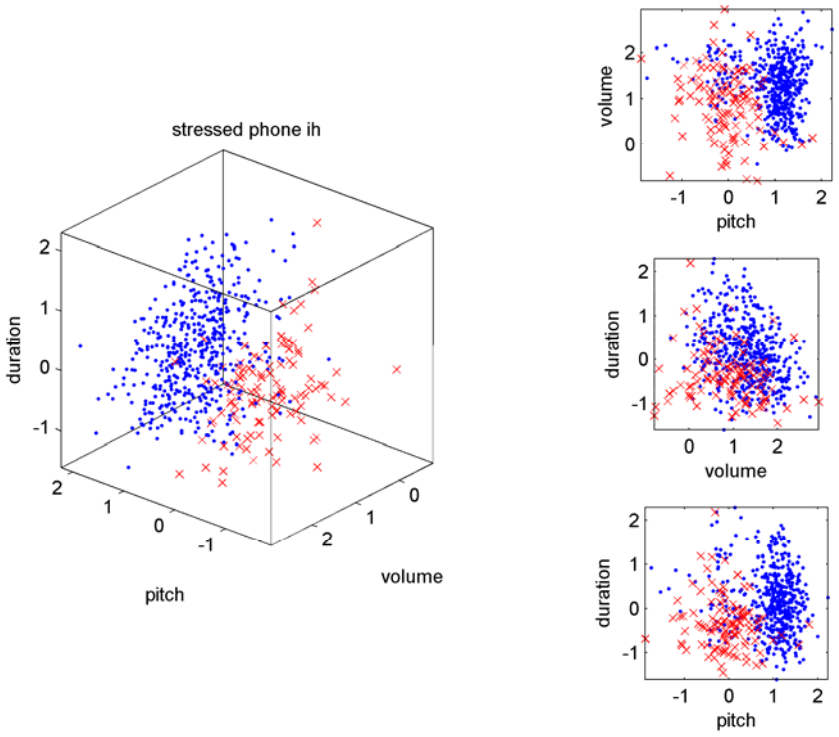

圖 35：ih 為重音的分布圖；x 點代表辨識錯誤，• 點代代表辨識正確

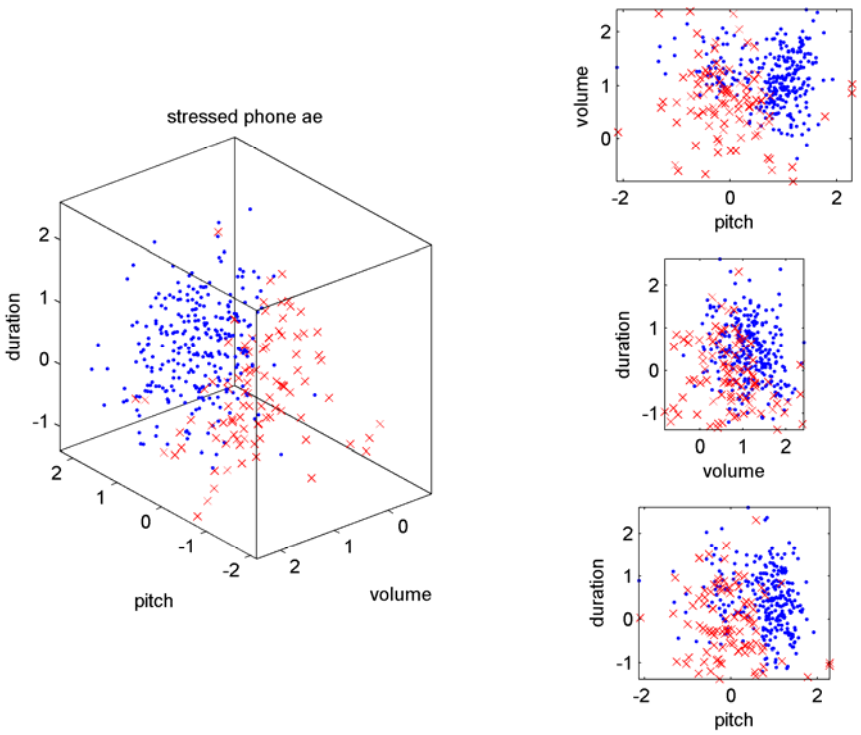

圖 36：ae 為重音的分布圖；x 點代表辨識錯誤，• 點代代表辨識正確

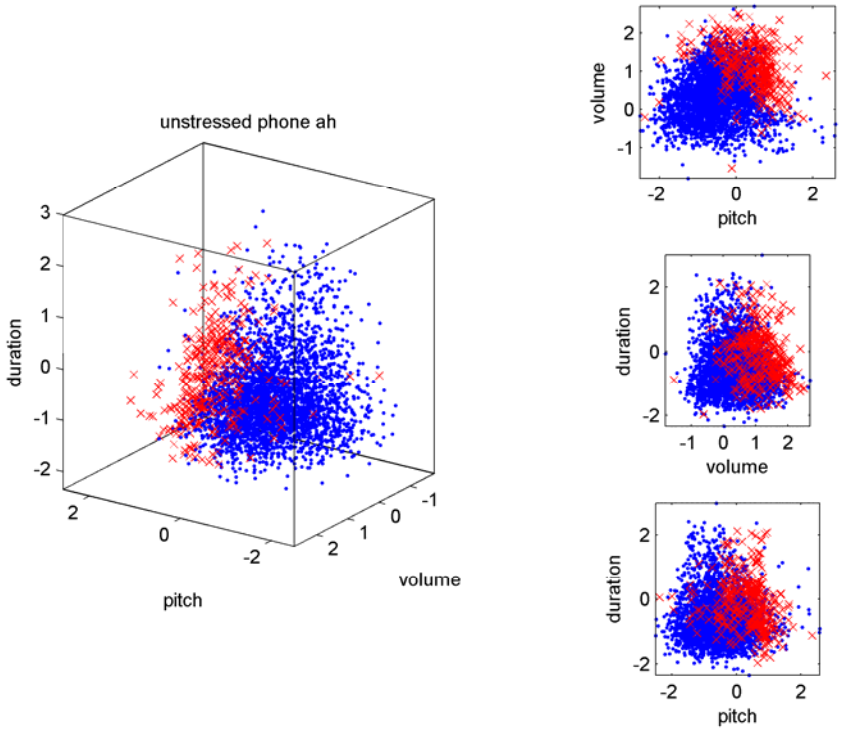

圖 37：ah 為非重音的分布圖；x 點代表辨識錯誤，• 點代表辨識正確

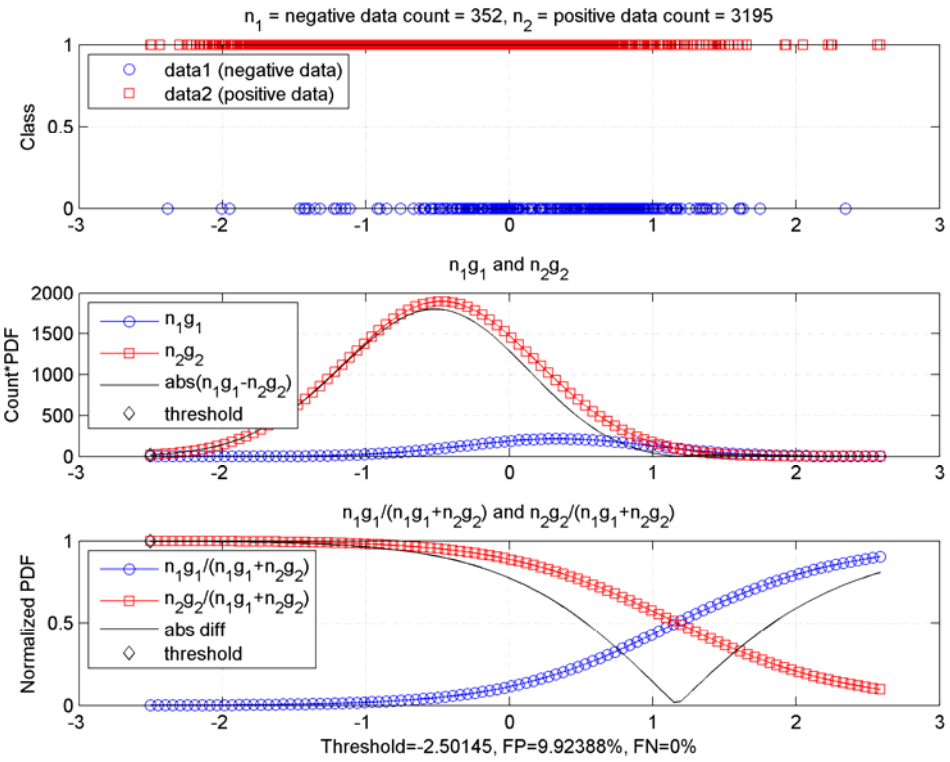

