中華民國第 57 屆中小學科學展覽會

作品說明書

高級中等學校組　電腦與資訊學科

第三名

052502

語音辨識之忠狗

學校名稱：臺中市立臺中女子高級中等學校

作者：	指導老師：
高二 賴湘縈	李文靖

關鍵詞：梅爾倒頻譜係數、音高、相關係數

摘要

本作品為一隻電腦上虛擬的「狗」，牠會聰明的聽懂主人的話，陌生人則不然。其主要的技術是利用錄製的語音訊號，擷取其梅爾倒頻譜係數(Mel-Frequency Cepstrum Coefficient , MFCC)、音高(Pitch)以及週期性聲波等三種特徵值來作為語音模型之建立，並以最少誤差來進行語者辨識。本實驗有 20 位語者，錄製語料共 207 個句子作測試。實驗結果顯示，自我語者之正確平均辨識率為 96.62%，而其他語者之平均辨識率為 99.62%，兩者之總平均為 99.47%。整體而言，本作品除了高準確度外亦含有高應用性之實測作品。

壹、 研究動機

隨著今日資訊科技的突飛猛進，人們也對語音信號愈來愈有研究，今天我們已能夠藉著語音和電腦分辨出人類所說的話以及說話者的身分。由於每個人聲音的特性以及說話的習慣都不相同，因此聲音可說是人類最為自然的特徵之一，且具備容易產生、擷取等特性，所以適合用於身分辨識。語者辨識已經廣泛應用在各種領域，例如:安全管制方面，如門禁系統、金融系統的身分辨識 [1]。手機也已可以透過指紋解碼，相信若可用聲音應該會更方便吧!!

貳、研究目的

透過做出的程式，能當場錄製聲音並可利用之前所得出的語者模型分辨出自己和他人聲音的差別，測試者若結果正確，則可使介面中的狗停止持續不停的吠聲，但若結果錯誤則會使介面中的狗吠聲更為頻繁。

參、研究設備及器材

一、操作界面

本作品採用的程式語言為 Visual Base 6.0 [2]。此軟體在開發環境中，介面設計與程式撰寫比其他語言容易，非常適合初學者使用。Visual Basic 6.0 有許多套裝物件可以使用，介面元件的外觀與配置可由工具箱直接取用及物件屬性來設定 ，設計者只須著重於演算法與介面排版。圖 1 及 2 為實測作品及撰寫程式之介面

圖 1. 實測介面

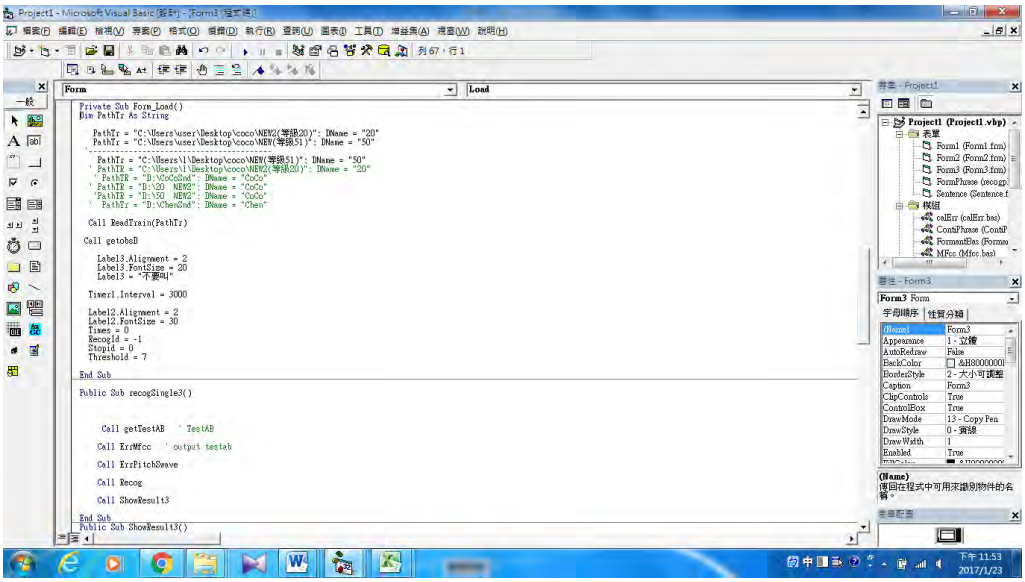


圖 2. 撰寫程式介面

二、語音資料來源

本作品之語音資料來源，是由熱心的同學和實驗室的學長姐們一起幫忙錄製。錄製內容為「不要叫」三個國語單字，共 20 位同學，207 句語料。

肆、研究方法

中文的語音可以分為「子音」、「母音」兩大部分。子音(consonant)是指發音的時候，氣流受到咽頭、口腔與鼻腔等部位的阻塞，導致語音的波形震動凌亂音強小。母音(vowel)是指發音的時候，氣流在口腔內會有共振的現象，導致語音波形具有週期性，和子音相比其音強較大。本作品將取用「母音」的語音資料做語者辨識之用。圖 3 為語音之聲波圖。

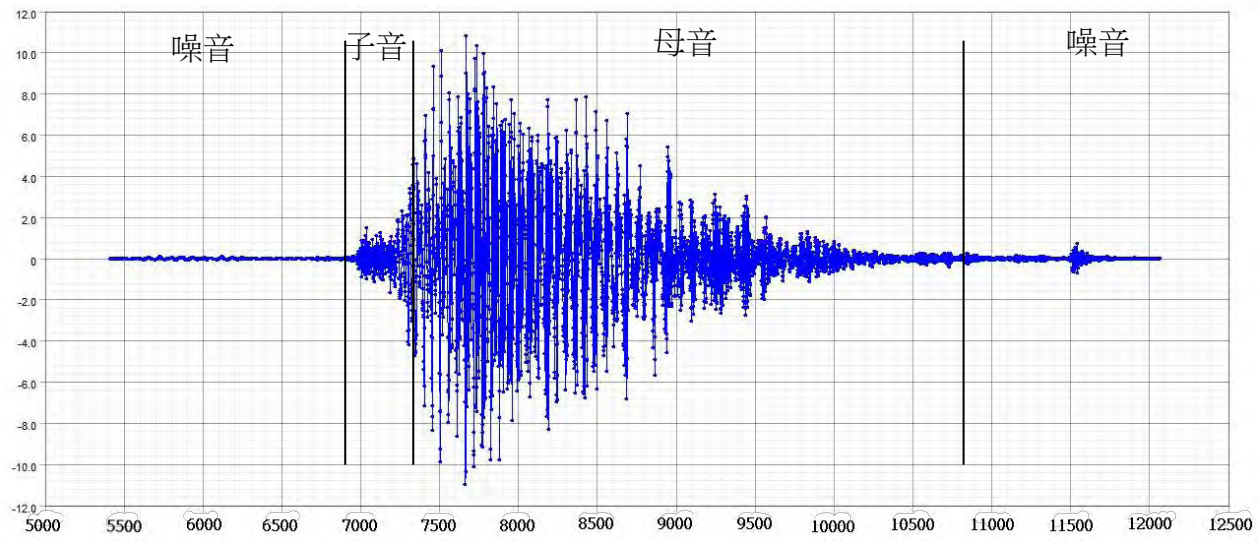


圖 3. 語音「F070 叫」之聲波圖

當語者錄製一段語音，在短短幾秒鐘的語音訊號裡，就包含相當龐大的資料。在整個語音資料中，真正有用的資料只佔其一小部分、長短不一，其餘都是噪音。因此若將所有語音資料進行比對，比對的結果不但準確率低而且耗時。為了節省時間和增加準確性，我們將對語音訊號進行前處理，一般過程為數位化、常態化、端點偵測、切割音框、預強調和視窗化。經過這些處理，將能提高我們的辨識率以及減少計算時間。

除了上述的初步處理，我們將擷取出代表語者的特徵參數(MFCC、Pitch、Corr)，做為將來辨識的資料。本作品中，主要架構分為兩大部分。 第一部分是語者特徵值的取得，包含了梅爾倒頻譜係數(MFCC)、基本週期(Pitch)、週期性聲波。 第二部分為語者模型的建立，並以最小誤差法作為辨識的依據。如圖 4 和 5 所示。

