運用語音鑑別技術之點名系統設計

魏吟芳*
高苑科技大學電子工程系
副教授
E-mail:t10014@cc.kyu.edu.tw
associate professor
corresponding author

翁國維[1]
高苑科技大學電子工程系
學生
E-mail:lan30198@gmail.com
undergraduate student

* Yin-Fang Wei, [1]Guo-Wei Weng, Department of Electronic Engineering, Kao Yuan University, No.1821, Zhongshan Rd., Luzhu Dist., Kaohsiung City 821, Taiwan (R.O.C.),TEL: 886-7-6077142, FAX: 886-7-6077000

摘要

本研究採用梅爾倒頻譜係數(Mel-Frequency Cepstral Coefficient, MFCC)以及動態時間扭曲(Dynamic Time Warping, DTW)技術進行語音鑑別，完成點名系統之設計，系統以 LabVIEW 軟體撰寫。研究的目的是希望在吵雜教室環境下能準確鑑別說話人。點名系統設計過程是將說話人預先在系統中進行語音訓練，產生六組參考語音向量字典庫，取 39 維的梅爾倒頻譜係數後，藉由動態時間扭曲比對得到良好之語音辨識度。

關鍵詞：梅爾倒頻譜係數、動態時間扭曲、LabVIEW、說話人鑑別。

Abstract

In this study, Mel-Frequency Cepstral Coefficient and Dynamic Time Warping are techniques used to implement a speaker identification system. The software LabVIEW is adopted as development environment and control system. Speaker verification is the aim of our system for any complex environment. 6 sets of reference voice vectors are database made by each speaker. Experiment results show the method for speaker verification is available.

Keywords: Mel-Frequency Cepstral Coefficient (MFCC), Dynamic Time Warping (DTW), LabVIEW, Speaker Verification.

1. 前言

語音辨識技術已日臻成熟，不論門禁裝置、防盜系統、居家保全…等，皆有其運用。在校園內，教師對學生點名一直是在課堂時間運用上的一項麻煩，因為十分廢時，有些課程學生人數甚至有超過百人的情形，點名方式漸漸衍生多樣的方式，例如採學生自主簽到、讀卡式線上點名、安排固定座位點名…等因應方式。有別於以上各種方式，結合語音辨識鑑別說話人特性來進行點名，讓學生進入教室門口對語音點名裝置進行點名，不但能大幅節省時間，也能間接防範頂替點名等弊端。

2. 梅爾倒頻譜係數(MFCC)

梅爾頻率倒譜係數，主要是在頻譜領域上討論，本研究使用的參數計算從預強調(Pre-emphasis)開始，經過漢明窗(Hamming window)處理到快速傅立葉轉換(Fast Fourier Transform)將時域信號轉換成頻域信號，將功率頻譜(Power Spectrum)經過梅爾頻率(Mel-Frequency)平均分布的三角濾波器組處理，最後對各個濾波器的輸出所形成的向量進行離散餘弦轉換[1-2]。將 MFCC 作為語音識別系統中的特徵，系統可以自動識別語音中的數字內容，同樣也適用於鑑別說話人[3-4]，在本研究的點名系統中，前半段採 MFCC 而後半段是採 DTW 做語音識別。

2.1 預強調(Pre-emphasis)

為了彌補聲音訊號的衰減，使用預強調的方式補償，讓訊號通過一個高通濾波器，補償高頻訊號的衰減。如式(2-1)所示：

$$\hat{x}[n]=[x]n-\alpha \cdot x[n-1] \qquad (2\text{-}1)$$

其中$\hat{x}[n]$表示聲音訊號，n 為時間係數，在這裡α取 0.95 [5-6]。圖 2-1 為以人名“翁國維”做原

始訊號範例，圖 2-2 為預強調之後的訊號。

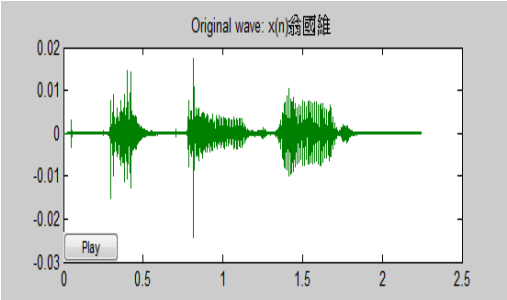

圖 2-1 原始訊號

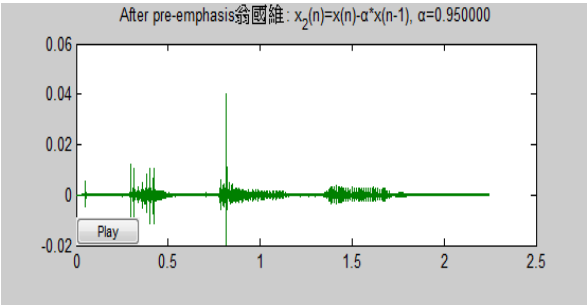

圖 2-2 經過預強調處理後之訊號

經過了預強調之後，很明顯地，聲音變的比較尖銳清脆，但是音量卻相對變小了。

2.2 漢明窗(Hamming Window)