圖 38：ah 的音高特徵之分布

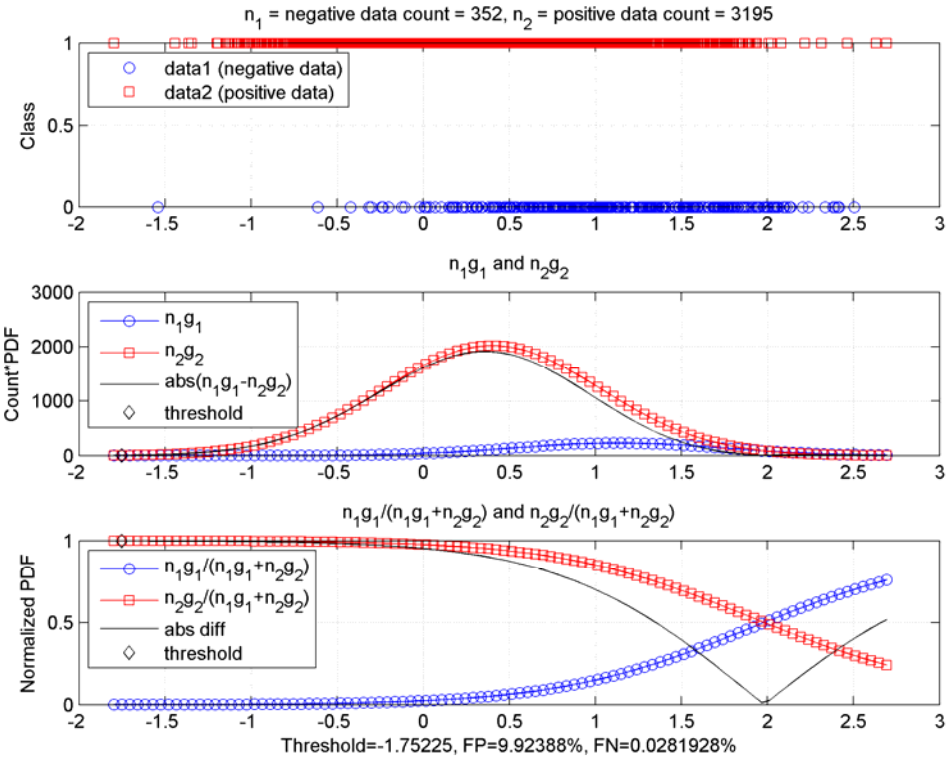

圖 39：ah 的音量特徵之分布

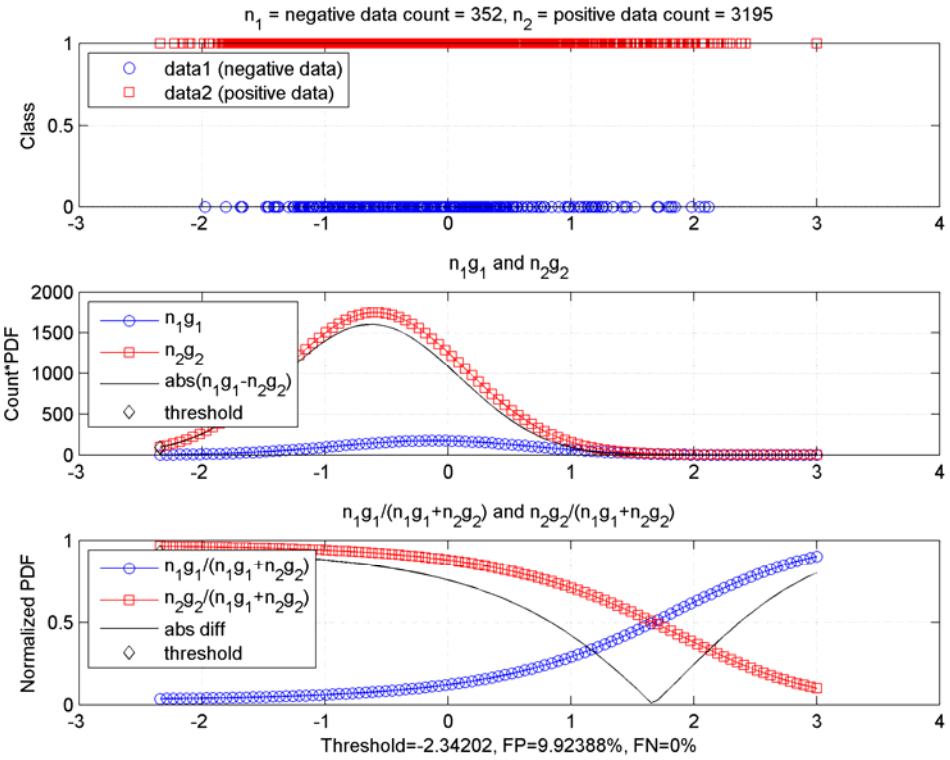

圖 40：ah 的時間特徵之分布

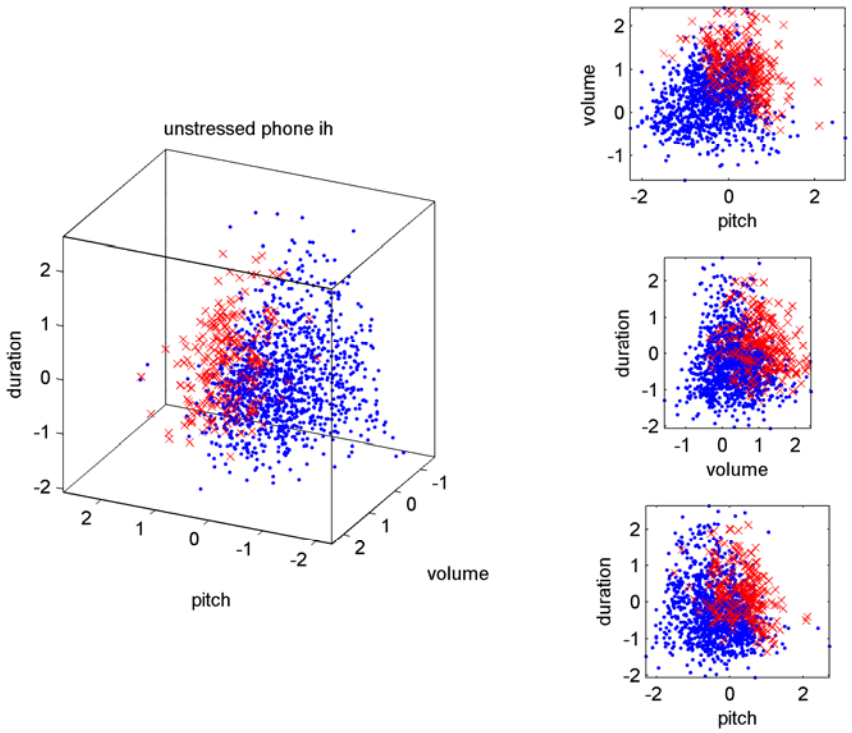

圖 41：ih 為非重音的分布圖；x 點代表辨識錯誤，• 點代表辨識正確

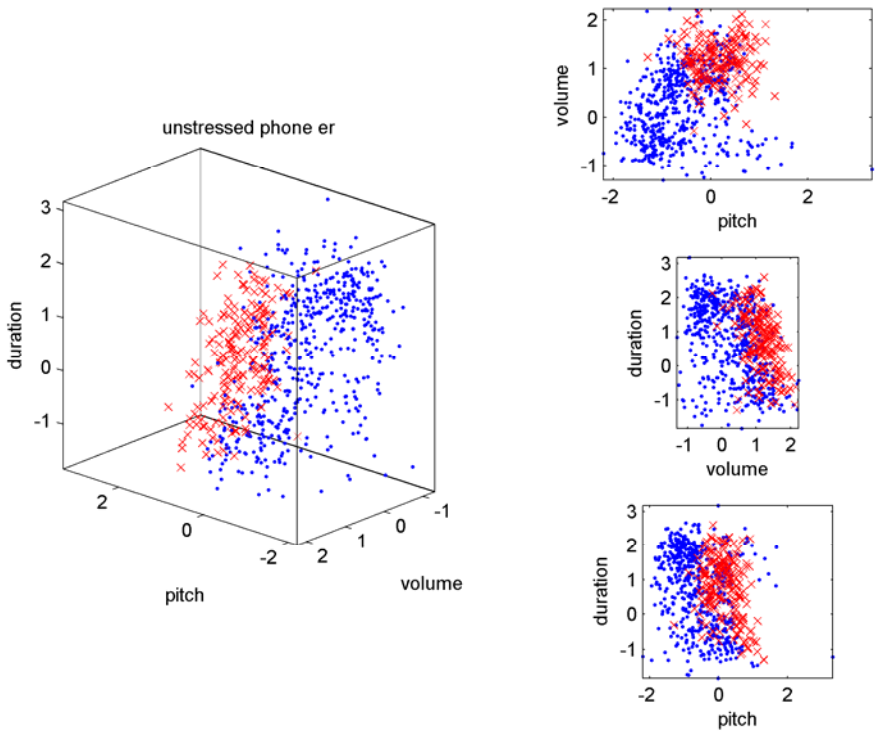