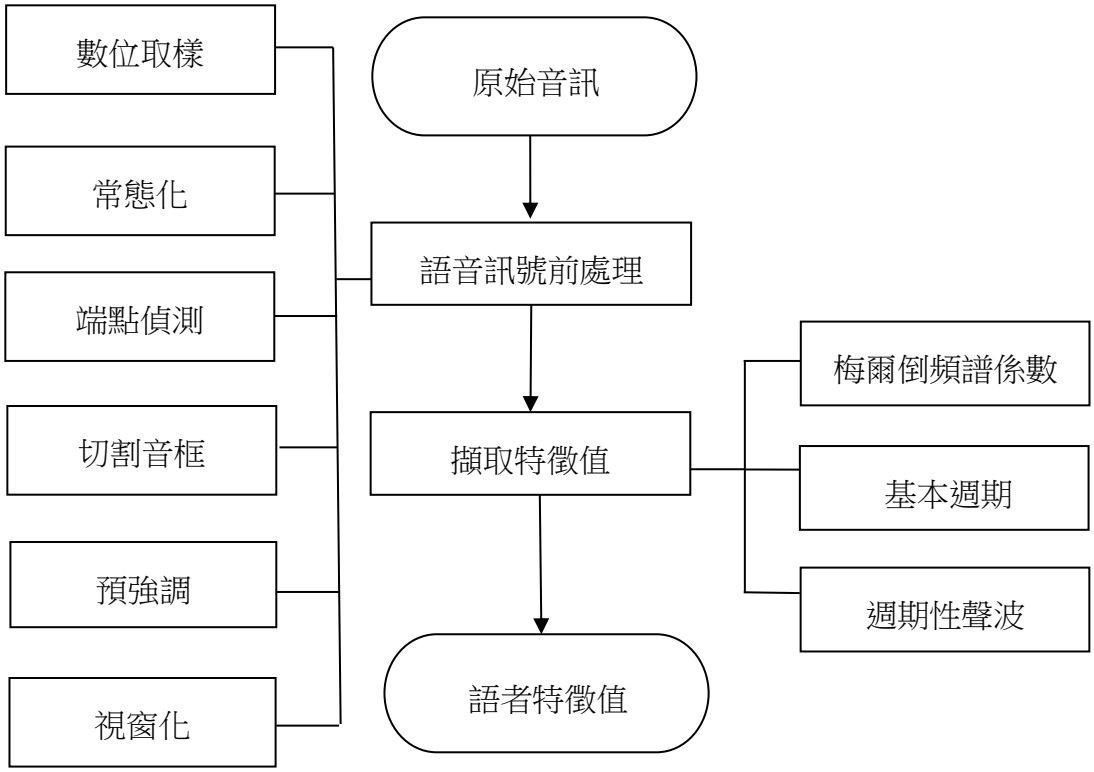


圖 4. 語者特徵值之擷取流程

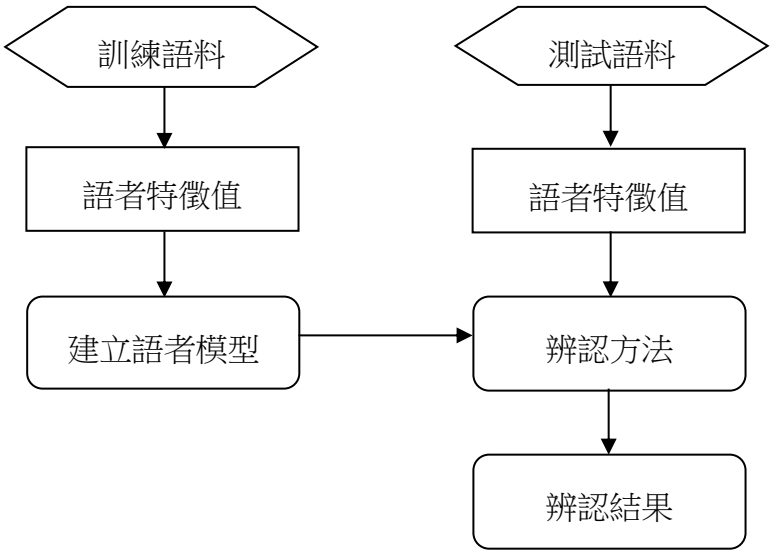


圖 5. 辨識流程

一、 語音訊號的前置處理

(一) 數位化

聲波是一個由空氣傳遞的縱波，此種語音訊號為『類比訊號』。由於電腦只能儲存 0 和 1 的數值訊號，所以我們必須將類比訊號量化成電腦能夠辨識的數位訊號。而這個過程我們稱之為數位化。在語音訊號經過數位化的處理後，會在固定時間重覆對聲音取樣，稱之為『取樣頻率』，而取樣之振幅大小則稱之為『取樣值』。本作品之語音訊號其取樣頻率為 11025Hz，即每秒鐘取樣 11025 個樣本點。

(二) 常態化

每次語者所錄製之語音，可能因說話的音量或者電腦上音量的設定而有所不同，其語音訊號之取樣值將會有很大的差異。為了提高辨識率，我們將取樣值常態化，使所有語音之音量都設在相同範圍之內。本作品之常態化之後的取樣值設定在範圍 [-10 , 10] 之間。

(三) 端點偵測

在錄製語音的過程中，說話者的速度或習性會造成真正語音的位置有所不同 ;有些音會獨立分開，有些音則會連在一起，再加上外部噪音的干擾及前後段不必要的靜音。因此為了達到最佳的辨識度，我們必須要去除雜訊並切割出真正的語音資訊。一般可以使用能量量測法及越零率法或者隱藏式馬可夫模型之強迫對齊法 (HMM forced alignment) 來達到端點偵測的效果。

(四) 切割音框

由於語音訊號包含著相當大的資料量，若對所有訊號直接進行分析，其波形或者頻譜的變動性會過大導致不容易分析。因此為了觀察其細微的頻譜變

化，我們會將訊號切割成很多段並以固定的視窗長度來分別處理，切割後的每一段語音，即稱為音框(frame)，本作品之音框的取樣點為 256 個點組成。

(五) 預強調

一般而言，語音訊號從口中說出，透過空氣傳遞到耳朵，這接收的過程中會有- 6 dB/oct 高頻的衰減 ; 但是人類耳朵會做出彌補的功能自動提高 + 6 dB/oct 的高頻，以利清楚地接收到訊息，為了讓電腦像人類耳朵一樣，我們必須讓語音訊號通過一個高通濾波器，稱之為預強調，以補回高頻的損失。

(六) 視窗化

語音本身是有連續性，但是切割音框會造成訊號的不連續性。因此為了增加音框與音框間的連續性，我們必須將音框內的能量值集中在中間段，減少邊邊的不連續。此過程則稱視窗化。一般我們是用漢明窗(Hamming window)來處理音框。

二、 語者特徵值的擷取

語音訊號在經過前處理後，我們考慮擷取梅爾倒頻譜係數、音高及週期性聲波等三種特徵值作為語者辨識之用。此三種特徵值詳述如下:

(一) 梅爾倒頻譜係數(MFCC) [2]

梅爾倒頻譜係數主要包含三部分流程：1. 離散傅立葉轉換， 2. 三角濾波器和 3. 離散餘弦轉換。

1. 離散傅立葉轉換

由於電腦語音訊號是數位存取的，而且聲音是由不同頻率所組成，離散傅立葉轉換主要目的是將語音訊號從「時間領域」轉換到「頻譜領域」來做處理。此種轉換可增加辨識之準確性極不穩定性，其公式如(1)式所示：

$$X[k] = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}nk} \quad , 0 \le k \le N-1 \qquad (1)$$

其中X[k]為頻域上的樣本，x(n)為時域視窗化後的語音信號， N為音框取樣點數。因為離散傅立葉轉換需耗費很多計算的時間，目前都改用快速傅立葉轉換(Fast Fourier Transform)，它加快了計算的速度，只不過是音框取樣點必須為 2 的次方倍數。

2. 三角濾波器

接下來將語音信號在頻譜領域的能量通過一組濾波器組，這濾波器組是由 M 個三角形濾波器所組成，其中每個三角濾波器的高度在[0 , 1]之間。

3. 離散餘弦轉換

對全部 M 個濾波器所輸出的對數能量，S[m], m=1,2,……,M, 利用離散餘弦轉換就可以得到梅爾頻率倒頻譜係數，其公式如(2)式所示：

$$c[n] = \sum_{m=1}^{M} S[m] \cos\left(\frac{\pi n\left(m-\frac{1}{2}\right)}{M}\right), \quad 0 \le n \le M-1 \qquad (2)$$

其中$c[n]$, n=1,2,……,M, 就是最後求得的梅爾頻率倒頻譜參數，M 為三角濾波器的個數。本作品 M=40 根據以上特徵求取的流程，圖 6 為錄製語音訊號『F702 叫』之一段母音波形圖，經過離散傅立葉轉換後的頻譜能量圖，當頻譜能量經過上述之三角濾波器和離散餘弦轉換後，我們可以得到一組梅爾倒頻譜係數，圖 7 為其示意圖。