為了使音框之間的訊號變化量不致於太大，通常取 1/2 的比例來重疊音框[7]，第 m 個音框的訊號如式(2-2)所示：

$$\hat{x}_m = \hat{x}[n] \cdot w[n] \qquad (2\text{-}2)$$

其中$\hat{x}[n]$是預強調後訊號，w[n]是窗函數(Windows function)，本研究採用的 Windows function 如式(2-3) [8-9]。

$$w[n]=\begin{cases}(1-\beta)-\beta\cdot\cos\left(\frac{2\pi n}{N-1}\right)\\0\end{cases} \qquad (2\text{-}3)$$

且 0≤n<N-1。

圖 2-3 為漢明窗曲線圖，圖 2-4 為經過漢明窗處理後之訊號。

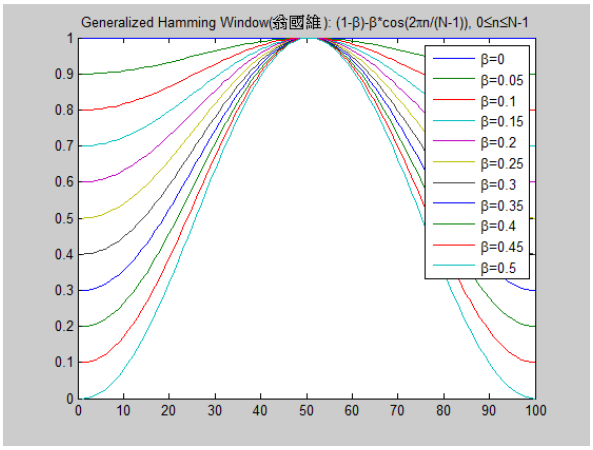

圖 2-3 漢明窗曲線圖

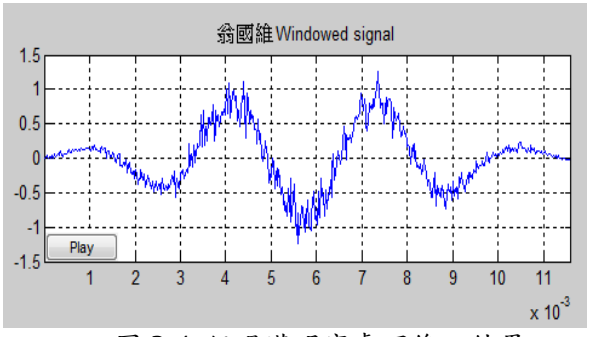

圖 2-4 經過漢明窗處理後之結果

2.3 快速傅立葉轉換(FFT)

經過預強與漢明窗的處理動作後，接著使用快速傅立葉轉換將時域訊號轉換到頻域訊號，可觀察出頻域訊號的特性。圖 2-5 為經過 FFT 之訊號結果。

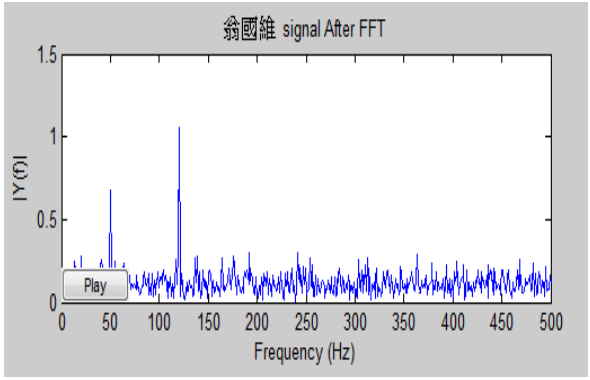

圖 2-5 FFT 處理後之訊號結果

2.4 梅爾倒頻譜參數

人類聽覺接收頻率範圍為 20 至 20 kHz，梅爾頻率模仿人耳對於頻率的感受，以對數變化表示。三角帶通濾波器（Triangular Band-pass Filters）主要目的是將頻譜進行平滑化，使原

始訊號共振峰值較明顯，且可以降低資料量。先將頻譜能量乘上一組 20 個三角帶通濾波器，求出濾波器的對數能量，這 20 個三角帶通濾波器是平均分布在梅爾頻率上的。濾波器在頻率軸的間隔稱為梅爾刻度(Mel Scale) [10-11]，轉換公式為式(2-4)，

$$f_{mel}=2595\cdot log_{10}\left(1+\frac{f}{700}\right) \qquad (2\text{-}4)$$

圖 2-6 為 Labview 程式碼，將頻率(Hz) 轉換至梅爾刻度

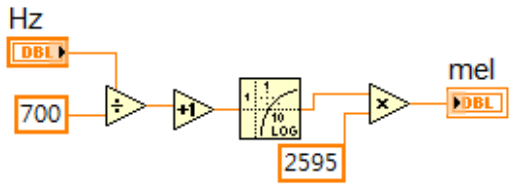

圖 2-6 將頻率(Hz) 轉換至梅爾刻度

2.5 離散餘弦轉換(DCT)