圖 42：er 為非重音的分布圖；x 點代表辨識錯誤，• 點代表辨識正確

接著分析非重音卻辨識成重音的分布情況。<圖 37>為母音 ah 為非重音的資

料，x 點代表辨識正確，•點代表非重音卻辨識錯誤成重音。同樣的左邊的圖為 3 維特徵的圖形，右邊的圖為兩兩特徵為一組的分布圖，由此圖可看出•區塊大部分比 x 區塊小，表示音高、音量及持續時間過大。<圖 41>為母音 ih 為非重音的資料，<圖 42>為母音 er 為非重音的資料，均有相似的分布情況。<圖 38>至<圖 40>分別代表了母音 ah 中音高、音量、時間的分布圖形，主要使用接收者操作特徵曲線(Receiver operating characteristic curve,簡稱 ROC)來觀察資料分布情況，而這三張圖中可以看到錯誤的資料(◯)與正常的資料(□)都為常態分布的圖形，然而正常資料的最高點均小於錯誤資料的最高點。

對於母音音素辨識錯誤的部分總括會錯誤的原因可能是因錄音者在念這些母音時並未發音完全，例如 eh(ɛ)與 ae (æ)如果發音不正確，兩者可能會混淆。例如<圖 43>，landmark 的兩個母音為 ae(æ)與 aa(ɑ)，但錄音者的 ae(æ)念成 eh(ɛ)，音量明顯較 aa(ɑ)小。