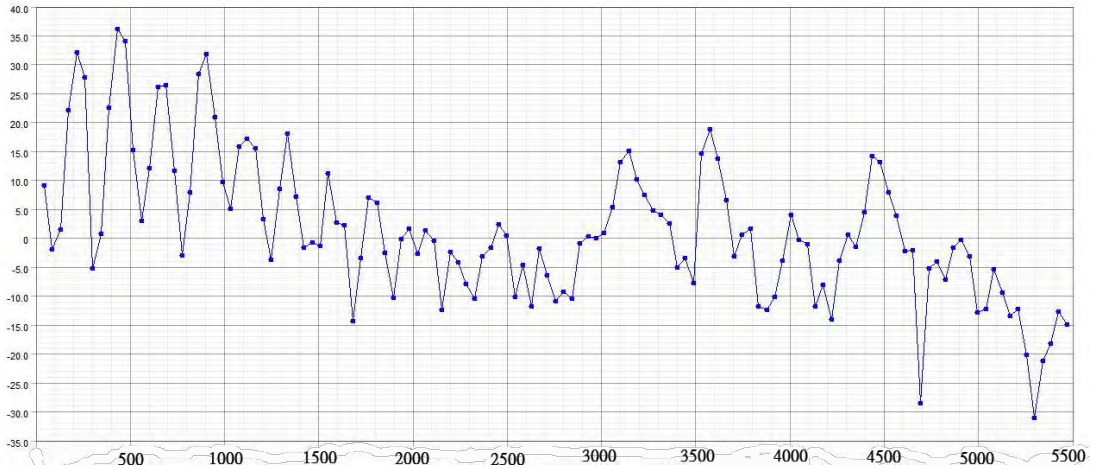


圖 6. 語音訊號「F702 叫」之頻譜圖

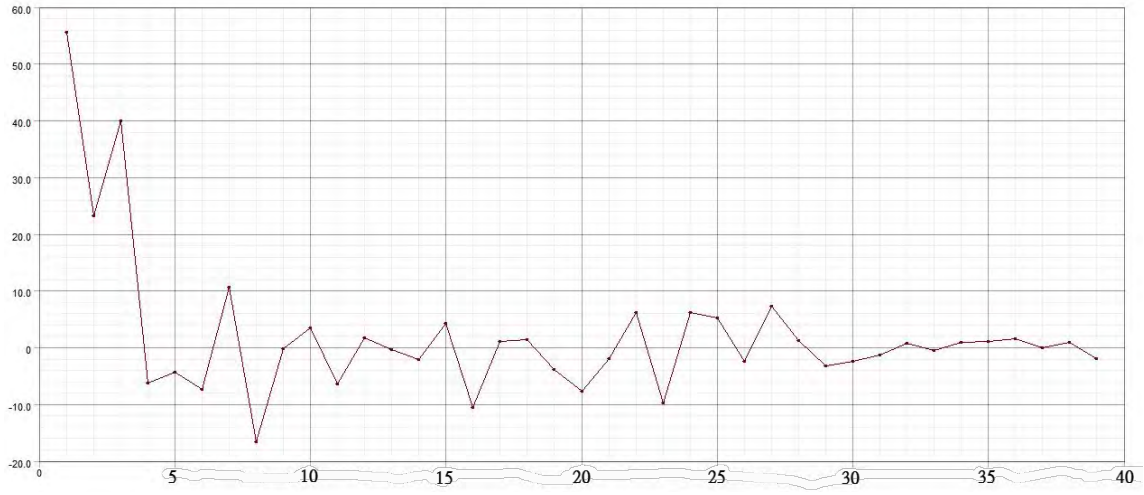


圖 7. 語音訊號「F702 叫」之梅爾倒頻譜係數

(二)音高(Pitch) [4]

男女生音高不同。一般而言，男生之音高比女生低；亦即，說話時，男生喉嚨震動的次數比女生低。因此，音高可以作為男女語者分辨的特徵。由於音高是一種頻率，所以其倒數為週期，我們稱為基本週期。

聲音之基本週期可由自相關函數得知。其方法是由 Rabiner (1997) [5]所提出，主要是基於時域的演算法作為音高的研究。自相關函數常用於時域上波形的分析，主要是用來計算一個音框在不同平移量的相關程度。其公式如(3)式所示：

$$ACF(\tau) = \sum_{n=0}^{N-1-\tau} S(n)S(n+\tau) \quad (3)$$

其中 S(n)是語音訊號，τ 是平移量和 N 為音框大小。式子(3)表示語音訊號平移 τ 時，語音訊號 S(n +τ)和原始語音訊號 S(n)之內積值。對平移量 τ ，我們將找尋最佳值使得 ACF 序列最大。此最大之 τ 值就可以當作該音框之基本週期，圖 8 表示基本週期為 50 點之聲波示意圖。

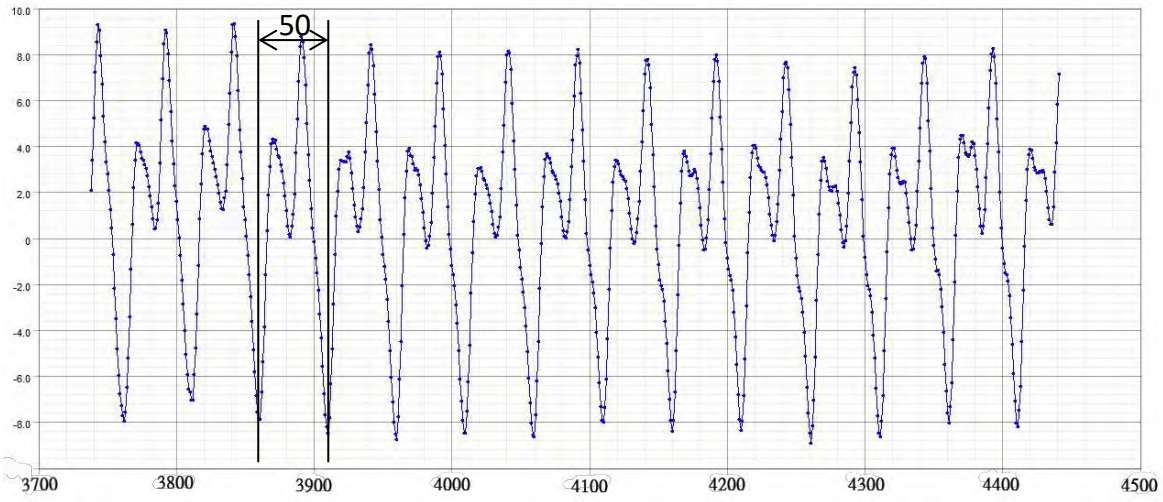


圖 8. 聲波之基本週期

(三)週期性聲波

中文母音之語音聲波具有週期性，可以被用來做語音辨識之特徵。當擷取一段週期性聲波，我們將可進行相關係數之比對，其計算公式如(4)式所示

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

(4)

這裡 $X=(X_1, X_2, X_3, \ldots\ldots, X_n)$ (語者的音)和 $Y=(Y_1, Y_2, Y_3, \ldots\ldots, Y_n)$(其他人的音)代表兩段週期性之聲波，$\bar{x}$ 和 $\bar{y}$ 為其平均值。本作品之擷取聲波點數 n 設為 80。圖 9 為一段週期性聲波之擷取圖(80 點)。

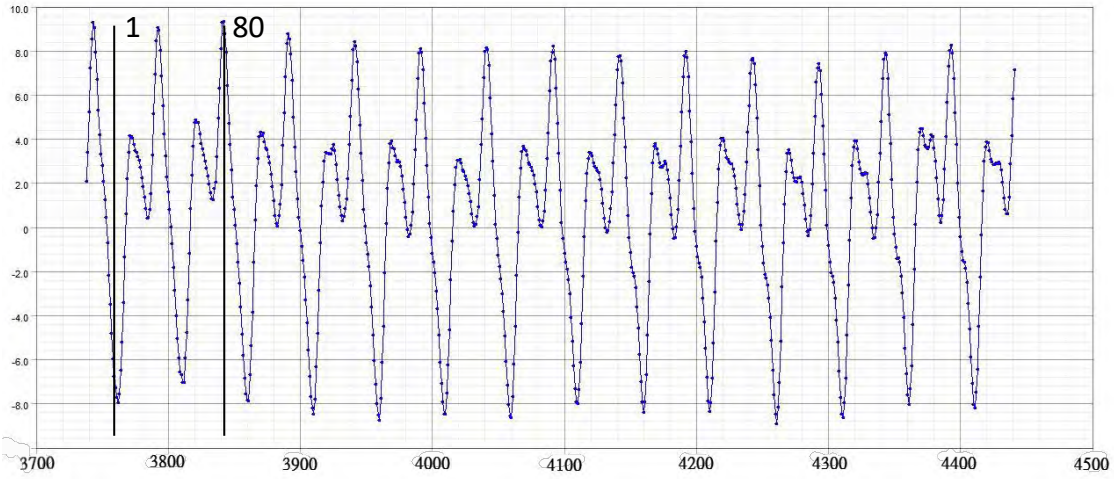


圖 9. 週期性聲波之擷取圖(80 點)