將上述的 20 個對數能量E_k帶入離散餘弦轉換，求出 L 階的 Mel- scale Cepstrum 參數，這裡 L 的值以 12 帶入。離散餘弦轉換公式如式(2-5)所示：

$$C_m=\sum_{k=1}^{N}\cos\left[m*(k-0.5)*\frac{\pi}{N}\right]*E_k \qquad (2\text{-}5)$$

且 m=1，2…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 轉換是期望轉回類似時域的情況來看，又稱逆頻域(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），對數能量表示一個音框的音量，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，如式(2-6)所示：

$$\Delta C_m(t)=\frac{\left[\sum_{t=-M}^{M}C_m(t+\tau)\tau\right]}{\sum_{t=-M}^{M}\tau^2} \qquad (2\text{-}6)$$

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。一般我們在電腦上進行的語音辨識，就是使用 39 維的特徵向量，在本研究中，M 取 3 為值，也就是 39 維向量[12]，如圖 2-7 所示。

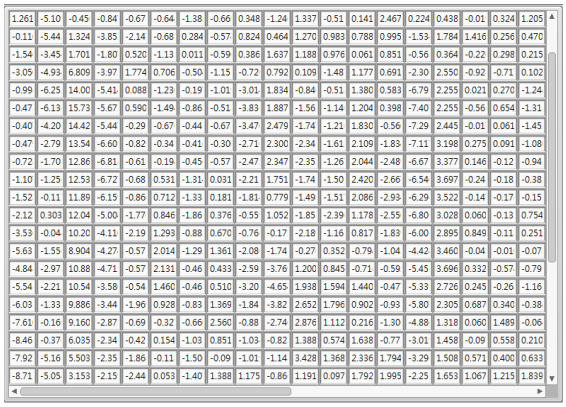

圖 2-7 Labview 取 39 維特徵向量

2.6 動態時間扭曲(DTW)

動態時間扭曲(DTW)使用動態規劃(Dynamic Programming, DP)的方式，在比對特徵時，必須對經由梅爾頻率倒頻譜係數的離散餘弦轉換得到的結果搜尋最佳路徑，且要不斷確認先前路徑的最佳值。如圖 2-8 所示。

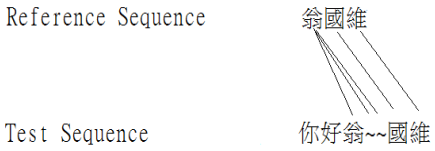

圖 2-8 比對示意圖

由於 Test Sequence 的聲音在時間上發生了扭曲，「翁」這個音延長了幾秒，Dynamic time warping 正是用來匹配兩者之前的距離。假設 Test Sequence 與 Reference Sequence 之間的距離為 distance(t, r)，歐幾里德距離[13-14]，如式(2-7)所示，

$$Dist(t,r)=\sqrt{\sum_{1}^{d}(a_i-b_i)^2} \qquad (2\text{-}7)$$

相似性和最短距離呈負相關，最短距離越大，相似性越低，反之亦然。在大部分音頻處理的過程中，DTW 所謂的最短的要求是忽略上時間上發生的扭曲，使得 Test Sequence 與 Reference Sequence 是相似的結果或者相同的結果。路徑圖如表 1 所示：

表 1 路徑比對圖

	你	好	翁	翁	翁	國	維
翁	0	0	1	1	1	0	0
國	0	0	0	0	0	1	0
維	0	0	0	0	0	0	1

在表 1 中的 1 表示所走的路徑，從左上方走到右下方的最短路徑所消耗的 Cost 為 DTW 的最短距離。因為時間具有不可逆性，所以走法分為三種，右、右下、下，稱為「局部關係」。因為 Test Sequence 多了「你好」兩字，使得距離為 2。

3. 結果與討論

對於不同的上課環境，須考慮面臨背景複雜的噪音進而影響比對辨識率，本研究以“翁國維”、“陳盈秀”、“Bella”、“黃伯傑” 四位同學個別採有雜音與靜音之比對辨識率來探討。每位同學在系統中進行語音訓練六次，並將音訊轉為向量儲存於系統資料庫當中成為字典檔，藉由 DTW 比對輸入訊號與六組的字典檔找出最佳結果(Best Match)，並以人機介面及系統介面呈現出研究結果。