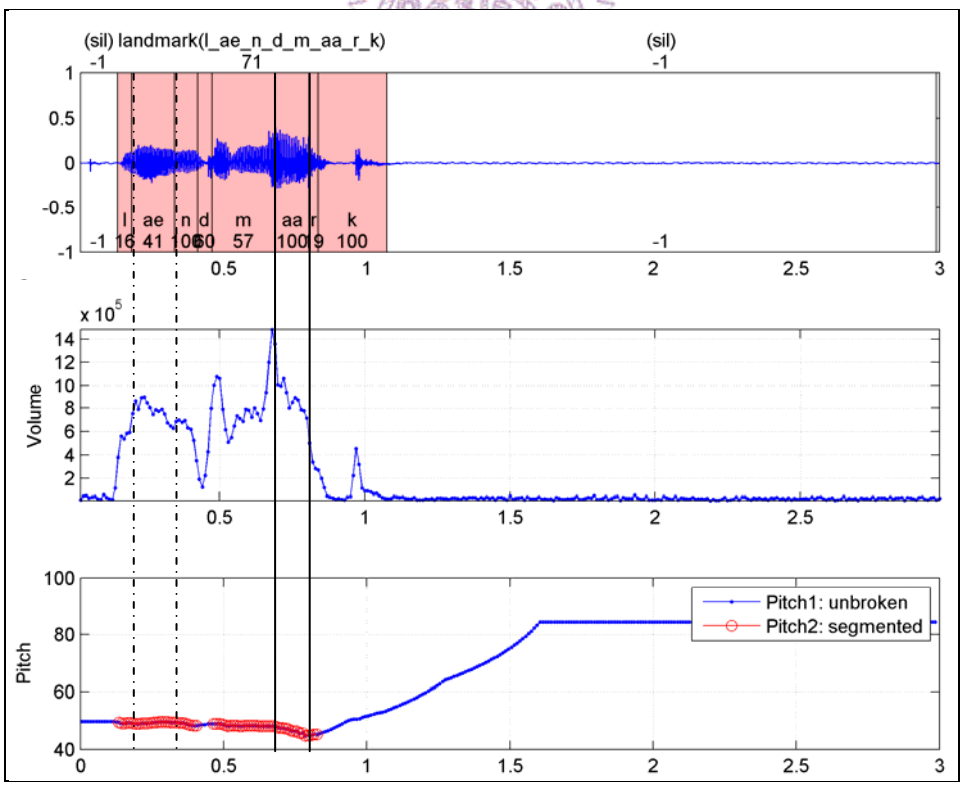

圖 43：母音念不正確之範例

在 Forced Alignment 切音時，少部分資料有音素切錯的情況，切錯的判定為以 2 位觀察者的主觀判斷來決定均錯誤則判定為錯誤，如<圖 44>所示，可以看見黑色方框的部分有一部分靜音被切入子音 ng 裡；<圖 45>中被框起的部分，實際聽到的音有 p_eh_r 整個音節的聲音，但被均列入母音 eh。音素若切錯的話，可能會導致特徵擷取上面會出現錯誤。