三、 辨認方法

當所有語料經前述(一) ~ (三)步驟之特徵擷取，我們必須要建立模型並對測試音進行比對。由於梅爾倒頻譜係數(MFCC)、音高(Pitch)及週期性聲波(Corr)在計算誤差時有不同的度量值，所以我們必須要轉換至同一度量上來做比對。本作品考慮之度量值為三者轉換後之總和，每個音其 Score 值之計算如(5)式所示。

$$\text{Score} = \text{MFCC}(X \mid \text{model}) + \text{Pitch}(Y \mid \text{model}) + \text{Corr}(Z \mid \text{model}) \quad (5)$$

其中三者公式定義如下:

令 $dx = \min_{a \in model} \|x - a\|$ 表示 x 和模型間最小之梅爾倒頻譜係數誤差，

$dy = \min_{p \in model} |y - p|$ 表示 y 和模型間最小之基本週期差，

$dz = \min_{s \in model} corr(z, s)$ 表示聲波 z 和模型間最大之相關係數值。

此三者轉換函數為

$$\text{MFCC(X|model)} = \begin{cases} 1 & , \ if\ dx \le 40 \\ \frac{45-dx}{5} & , \ if\ 40 \le dx \le 45 \\ 0 & , \ if\ 其他值 \end{cases} , \quad (6)$$

$$\text{Pitch(Y|model)} = \begin{cases} 1 & , \ if\ y \le 5 \\ \frac{10-dy}{5} & , \ if\ 5 < dy \le 10 \\ 0 & , \ if\ 剩餘的值 \end{cases} , \quad (7)$$

和 $$\text{Corr(Z|model)} = \begin{cases} 2 & , \ if\ 0.9 \le dz \le 1 \\ 1 & , \ if\ 0.8 \le dz \le 0.9 \\ 0 & , \ if\ 剩餘的值 \end{cases} 。 \quad (8)$$

由於我們是進行語者辨識，所以週期性聲波之相關係數將給予較高之權重，介於[0,2]之間，其餘兩個介於[0,1]之間。本作品之語音訊號為「不要叫」三個音，因此整個值為三個單音之 Score 值總和，為了方便，我們將以 Score 代表三個單音之總和。

對於每一個測試音，我們要判斷是否為自我語者或其他語者之語音，我們將給予一門檻值 α，若 Score 值大於 α 值，則為語者之音。由於所有測試語料有兩種錯誤，一種為自我語者本身的錯誤，另一種則為其他語者本身的錯誤，因此門檻值 α 之設定將

使兩者之平均錯誤率最小，如公式(9)所示:

$$\alpha = \arg\min_{\alpha}\left[\left(\text{語者錯誤個數} * 3 \middle| \alpha\right) + \left(\text{非語者錯誤個數} \middle| \alpha\right)\right] \quad (9)$$

由本實驗之測試資料庫可得出α值介於 5 至 7 之間。

伍、研究結果與討論

本實驗有 20 個語者共錄製 207 句子「不要叫」。表 1 和 2 為以作者為測試自我語者及其他語者之部分特徵值如梅爾倒頻譜係數(MFCC)、音高(Pitch)和相關係數(Corr)。由表 1 可知，自我語者之 Mfcc 誤差值大都介於[30 , 40]中 ；Pitch 之誤差則介於[0 , 2]中 ；Corr 則介於[0.8 , 0.9]左右。由式子(5)之計算，自我語者之 Score 值將介於 7 到 11 之間 ； 其中除了一個 7 以外，其餘 Score 值皆大於 8，甚至很多都大於 10。表 2 顯示，其他語者之 Mfcc 和 Pitch 之誤差值很大，而且，Corr 之值也很少，所以其他語者之 Score 值介於 0 到 5 之間，且皆低於 7，大都小於 5，因此，當門檻值為 7 時，自我語者及其他語者兩者辨識率為 100%。 圖 9 為所有語料(207 句子)之 Score 值，很明顯地我們可以發現其兩者間的差異，當門檻值為 7 時，我們能清楚的區別出自我語者和其他語者之 Score 值，自我語者之 Score 值皆大於 7，其他語者則皆低於 6。

表 1. 自我語者之特徵值誤差及 Score 值

	MFCC 誤差			Pitch 誤差			Corr 相關係數			Score 值
	1	2	3	1	2	3	1	2	3	Score
1	33.51	33.4	34.05	0	0	0	0.92	0.96	0.84	11
2	32.12	33.71	36.46	1	0	0	0.82	0.93	0.87	10
3	32.61	47.47	34.05	0	0	52	0.94	0.88	0.57	7
4	29.38	39.83	35.94	0	2	0	0.88	0.85	0.94	10
5	42.43	34.6	34.65	0	1	1	0.88	0.96	0.61	8.51
6	41.16	41.16	32.56	0	0	1	0.94	0.89	0.91	10.5
7	39.89	36.85	33.91	0	0	0	0.94	0.89	0.91	11
8	38.89	27.87	29.71	0	0	0	0.94	0.89	0.9	11
9	35.79	36.24	28.87	0	0	0	0.8	0.9	0.87	10
10	33.61	28.83	34.62	1	0	0	0.95	0.9	0.86	11
11	33.49	33.99	29.86	1	0	0	0.92	0.9	0.89	11
12	29.77	30.09	30.17	0	0	2	0.91	0.88	0.42	9

表 2. 其他語者之部分特徵值誤差及 Score 值

	MFCC 誤差			Pitch 誤差			Corr 相關係數			Score 值
	1	2	3	1	2	3	1	2	3	Score
13	45.5	60.33	55.34	11	52	7	0.23	0.37	0.48	0.5
14	46.96	55.83	55.43	10	33	9	0.15	0.22	0.52	1
15	47.03	55.33	5.1	7	45	25	0.45	0.13	0.38	1.5
16	54.06	59.44	56.33	9	24	6	0.11	0.24	0.46	0
17	46.55	56.96	53.52	30	27	21	0.36	0.38	0.2	2
18	46.56	58.87	54.63	11	6	4	0.27	0.09	0.58	2
19	56.56	57.5	55.32	6	35	0	0.67	0.16	0.36	0
20	55.57	57.77	54.29	13	30	11	0.19	0.3	0.45	1
21	48.52	58.51	54.61	14	42	6	0.25	0.24	0.48	0.5
22	47.72	58.74	53.83	9	16	22	0.38	0.35	0.23	3
23	59.99	52.41	57.42	1	0	0	0.62	0.39	0.5	4
24	50.31	58.9	58.65	0	0	0	0.81	0.43	0.44	3.5