圖 3-1 及圖 3-2 為第一位學生有雜音環境人機介面與系統介面之比對結果，人機介面顯示完成簽到，系統介面顯示 Best Match =5，表示輸入的音訊與當初訓練的第 5 組 Cost 最低，完成正確鑑別，圖 3-3 及圖 3-4 為第一位學生靜音環境比對結果，Best Match =4 比對出第 4 組字典檔 Cost 最低。圖 3-5 及圖 3-6 為第二位學生有雜音環境比對結果，Best Match =4 表示輸入的音訊與當初訓練的第 4 組 Cost 最低，圖 3-7 及圖 3-8 為第二位學生靜音環境比對結果，Best Match = 1 輸入的音訊與當初訓練的第 1 組 Cost 最低。圖 3-9 及圖 3-10 為第三位學生有雜音環境比對結果，圖 3-11 及圖 3-12 為第三位學生靜音環境比對結果。圖 3-13 及圖 3-14 為第四位學生有雜音環境比對結果，圖 3-15 及圖 3-16 為第四位學生靜音環境比對結果。

在有雜音環境當中，辨識率的提升可透過適當選擇參考語音字典庫之音高、清晰度或音量來當控制變因，在本研究中，選用音量來當控制變因，當測試學生提高音量，系統會有較佳的辨識率，以本研究的 Best Match 參數而言，四位說話人皆可得到成功的鑑別。