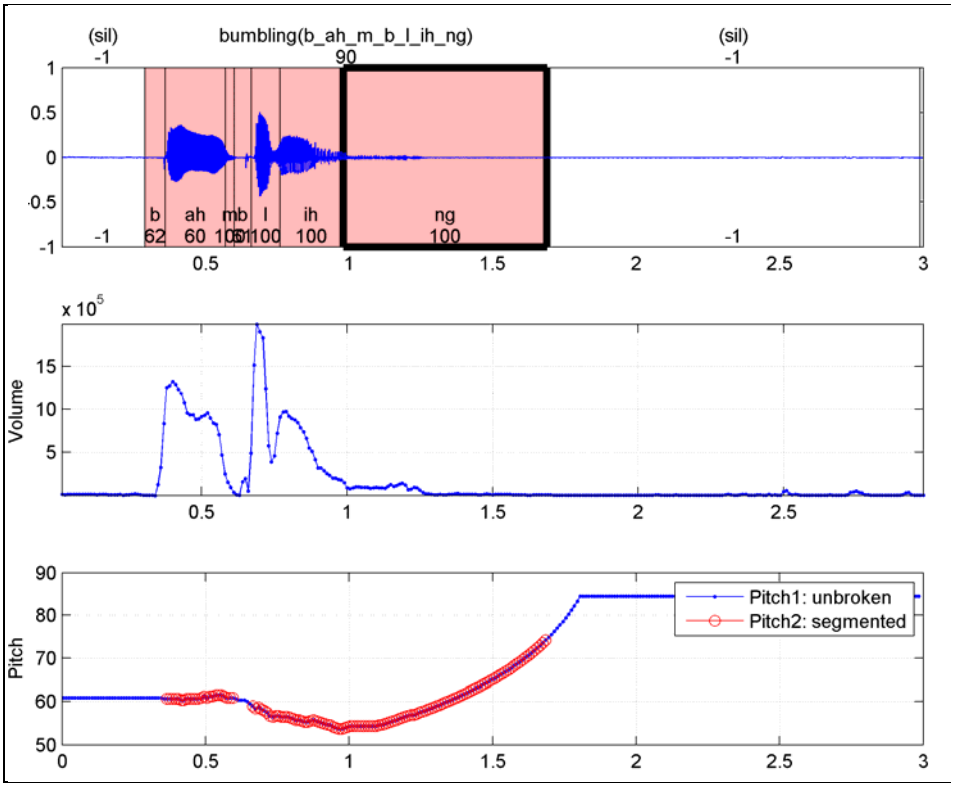

圖 44：切音錯誤範例 1

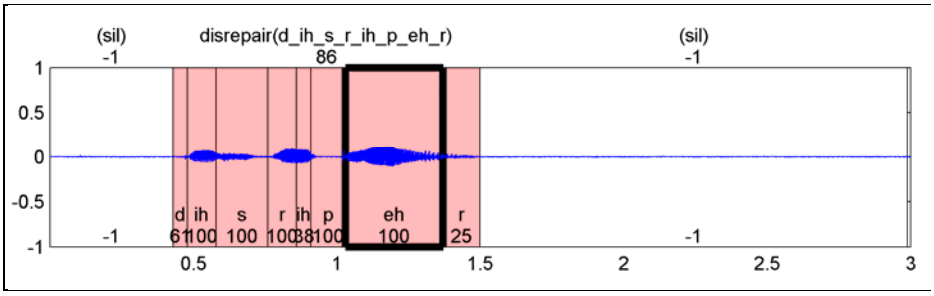

圖 45：切音錯誤範例 2

在使用 Forced Alignment 切音時，使用 Unbroken Pitch Determination Using Dynamic Programming(UPDUDP)擷取音高向量資訊，大部分的錯誤資料發生在擷取音高特徵時出現錯誤，主要的錯誤分為兩個部分，一個是在起始時音高不正常過高，一個則是在尾端不正常變高。如<圖 46>所示，可以看到音高資訊於起始端有一部分不正常的高起，而擷取時若取到該段音高，則音高的擷取就會是不正確的資訊。正常的音高為黑色線條(—)所表示，明顯看出差異很大。

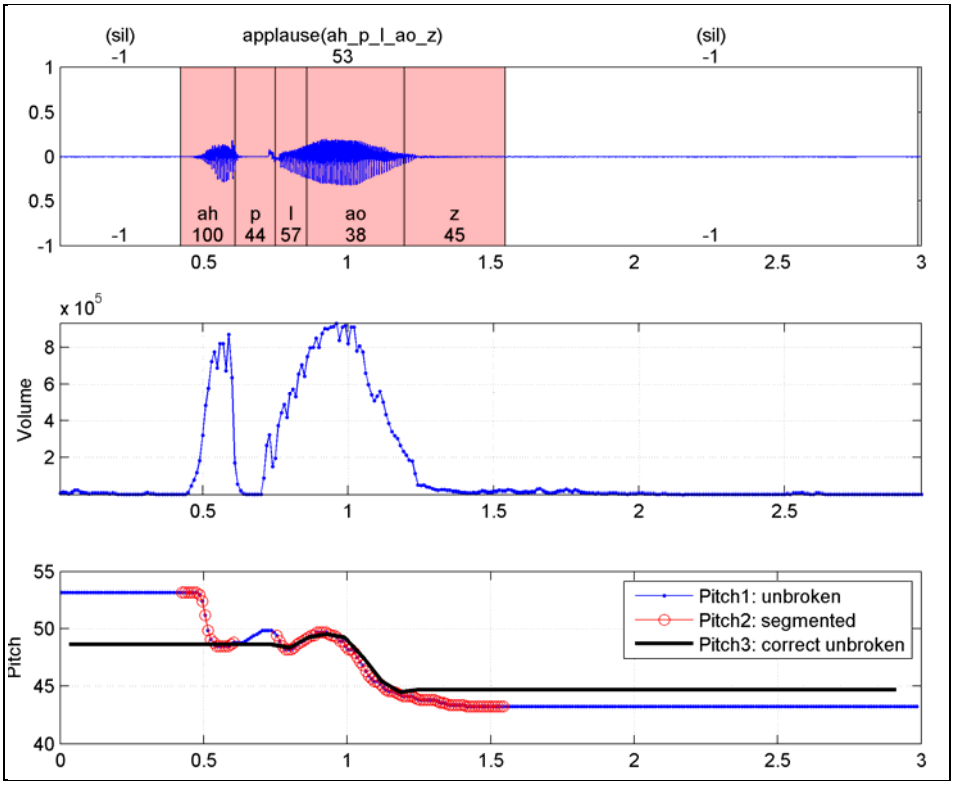

圖 46：音高擷取異常範例 1

<圖 47>則是尾端異常高起的範例，很明顯可以看到在音高的示意圖中，尾端異常增高，而擷取資料時取到一段增高的錯誤資訊。黑色的線條明顯看出在尾端下降，與程式取出的音高向量差異很大。