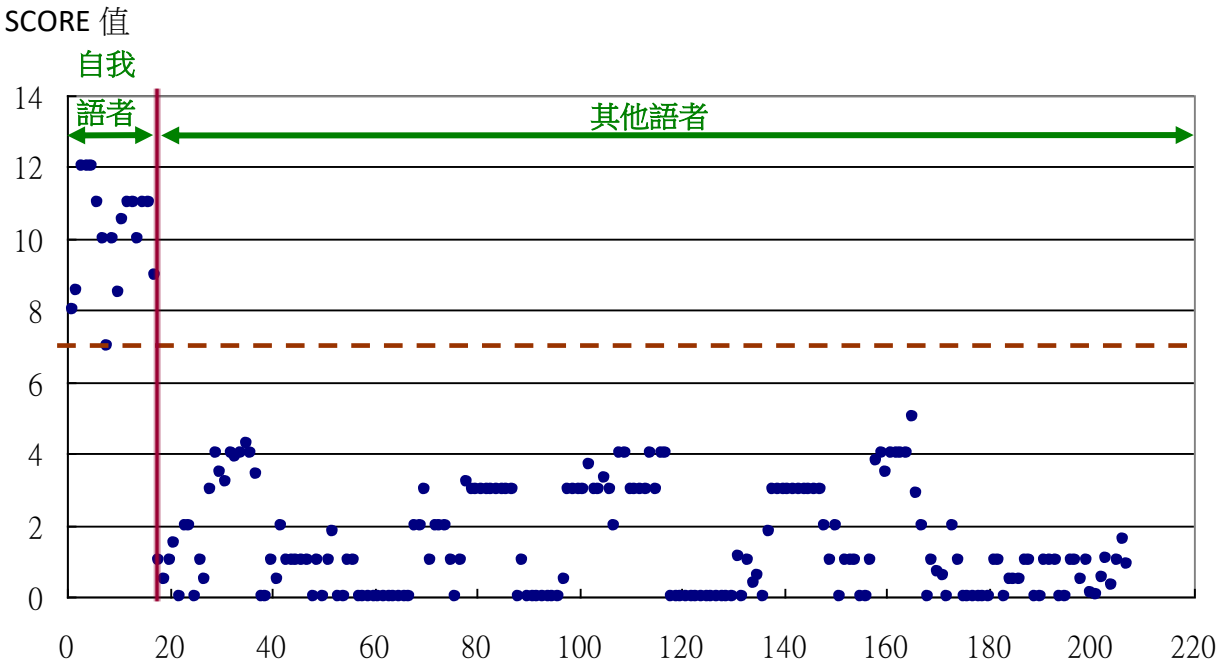


圖 9. 所有語料「不要叫」207 句子之 Score 值

表 3 為每個人的資料庫與自己和其他人的資料庫做比對的結果。表 3 顯示自我資料比對的結果，其辨識率平均為 96.62%，而與他人資料集比對的結果，其辨識率則為 96.62%。整體辨識率為 99.47%，顯示出本作品有非常高的正確率。

表 3. 資料庫之語者辨識率(個數)

語者	自我語者辨識率 (個數)	其他語者辨識率 (個數)	最佳門檻值
A	0.9 (9)	0.97 (192)	5.5
B	1 (10)	1 (197)	6
C	0.9 (9)	0.99 (196)	6
D	1 (10)	0.99 (196)	7
E	1 (10)	0.99 (196)	6
F	0.9 (9)	1 (197)	7
G	1 (10)	1 (197)	7
H	0.9 (9)	0.99 (196)	6
I	1 (10)	1 (197)	6
J	1 (10)	1 (197)	7
K	1 (10)	1 (197)	6
L	1 (10)	1 (197)	6
M	1 (10)	0.99 (196)	5.5
N	0.9 (9)	0.99 (196)	5.5
O	1 (10)	1 (197)	7
P	1 (10)	1 (197)	7
Q	0.8 (8)	0.99 (196)	6.5
R	1 (10)	0.99 (196)	6
S	1 (17)	1 (190)	7
T	1 (10)	1 (197)	7
平均值	0.9662 (200)	0.9962 (3918)	0.9947

有關特徵項目權重之選取，本作品也做了簡易之探討。首先設 Mfcc 特徵項目之權重為 1，Pitch 及週期性聲波兩項目之權重設定為[0.5，2](每次增加 0.25)，門檻值之範圍則擴大為[4，12](每次增加 0.5)。實驗結果顯示各種權重值之辨識率介於[98.91%，99.61%]如表 4 所示。

本作品採用權重(1,1,1)，辨識率為 99.47%，其主要原因為週期性聲波之特徵比 Pitch 和 Mfcc 之特徵較難模仿。

表 4. Mfcc、Pitch 和週期性聲波在各種權重下之辨識率(Mfcc 權重為 1)

		週期性聲波權重						
		0.5	0.75	1	1.25	1.5	1.75	2
Pitch 權重	0.5	0.9966	0.9964	0.9954	0.9947	0.9925	0.9903	0.9889
	0.75	0.9966	0.9959	0.9952	0.9947	0.9918	0.9913	0.9894
	1	0.9952	0.9957	0.9947	0.9944	0.9923	0.9896	0.9894
	1.25	0.9952	0.9952	0.9949	0.9944	0.9915	0.9908	0.9899
	1.5	0.9957	0.9954	0.9949	0.9944	0.9923	0.9896	0.9889
	1.75	0.9957	0.9949	0.9947	0.9942	0.9915	0.9911	0.9899
	2	0.9947	0.9949	0.9944	0.9940	0.9920	0.9899	0.9879

陸、結論及未來展望

本作品中，我們使用了前置處理得出了真正所需的語音資訊，並以梅爾倒頻譜係數、基本週期、週期性聲波三種辨認方式所得的數據，透過轉換函數以及權重比例去加成為一 Score 值，並以最佳門檻值作為語者辨識。本方法簡易、有趣又有效。整體辨識率為 99.47%，準確率非常高。在實用性上，本作品的優點為:

1. 語者可自我設定語句，如「不要叫」、「隨便」……等，由於中文字組合有太多型態，因此其他語者要猜對的可能性很低。
2. 男女生的音高不同，若要偽裝成相似之音高，以生物特徵而言相當困難。
3. 講同樣的句子，每個人在聲波型態上的變化也會不同，因此要有 8 成以上相似性更是難上加難。
4. 門檻值可根據語者做彈性調整。當門檻值愈高，系統之安全性愈高，因為其他語者要有較高的 Score 值 ; 反之，則安全性降低。

在系統穩定性上，建議:

1. 增加實驗取樣人數。
2. 思考外來狀況(如感冒、情緒等)對辨識可能的影響。

3. 可多試幾個權重比以提高準確度

整體而言，本作品具實用性可應用於門禁系統或金融系統等相關領域。實測作品及影片示範亦可參閱。

柒、參考資料

[1] http://ir.lib.nchu.edu.tw/bitstream/11455/18757/1/nchu-102-7100018006-1.pdf。

[2] 王國榮 (2000)，「Visual Basic 6.0 實戰講座 」，全華。

[3] 王小川 (2012)，「語音訊號處理 修訂三版」，全華。

[4] 吳俊甫 (2012)，“時域上基頻軌跡演算法的改良與探討”，國立中興大學應用數學系碩士學位論文。

[5] L.R. Robiner. “On the Use of Autocorrelation Analysis for Pitch Detection,” IEEE Trans. Acoustics, Speech, and Signal processing, vol. ASSP-25, no. 1, pp. 24-33, Feb. 1997.

【評語】052502

本作品主要針對如何辨識出不同使用者所說「不要叫」。

主題清楚，且完整性高。以 Mel-Frequency cepstrum frequency and HMM 為基礎進行實作。

作品目前能分辨出一名語者，若能分辨出多名使用者將更佳。

作品海報

摘要

本作品為一隻電腦上虛擬的「狗」，牠會聰明的聽懂主人的話，陌生人則不然。其主要的技術是利用錄製的語音訊號，擷取其梅爾倒頻譜係數(Mel-Frequency Cepstrum Coefficient, MFCC)、音高(Pitch)以及週期性聲波等三種特徵值來作為語音模型之建立，並以最少誤差來進行語者辨識。本實驗有21位語者，錄製語料共217個句子作測試。實驗結果顯示，自我語者之正確平均辨識率為96.61%，而其他語者之平均辨識率為99.61%，兩者之總平均為99.46%。整體而言，本作品除了高準確度外亦含有高應用性之實測作品。

壹、研究動機

隨著今日資訊科技的突飛猛進，人們也對語音信號愈來愈有研究，今天我們已能夠藉著語音和電腦分辨出人類所說的話以及說話者的身分。由於每個人聲音的特性以及說話的習慣都不相同，因此聲音可說是人類最為自然的特徵之一，且具備容易產生、擷取等特性，所以適合用於身分辨識。語者辨識已經廣泛應用在各種領域，例如:安全管制方面，如門禁系統、金融系統的身分辨識[1]。手機也已可以透過指紋解碼，相信若可用聲音應該會更方便吧!!