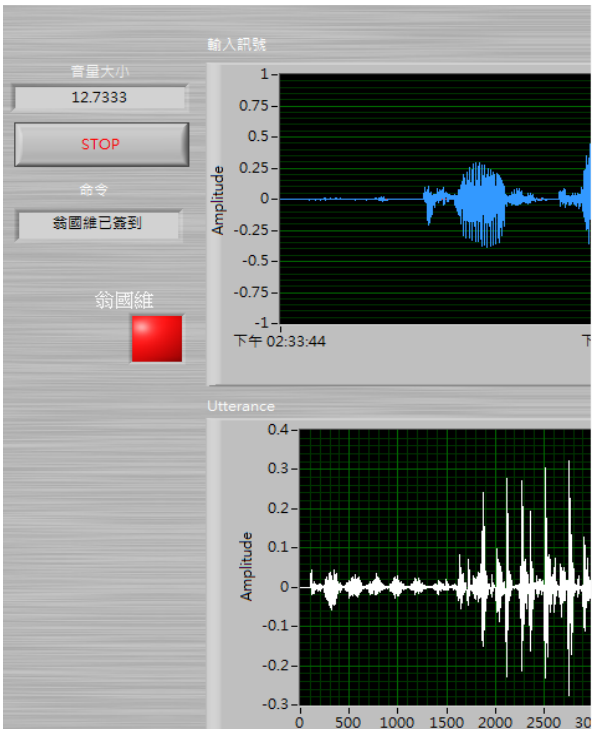

圖 3-1 翁國維有雜音環境人機介面

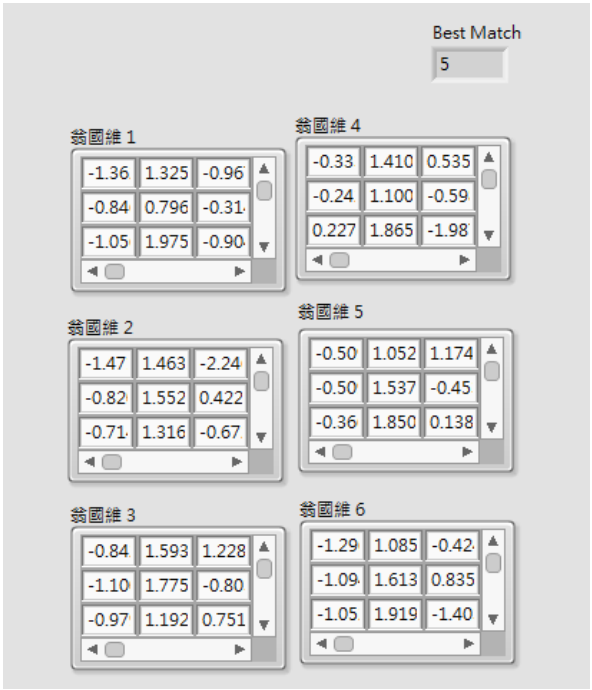

圖 3-2 翁國維有雜音環境系統介面

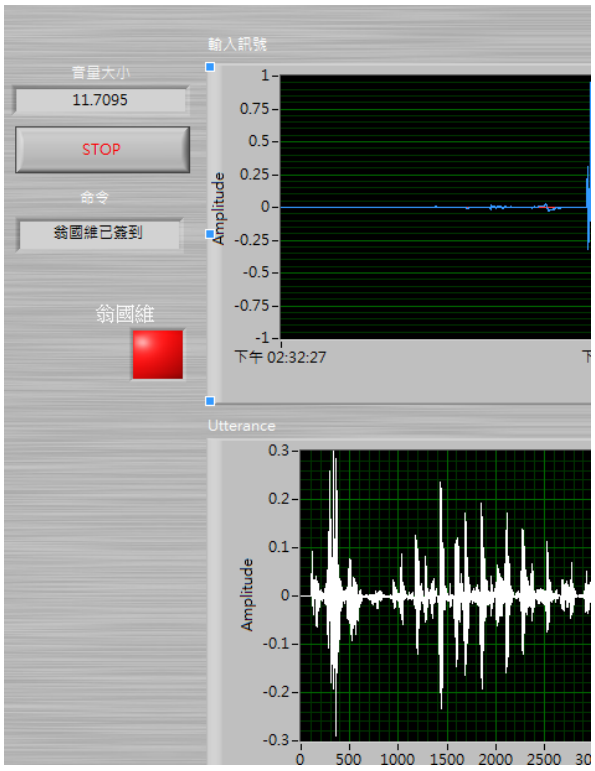

圖 3-3 翁國維靜音環境人機介面

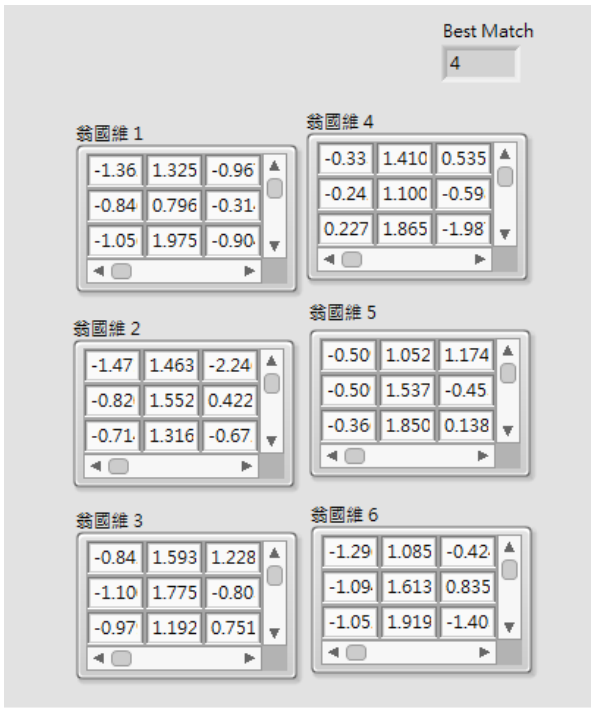

圖 3-4 翁國維靜音環境介面

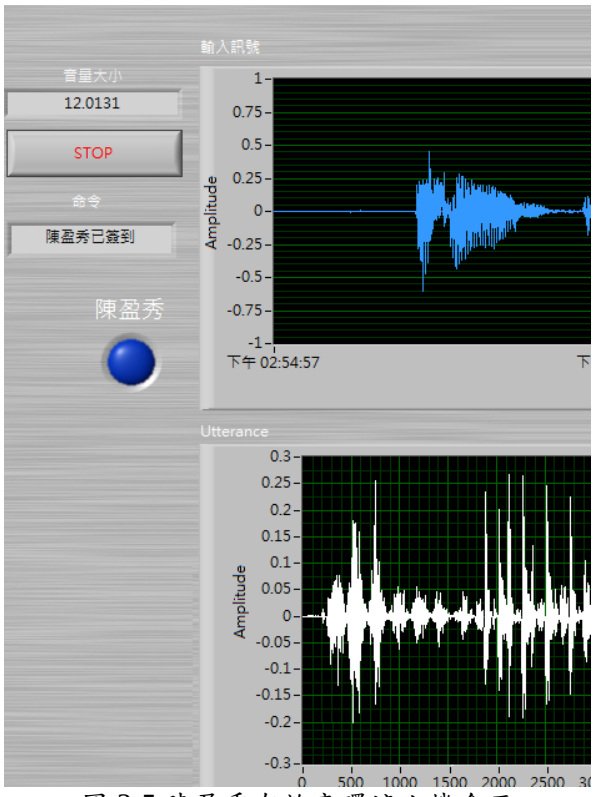

圖 3-5 陳盈秀有雜音環境人機介面

圖 3-6 陳盈秀有雜音環境系統介面

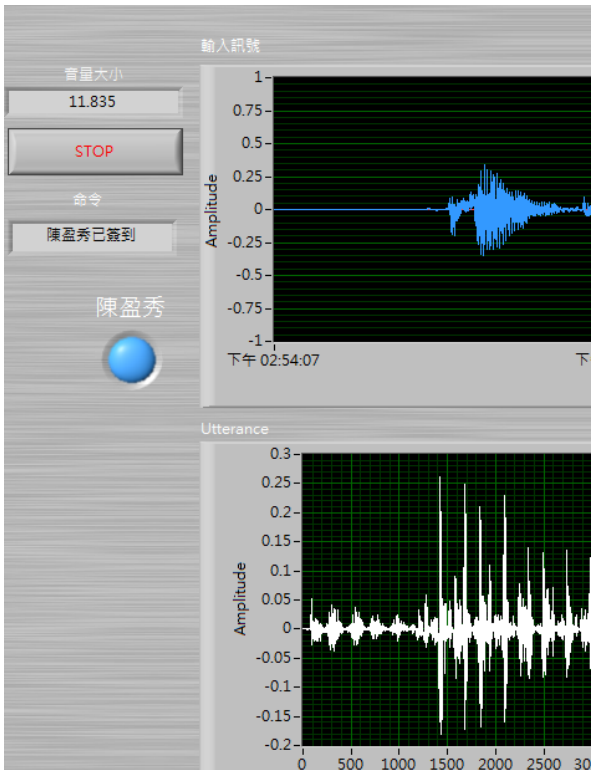

圖 3-7 陳盈秀靜音環境人機介面

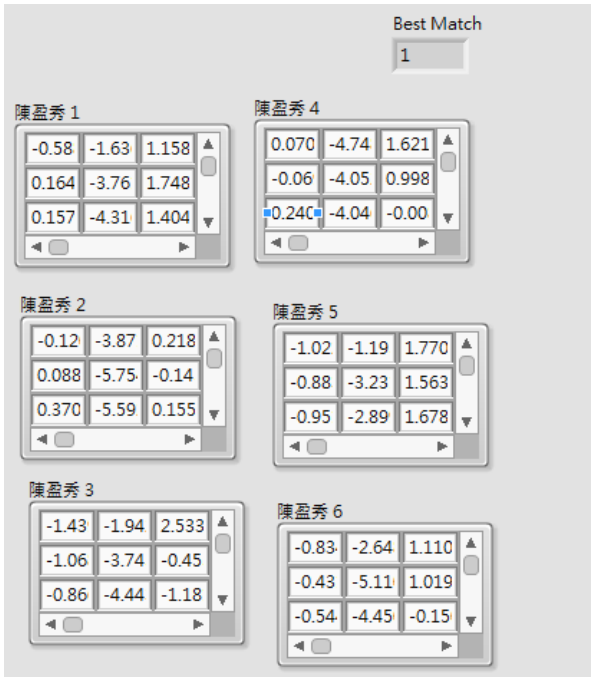

圖 3-8 陳盈秀靜音環境系統介面

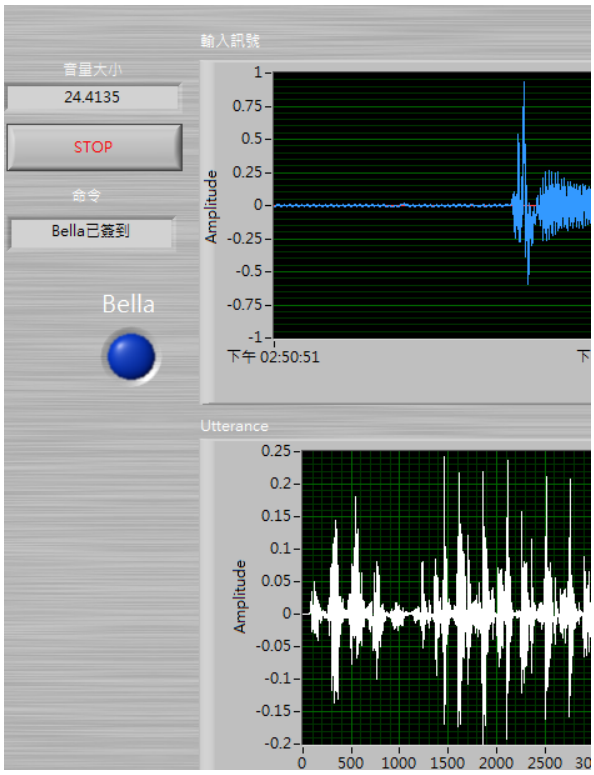

圖 3-9 Bella 有雜音環境人機介面

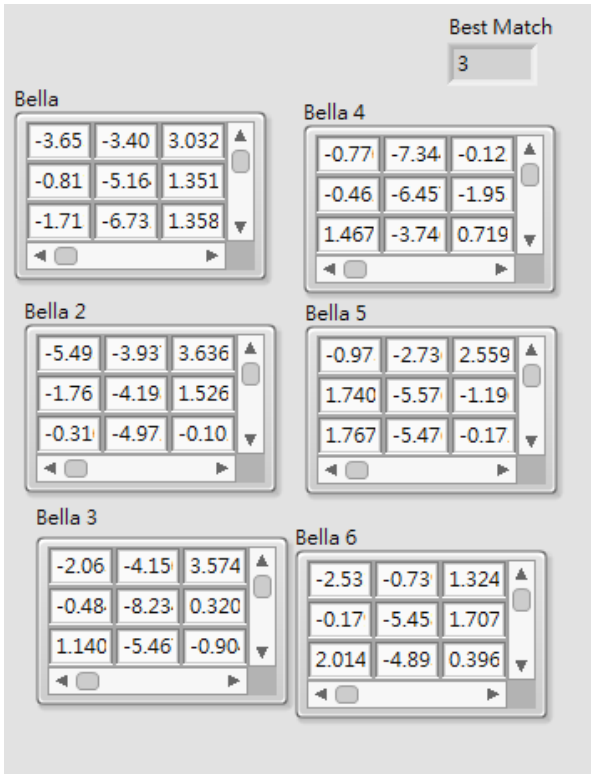

圖 3-10 Bella 有雜音環境系統介面