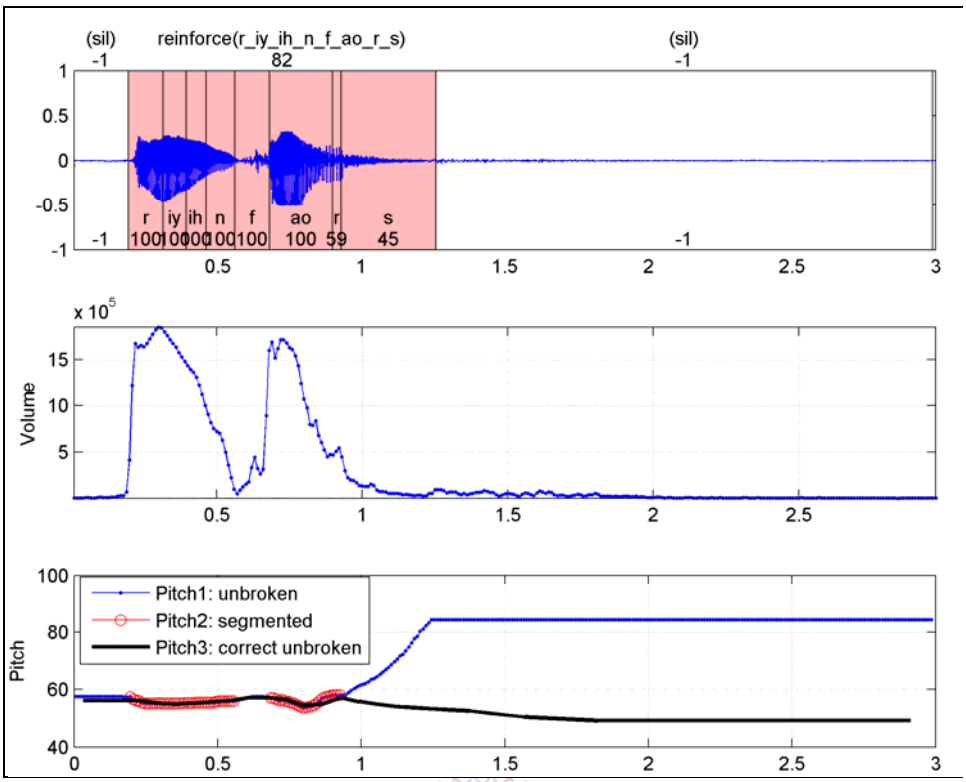

圖 47：音高擷取異常範例 2

而在分析錯誤資料時，發現有 4%的錯誤發生在重音是最後一個音節時，有可能的原因是因最後一個音節聲音會變小聲，使得在取得音高、音量向量的資料後，取平均值或中位數的處理時會受到影響。<圖 48>中，iy 是重音而 ey 是非重音，但可以看到 iy 的音高、音量後半段急速下降，使得取平均值或中位數時，該段資料下降。

而可能的解決方法就是改取最大值，這部分於實驗一作單一特徵參數的辨識方法可見(3.2 節)，其音高及音量各自的辨識率只有 76.31%、56.05%，比取平均值或中位數的方法低 3.5%~4%。實驗二使用多維特徵參數的辨識方法也使用到最大值來辨識，但是效果並不顯著。