貳、研究目的與介面

本作品透過程式語言Visual Base 6.0[2]之撰寫，能當場錄製聲音並利用本作品提供之語者模型來分辨出自己或他人聲音。測試者若結果正確，則可使介面中的狗停止持續不停的吠聲，但若結果錯誤則會使介面中的狗吠聲更為頻繁。實測作品及撰寫程式之介面如圖1和2所示。

圖1. 實測介面

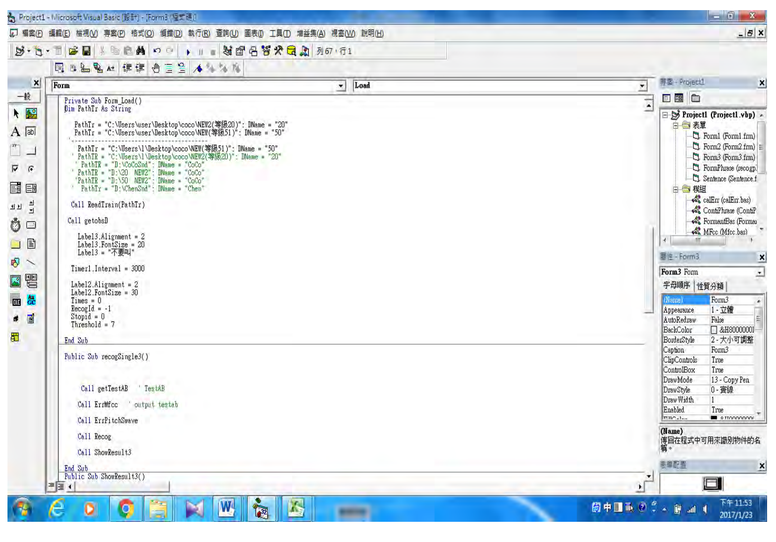
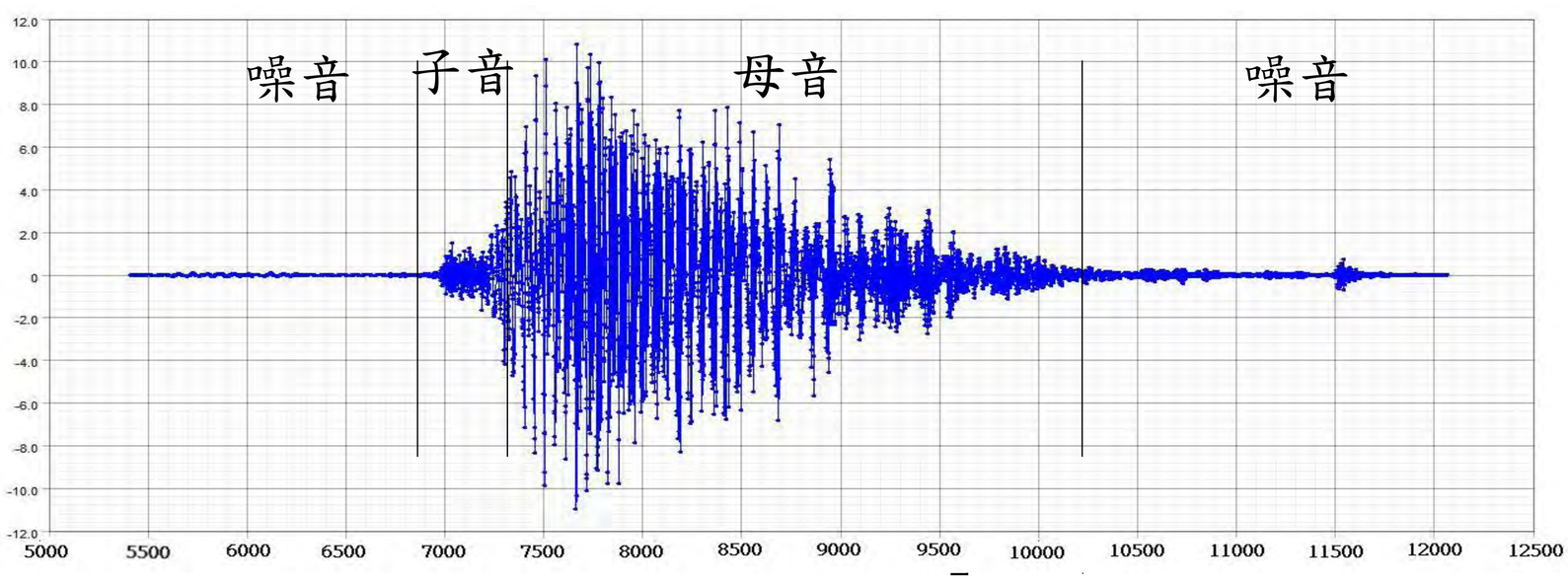


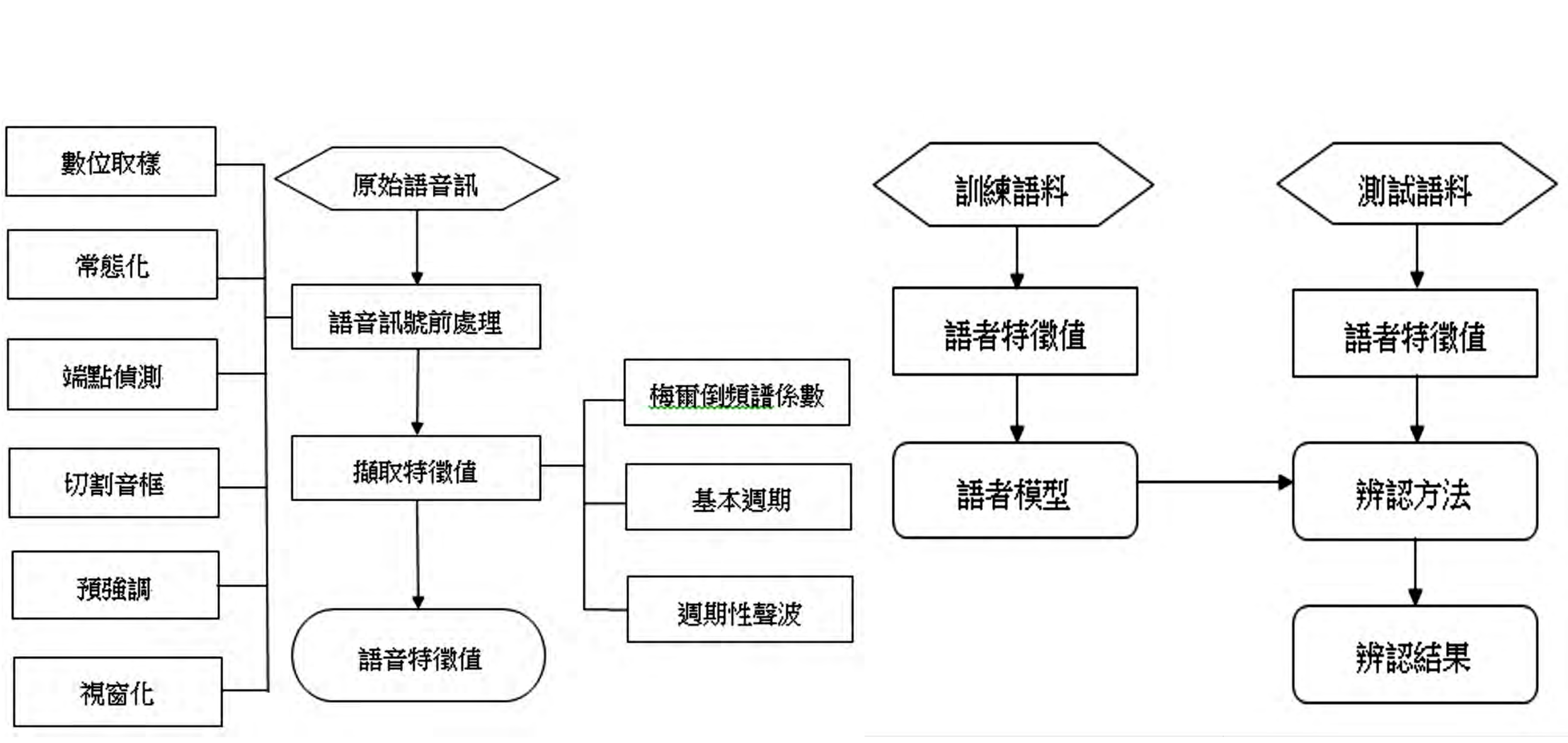
圖2. 撰寫程式介面

參、研究方法

中文的語音可以分為「子音」、「母音」兩大部分。本作品將取用「母音」的語音資料做語者辨識之用。圖3為語音之聲波圖。我們將對語音的特性，進行前處理，如數位化、常態化、端點偵測、切割音框、視窗化。除了上述的初步處理，我們將擷取出代表語者的特徵參數，做為將來辨識的資料。



本作品中，主要架構分為兩大部分。 第一部分是語者特徵值的取得，包含了梅爾倒頻譜係數(MFCC)、基本週期、週期性聲波。 第二部分為語者模型的建立，並以最小誤差法作為辨識的依據。如圖4和5所示。