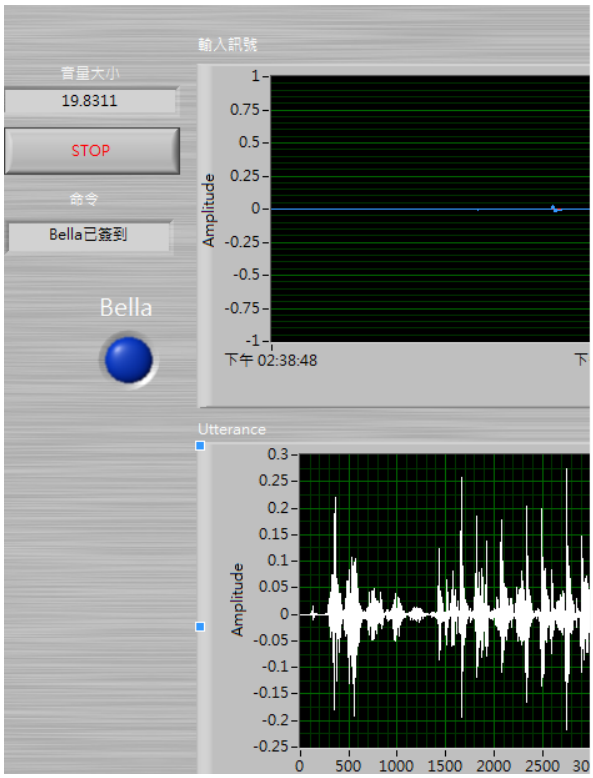

圖 3-11 Bella 靜音環境人機介面

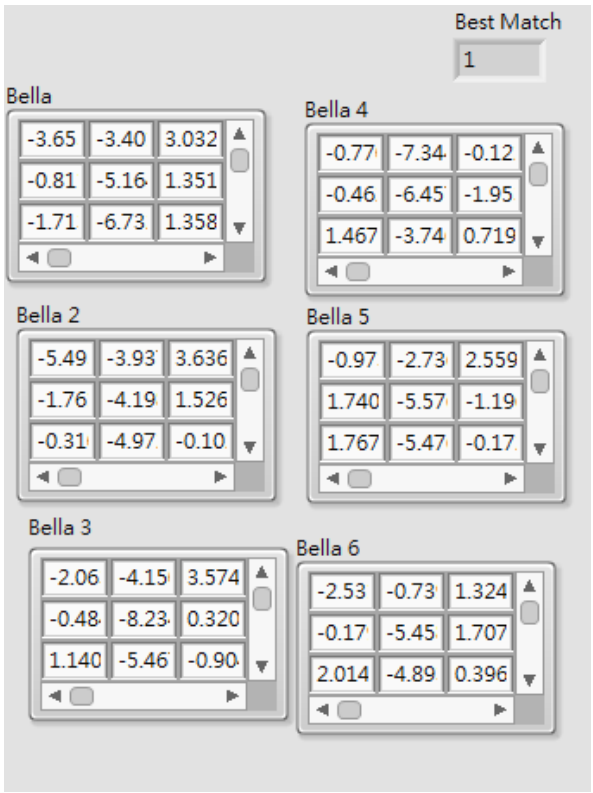

圖 3-12 Bella 靜音環境系統介面

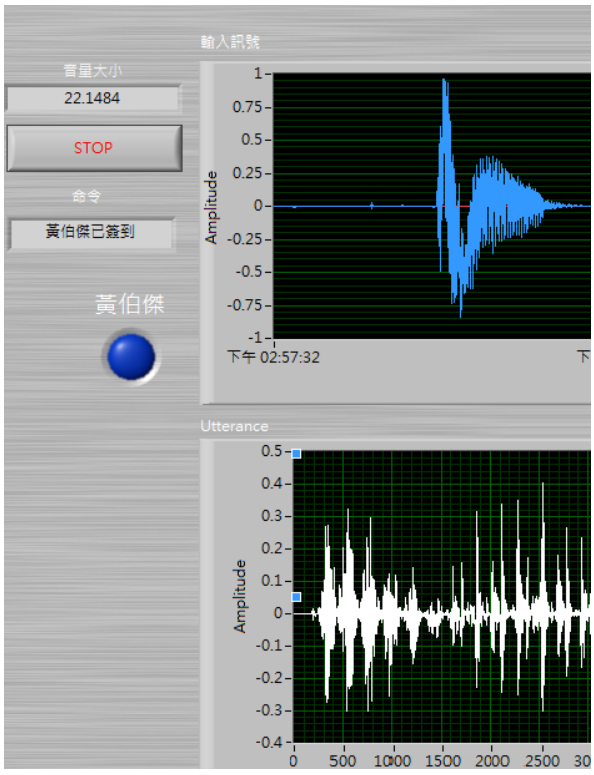

圖 3-13 黃伯傑有雜音環境人機介面

圖 3-14 黃伯傑有雜音環境系統介面

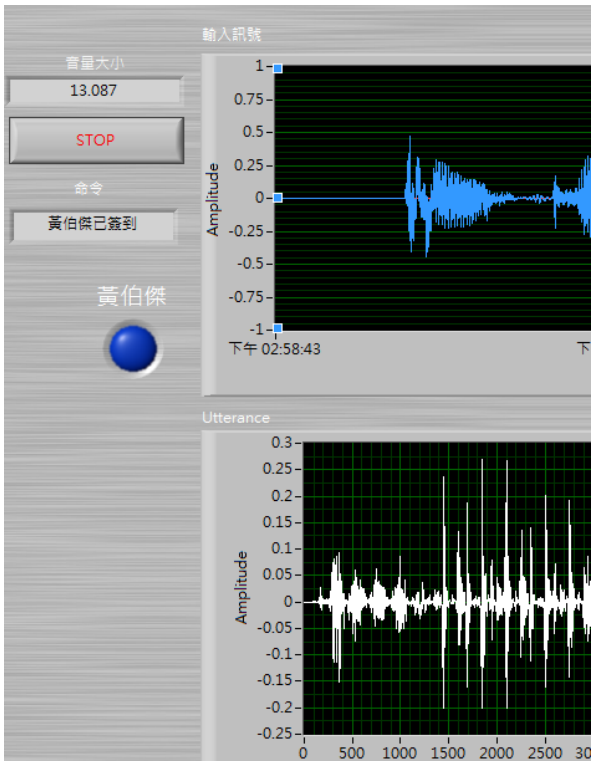

圖 3-15 黃伯傑靜音環境人機介面

圖 3-16 黃伯傑靜音環境系統介面

參考文獻

[1] Wan-Yu Chen, “ The Investigation of Capturing Mel-Frequency Cepstrum Coefficient Features on Mandarin Consonant Word Recognition”, Master dissertation, National Chung Hsing University, pp.17-19, 2011

[2] 陳柏琳,“以能量為基礎之語音正規化方法研究及其於語音端點偵測之應用”，碩士論文，國立臺灣師範大學，pp.9-10，2007