另外可以解決的方法則是用第三四分位數取代最大值，這部分可於單一維特徵參數的辨識方法中看出第三四分位數的辨識結果均比最大值的辨識結果佳。

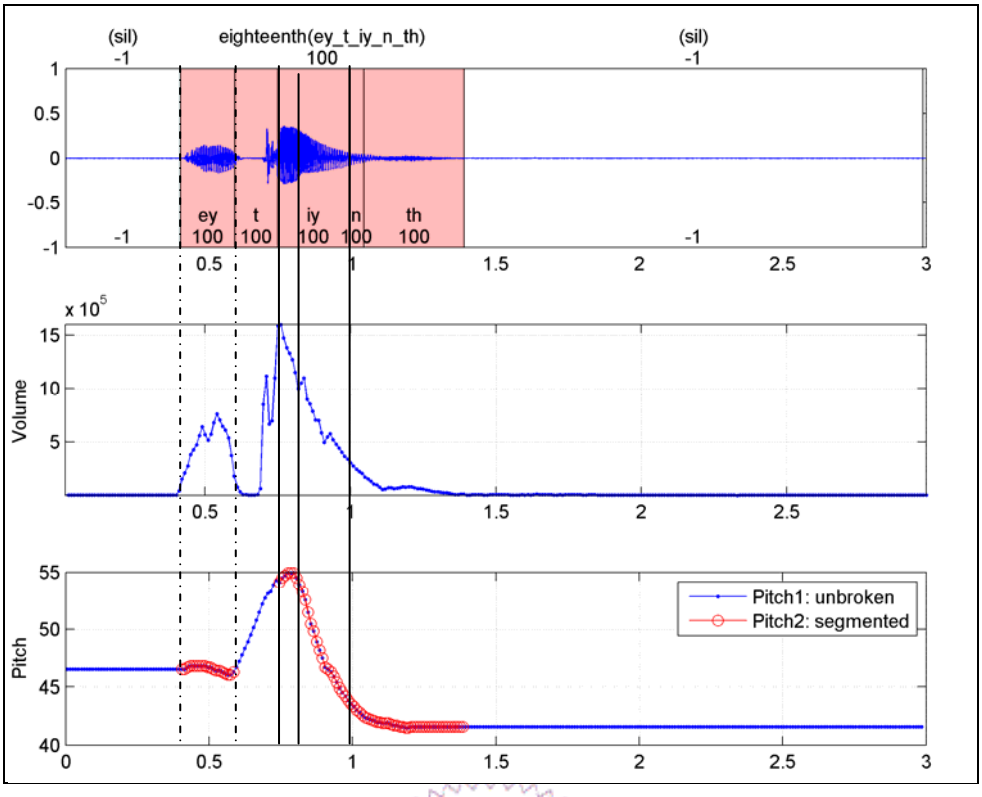

圖 48：重音在最後一個音節

第4章 結論與未來工作

本篇論文主要研究英語詞彙之重音辨識，主要使用強制對位(Forced Alignment)的方法對各詞彙進行切音，產生各音素(phone)，對各音素取出三種特徵參數：音高向量(pitch vector)、音量向量(volume vector)以及持續時間(duration)。使用不同的處理方法，如取中位數、平均值、最大值等等，將各個音高及音量向量轉換為各種數值，而每種處理方法均會產生一種特徵參數。主要的辨識方法分成兩種，8 組一維特徵參數的辨識方法，以及 11 組多維特徵參數的辨識方法。前者的辨識方法是以一個詞彙為單位，找出母音音素下有最大值的特徵參數為重音；而後者的辨識方法分為兩個步驟，第一步驟是對各個母音音素的特徵參數做 GMM 分類，分出重音與非重音兩類；第二步驟則是對 n 個音節詞彙使用第一步驟所產生的 log likelihood 產生 2n 個特徵，分別使用 GMM 與 SVM 分類器進行 n 個音節詞彙的重音辨識。

根據實驗結果可知兩步驟分類的辨識方法中，第二步驟使用 SVM 的辨識效果均比使用 GMM 來的佳，此部分可以再作 GMM 前先使用 PCA 進行資料的主要分量分析再進行 GMM，則可以改善 GMM 的辨識結果，而 2、3、4 個音節的詞彙辨識率分別為 90.36%、86.85%、85.65%，由此可知音節詞彙越多則辨識率會下降，若要得到較佳的辨識結果，各個音節詞彙要使用不同的特徵組合(feature set)。第二步驟中 2 個音節詞彙的辨識結果比一維特徵的辨識方法提高 7.78%的辨識率。

未來對於重音辨識的研究可以朝下列幾個方向做更細部的處理以提升辨識率及實用性：

1. 對完整的語句中辨識各個詞彙的重音，這部分可能會受到語調的影響，與本論文單一個多音節詞彙所使用的辨識方法將可能會有所差異。
2. 可以考慮使用其它的分類器，如：Neural Network(NN)，進行比較。
3. 對於音高擷取所使用的方法進行改良，降低特徵擷取時的錯誤。

4. 探討音高、音量、持續時間對於重音的重要性。

5. 對於進行 GMM 分類前可以先對資料作 PCA 進行主要分量分析。

6. 對於兩步驟分類的辨識方法加入更多的特徵，例如：Q_1，Q_3

參考文獻

[1] D. Wang and S. Narayanan, “An acoustic measure for word prominence in spontaneous speech,” *IEEE Trans. Speech, Audio, Language Process.*, vol. 15, no. 2, pp. 690–701, Feb. 2007.

[2] F. Tamburini and C. Caini, “An automatic system for detecting prosodic prominence in American English continuous speech,” *Int. J. Speech Technol.*, vol. 8, pp. 33–44, 2005.

[3] Jenkin, K.L. and Scordilis M.S., “Development and comparison of three syllable stress classifiers,” *ICSLP ‘96 roceedings*, Philadelphia, USA, pp. 733–736.

[4] Huayang Xie, Peter Andreae, Mengjie Zhang, and Paul Warren, “Detecting stress in spoken English using decision trees and support vector machines,” in *Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation*. 2004, pp. 145–150, Australian Computer Society, Inc.

[5] J. Tepeerman and S. Narayanan, “Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners,” in *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing*, Philadelphia, March 2005.

[6] C. Wang and S. Seneff, “Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the JUPITER domain,” in *Proc. 7th Eur. Conf. Speech Communication and Technology (EUROSPEECH ’01)*, vol. 4, Aalborg, Denmark, September 2001, pp. 2761–2764.

[7] Jyh-Shing Roger Jang, “DCPR (Data Clustering and Pattern Recognition) Toolbox”, available from the link at the author's homepage at "http://www.cs.nthu.edu.tw/~jang".

[8] Zhi-Sheng Chen, Jia-Min Zen, Jyh-Shing Roger Jang, and Liang-Yu Chen, “A Two-stage Classification Framework for Stress Detection in English Word Utterances”, April, 2009

[9] 高斯混合模型 (GMM), http://neural.cs.nthu.edu.tw/jang/books/dcpr/doc/08gmm.pdf

[10] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines,2001. Software available at

[11] 林長青, “支撐向量機應用於科學探索”, 雲科大碩士論文, 2003.