一、語者特徵值

語音訊號在經過前處理後，我們考慮擷取梅爾倒頻譜係數、音高及週期性聲波等三種特徵值作為語者辨識之用。此三種特徵值簡述如下:

(一)梅爾倒頻譜係數(MFCC)[3]

梅爾倒頻譜係數主要包含三部分流程：1.離散傅立葉轉換,2.三角濾波器,3.離散餘弦轉換。首先我們將語音信號從時間領域轉換成頻譜領域，並通過M個三角形濾波器所輸出的對數能量，然後利用離散餘弦轉換就可以得到梅爾頻率倒頻譜係數。

(二)音高(Pitch)[4]

聲音之基本週期可由自相關函數得知。其方法是由 Rabiner[5]所提出，主要是基於時域的演算法作為音高的研究。自相關函數常用於時域上波形的分析，主要是用來計算一個音框在不同平移量的相關程度。所用到的公式如下:

$$ACF(\tau) = \sum_{n=0}^{N-1-\tau} S(n)S(n+\tau)$$

其中S(n)是語音訊號，τ是平移量和N為音框大小。當語音訊號平移τ時，此式可求出語音訊號S(n +τ)和原始語音訊號S(n)之內積值。對平移量τ，我們將找尋最佳值使得ACF序列最大。此最大之τ值就可以當作該音框之 基本週期，圖6表示基本週期為50點之聲波示意圖。

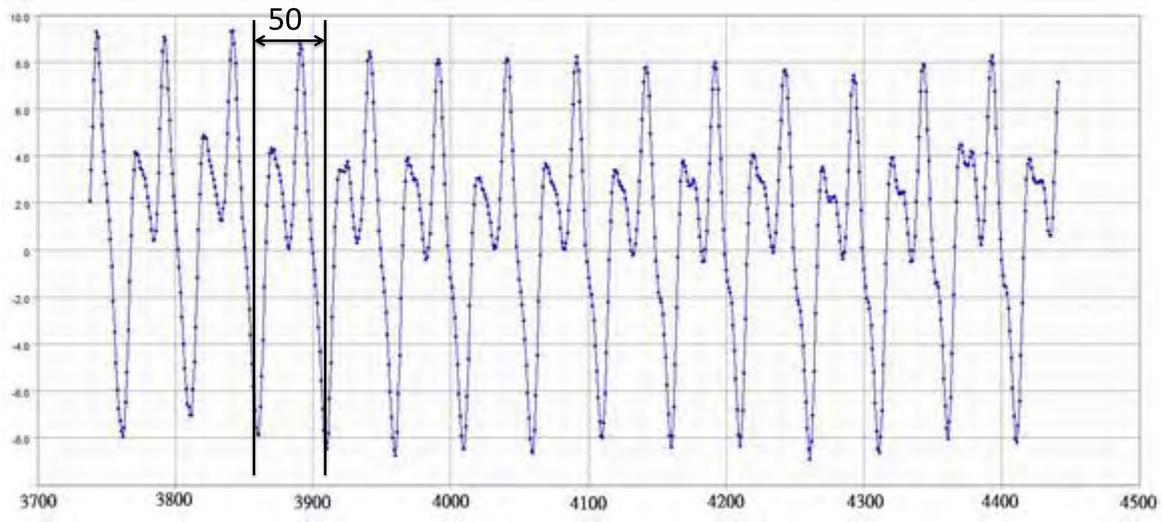


圖6. 聲波之基本週期

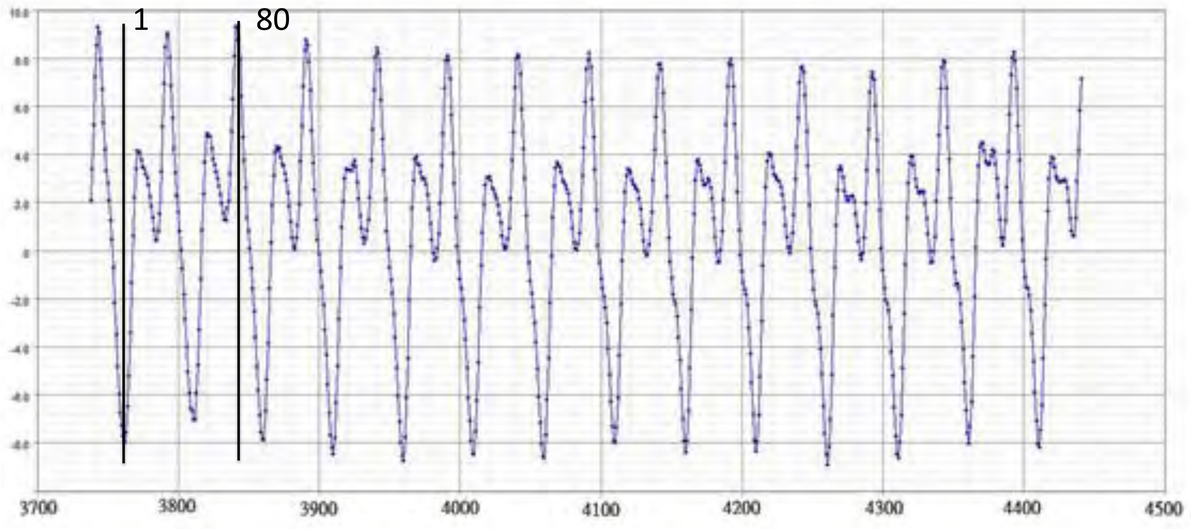


圖7. 週期性聲波之擷取圖(80點)

(三)週期性聲波

中文母音之語音聲波具有週期性，可以被用來做語音辨識之特徵。當擷取一段週期性聲波，我們將可進行相關係數之比對，其計算公式如下所示:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

這裡$x=(x_1, x_2, x_3, \ldots\ldots, x_n)$ (語者的音)和$y=(y_1, y_2, y_3, \ldots\ldots, y_n)$(其他人的音)代表兩段週期性之聲波，$\bar{x}$ 和 $\bar{y}$ 為其平均值。本作品之擷取聲波點數n設為80。圖7為一段週期性聲波之擷取圖(80點)。

二、辨認方法

由於梅爾倒頻譜係數、音高及週期性聲波在計算誤差時有不同的度量值，所以我們必須要轉換至同一度量上來做比對。本作品考慮之度量值為三者轉換後之總和，每個音其Score值之計算如下所示。

$$\text{Score} = \text{MFCC}(X \mid \text{model}) + \text{Pitch}(Y \mid \text{model}) + \text{Corr}(Z \mid \text{model})$$

此三者轉換函數分別為

$$\text{MFCC}(X|\text{model}) = \begin{cases} 1 & , \ if\ dx \le 40 \\ \frac{45-dx}{5} & , \ if\ 40 \le dx \le 45 \\ 0 & , \ if\ \text{其他值} \end{cases}$$

$$\text{Pitch}(Y|\text{model}) = \begin{cases} 1 & , \ if\ y \le 5 \\ \frac{10-dy}{5} & , \ if\ 5 < dy \le 10 \\ 0 & , \ if\ \text{剩餘的值} \end{cases}$$

$$\text{和}\ \ \text{Corr}(Z|\text{model}) = \begin{cases} 2 & , \ if\ \ 0.9 \le dz \le 1 \\ 1 & , \ if\ \ 0.8 \le dz \le 0.9 \\ 0 & , \ if\ \text{剩餘的值} \end{cases}$$

其中三者公式定義如下:

令 $dx = \min_{a \in model} \|x-a\|$ 表示x和模型間最小之梅爾倒頻譜係數誤差，

$dy = \min_{p \in model} |y-p|$ 表示y和模型間最小之基本週期差，

$dz = \min_{s \in model} corr(z,s)$ 表示聲波z和模型間最大之相關係數值。

由於我們是進行語者辨識，所以週期性聲波之相關係數將給予較高之權重，介於[0,2] 之間，其餘兩個介於[0,1]之間。本作品之語音訊號為「不要叫」三個音，因此整個值為三個單音之Score值總和，為了方便起見，我們將以Score代表三個單音之總和。對於每一個測試音，我們要判斷是否為自我語者或其他語者之語音，我們將給予 一門檻值α，若Score值大於α值，則為語者之音。由於所有測試語料有兩種錯誤，一種為自我語者本身的錯誤，另一種則為其他語者本身的錯誤，因此門檻值α之設定將使兩者之平均錯誤率最小，如公式(9)所示:

$$\alpha = arg\ \min_{\alpha}[(\text{語者錯誤個數} * 3|\alpha) + (\text{非語者錯誤個數}|\alpha)] \qquad (9)$$