[3] T. Ganchev, N. Fakotakis, G. Kokkinakis, “Comparative evaluation of various MFCC implementations on the speaker verification task”, Master dissertation, University of Patras, pp.1-4, 2005

[4] 劉炳哲、李宗葛，“基於 MFCC 和加權矢量量化的說話人識別系統”，碩士論文，上海復旦大學，pp.127-128，2004

[5] 周智勳、林秉韶，“最佳化梅爾倒頻譜係數之研究及其於音樂曲風辨識之應用”，資訊科技與應用期刊，Vol. 4，No. 1，pp. 53-58，2010

[6] 陳柏琳、張志豪，“強健性和鑑別力語音特徵擷取技術於大詞彙連續語音辨識之研究”，碩士論文，國立臺灣師範大學，pp.9-10，2005

[7] 李建樹、吳紹敬，“以階層式整合時間與非時間特徵之音樂曲風分類法”，第十二屆離島資訊技術與應用研討會論文集，pp.314，2013

[8] 林士翔，“數據擬合與分群方法於強健語音特徵擷取之研究”，碩士論文，國立臺灣師範大學 ，pp.12-13，2005

[9] Guo-Syun Huang, Hung-Yan Gu, “ Implementation of a Voice Command Recognition System for Mobile Devices”, Master dissertation, National Taiwan University of Science and Technology，pp. 1-4，2008

[10] 凌欣暉,“強健性語音辨識及語者確認之研究”，碩士論文，國立中央大學，pp. 44-60，2010

[11] 王孙武、林進發、陳俞多、陳恆鳴、翁傳翔、黃子豪 、連逸傑、曾聖傑、李昱緯，“夜行性動物聲音收錄與辨識系統 ”，行政院農業委員會林務局保育研究系列 99-21 號，pp. 24-30，2011

[12] Yu-Ching Liu, “Research on Text-dependent Speaker Identification with Screening Mechanism for Impropriate Speech Inputs”, Master dissertation, National Tsing Hua

University, pp.8-13, 2009
[13] 鄭琨燁,“在地理資訊系統環境下以行動裝置實現行程規劃”，碩士論文，國立雲林科技大學，pp.10-15，2007
[14] 陳延洛，“基因表現時間序列的叢集分析方法與系統實作”，碩士論文，國立成功大學，pp.20-25，2002