肆、研究結果

表1和表2為自我語者及其他部分語者(以作者為測試模型下)測試所得之特徵值誤差，如梅爾倒頻譜係數(MFCC)、音高(Pitch)、相關性係數(Corr)和Score值。表1顯示自我語者之Score值介於7到11之間，遠遠大於其他語者之Score值(介於0到4之間)，如表2所示。圖8為所有語料(207句子)之Score值，當門檻值為7時，我們能100%區別出自我語者和其他語者。表3為每個人的資料庫與自己和其他人的資料庫所比對的結果，表3顯示自我資料所比對結果的正確率平均為96.62%，而與他人資料集比對結果的正確率平均則為96.62%。整體準確率非常高為99.47%。

表1. 自我語者之特徵值誤差及Score值

	MFCC誤差			Pitch誤差			Corr相關係數			Score值
	1	2	3	1	2	3	1	2	3	Score
1	33.51	33.4	34.05	0	0	0	0.92	0.96	0.84	11
2	32.12	33.71	36.46	1	0	0	0.82	0.93	0.87	10
3	32.61	47.47	34.05	0	0	52	0.94	0.88	0.57	7
4	29.38	39.83	35.94	0	2	0	0.88	0.85	0.94	10
5	42.43	34.6	34.65	0	1	1	0.88	0.96	0.61	8.51
6	41.16	41.16	32.56	0	0	1	0.94	0.89	0.91	10.5
7	39.89	36.85	33.91	0	0	0	0.94	0.89	0.91	11
8	38.89	27.87	29.71	0	0	0	0.94	0.89	0.9	11
9	35.79	36.24	28.87	0	0	0	0.8	0.9	0.87	10
10	33.61	28.83	34.62	1	0	0	0.95	0.9	0.86	11
11	33.49	33.99	29.86	1	0	0	0.92	0.9	0.89	11
12	29.77	30.09	30.17	0	0	2	0.91	0.88	0.42	9

表2. 其他語者之部分特徵值誤差及Score值

	MFCC誤差			Pitch誤差			Corr相關係數			Score值
	1	2	3	1	2	3	1	2	3	Score
13	45.5	60.33	55.34	11	52	7	0.23	0.37	0.48	0.5
14	46.96	55.83	55.43	10	33	9	0.15	0.22	0.52	1
15	47.03	55.33	5.1	7	45	25	0.45	0.13	0.38	1.5
16	54.06	59.44	56.33	9	24	6	0.11	0.24	0.46	0
17	46.55	56.96	53.52	30	27	21	0.36	0.38	0.2	2
18	46.56	58.87	54.63	11	6	4	0.27	0.09	0.58	2
19	56.56	57.5	55.32	6	35	0	0.67	0.16	0.36	0
20	55.57	57.77	54.29	13	30	11	0.19	0.3	0.45	1
21	48.52	58.51	54.61	14	42	6	0.25	0.24	0.48	0.5
22	47.72	58.74	53.83	9	16	22	0.38	0.35	0.23	3
23	59.99	52.41	57.42	1	0	0	0.62	0.39	0.5	4
24	50.31	58.9	58.65	0	0	0	0.81	0.43	0.44	3.5

SCORE值

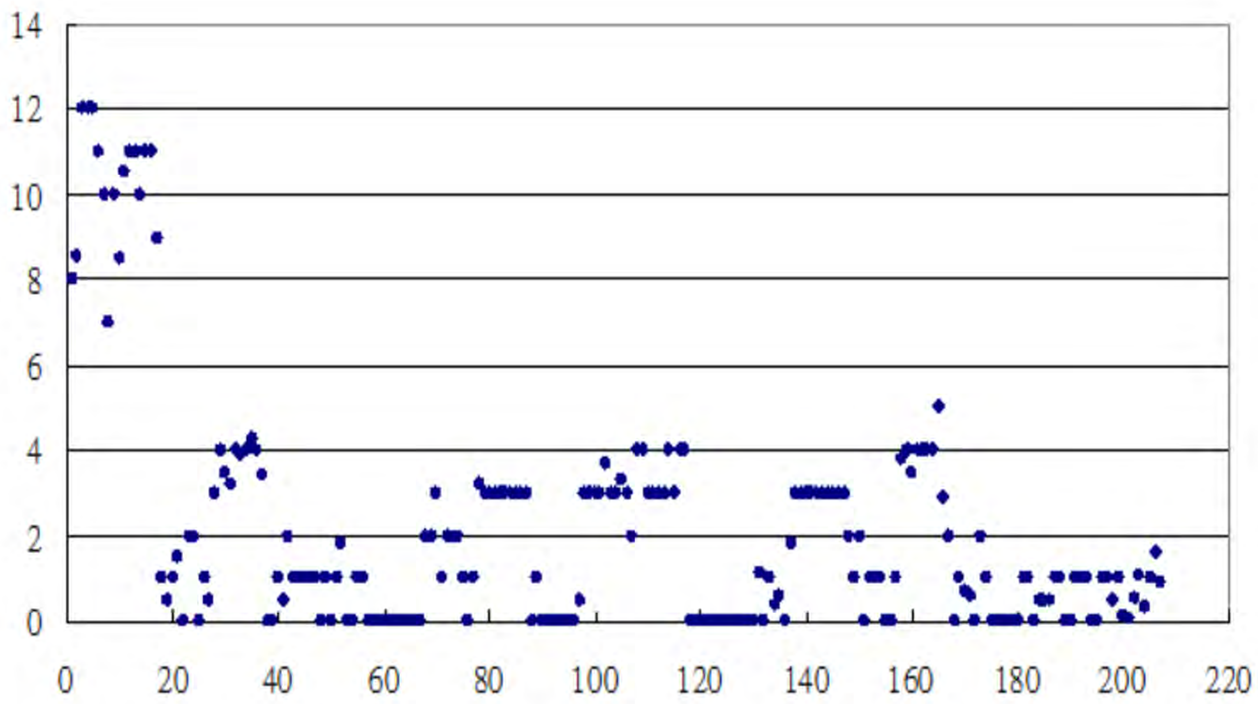


圖8. 所有語料「不要叫」207句子之Score值

表3. 資料庫之語者辨識率(個數)

語者	自我語者辨識率（個數）	其他語者辨識率（個數）	最佳門檻值
A	0.9 (9)	0,97 (192)	5.5
B	1 (10)	1 (197)	6
C	0.9 (9)	0.99 (196)	6
D	1 (10)	0.99 (196)	7
E	1 (10)	0.99 (196)	6
F	0.9 (9)	1 (197)	7
G	1 (10)	1 (197)	7
H	0.9 (9)	0.99 (196)	6
I	1 (10)	1 (197)	6
J	1 (10)	1 (197)	7
K	1 (10)	1 (197)	6
L	1 (10)	1 (197)	6
M	1 (10)	0.99 (196)	5.5
N	0.9 (9)	0.99 (196)	5.5
O	1 (10)	1 (197)	7
P	1 (10)	1 (197)	7
Q	0.8 (8)	0.99 (196)	6.5
R	1 (10)	0.99 (196)	6
S	1 (17)	1 (190)	7
T	1 (10)	1 (197)	7
平均值	96.62% (200)	99.62% (3918)	99.47%

伍、結論及未來展望

本作品中，我們使用了前置處理得出了真正所需的語音資訊，並以梅爾倒頻譜係數、基本週期、週期性聲波三種辨認方式所得的數據，透過轉換函數以及權重比例去加成為一Score值，並以最佳門檻值作為語者辨識。本方法簡易、有趣又有效。整體辨識率為99.47%，準確率非常高。

為了提高系統穩定及實用性，本作品可以:

1. 增加實驗取樣人數。
2. 思考外來狀況(如感冒、情緒等)對辨識可能的影響。

整體而言，本作品可應用於門禁系統或金融系統等相關領域。實測作品及影片示範亦可參閱。

陸、參考資料

[1] http://ir.lib.nchu.edu.tw/bitstream/11455/18757/1/nchu-102-7100018006-1.pdf。

[2] 王國榮 (2000)，Visual Basic 6.0 實戰講座 ，全華。

[3] 王小川 (2012)，語音訊號處理 修訂三版，全華。

[4] 吳俊甫 (2012)，時域上基頻軌跡演算法的改良與探討。國立中興大學應用數學系碩士學位論文。

[5] L.R. Robiner. “On the Use of Autocorrelation Analysis for Pitch Detection,”IEEE Trans.Acoustics, Speech, and Signal processing, vol. ASSP-25, no. 1, pp. 24-33, Feb. 1997.