電動滑板車語音辨識應用設計

陳茂林[1]鄭博懋[1]何建昌[1]施政樂[2]朱吉皇[2]陳鵬仁[3]
[1]建國科技大學自動化工程系暨機電光系統研究所
[2]建國科技大學電機工程研究所
[3]建國科技大學體育室
彰化市介壽北路 1 號
E-mail:mm122048@ctu.edu.tw

摘要

本文研究主要是利用動態時軸校正方法來設計電動滑板車語音辨識應用，期使讓低收入戶肢體殘障人士能有便利的應用交通工具。首先錄製與訓練語音資料庫，然後對語音訊號做前處理，採取梅爾倒頻譜方法擷取語音特徵值參數，再應用動態時軸校正方法比對辨識測試語音與基本語音資料庫做驗證結果。系統經由實驗驗證語音辨識應用於電動滑板車可以有效實現中低收入戶肢體殘障人士的交通工具使用普遍性，替代電動輪椅的高價位推展，真正落實照護的品質。

關鍵字: 動態時軸校正、語音辨識、梅爾倒頻譜、電動輪椅。

1.前言

本文根據P.K.Sharma等人發表的“Real time control of DC motor drive using speech recognition”[1]探討即時語音辨識控制直流馬達。語音處理方面，目前語音辨識多數使用梅爾倒頻譜係數(MFCC)方法擷取語音特徵值，此參數能夠表示人耳對高頻與低頻不同的感受程度，適合用在語音辨識。[2]擷取特徵值步驟需要經過下列計算：1.預強調、2.音框化、3.漢明窗、4.快速傅立葉轉換、5.梅爾濾波器組、6.離散餘弦轉換。

語音分析方面，根據文獻[2] T.B.Amin 發表的“Speech Recognition using Dynamic Time Warping”利用動態時間校準(Dynamic Time Warping，DTW)辨識分析比對驗證，找出測試語音和參考語音兩者語音向量之間最短的距離。系統實作方面，利用語音辨識微控制器 HM2007 為中央處理模組控制電動滑板車直流馬達運轉，進而探討語音辨識的準確性與語音控制電動滑板車駕駛的便利性。

本論文之研究透過語音辨識控制電動滑板車直流馬達。其研究目的為：1.語音辨識控制做為電動滑板車動力控制系統。2.中低收入肢體障礙者便利駕駛的交通工具。3.透過語音辨識模組系統，駕駛者可依自己喜好，錄製自己慣用牢記的辭彙。

2.語音辨識

任何人類所能聽見的聲音都稱為語音。語音信號原本在空氣中以聲波的形式傳播，我們將這種聲波訊號稱為類比訊號（Analog signal）。語音訊號的波形，在不同的時間區段上會出現不同的週期，這種隨時間變化的訊號，稱為非固定式（nonstationary）的訊號。非固定式的語音訊號，可以用處理固定式訊號的方式來對非固定式訊號作處理。語音訊號處理，經過麥克風將聲波轉換為電訊號，再透過ADC轉換成數位數字表示。在每一段取樣時間，將聲波的振幅轉換成數位化數值，即數位訊號處理描述語音訊號的波形。語音訊號處理目的是得到語音特徵參數以便電腦有效傳輸與存取，或者達到語音編碼、語音合成與語音辨識的應用功能[8]。

語音數位[8]取樣表示式為：

$$x(t) = e^t, t \in \Re$$

在數位系統的訊號，資料都是以位元為單位儲存，這些訊號都是離散的，稱之為「離散時間語音訊號」(Discrete Time Speech Signal)。表示式為：

$$u[n] = \sin n, n = 0,1,2,3,\ldots$$

將類比訊號轉換成數位訊號的過程，需要經過取樣(sampling)與量化(quantization)。取樣就是將類比訊號乘上一個周期性脈衝訊號，所得一序列的脈衝，脈衝大小也就是該時間點上類比訊號的振幅。表示式為：

$$a_p(t) = a_x(t)\,p(t)$$

時間函數 $a_x(t)$ 表示為一個類比訊號，取樣周期為 T_S ，取樣後的訊號為 $a_p(t)$ 。 $p(t)$ 為一個脈衝序列訊號，表示式為： $p(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT_s)$

描述語音波形的取樣如圖 1 所示。

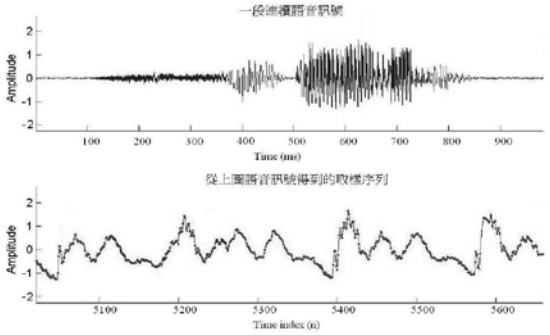

圖 1 語音訊號取樣波形圖

取樣後波形訊號 $a_p(t)$ ，只有在 $t = nT_S$ 時有值，其值為 $a_x(nT_s)$ ，也能寫成 $a(n)$ ， n 是一個整數常數，其餘的時間 $a_p(t)$ 為零。 $a_p(t)$ 表示式為：

$$a_p(t) = \sum_{n=-\infty}^{\infty} a_x(nT_s)\delta(t - nT_s) = \sum_{n=-\infty}^{\infty} a(n)\delta(t - nT_s)$$

此訊號在時間軸上為不連續訊號，稱之為離散時間語音訊號(Discrete Time Speech Signal)。 T_s 為取樣週期， $F_s = 1/T_s$ 為取樣頻率。

語音特徵值擷取[8]：採用梅爾倒頻譜方法(Mel-Frequency Cepstral Coefficients, MFCC)做擷取語音特徵參數，由於人類天生對低頻的聲音察覺較敏銳，而對高頻的聲音察覺較含糊，因此擷取特徵參數使用MFCC方法能夠使低頻的部份佔多數，而高頻的部份佔較少數。

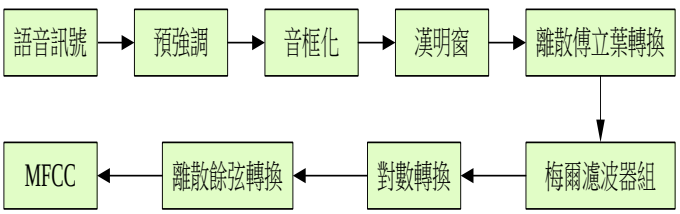

圖 2 MFCC 特徵擷取流程圖

圖 2 為梅爾倒頻譜方法擷取語音特徵值的流程圖，MFCC 的流程步驟如下：

(1)預強調[8]

預強調的作用主要是加強語音訊號高頻的部份，讓語音訊號通過一個高通濾波器(High-Pass Filter)，因為語音訊號在空氣中傳送時，頻率較高的部份會隨時間變化而衰減，預強調可以補償衰減的高頻損失，所以藉由預強調補償高頻衰減。常見用於預強調的高通濾波器之 Z-轉換(Z-Transform)其數學表示式如下：

$$Y(z) = H(z)X(z) \qquad where \qquad H(z) = 1 - \alpha z^{-1}$$

其中 α 為預強調的參數，α 值設定通常設為 0.95 左右，若將此高通濾波器之反 Z-轉換轉成時域(Time Domain)可以改寫成如下：

$$y[n] = x[n] - \alpha x[n-1] \quad 0.9 \le \alpha \le 1$$

其中 $y[n]$ 為語音訊號在時域上的第 n 個採樣點。經過預強調處理與未經過預強調處理的振幅比較圖如圖 3 所示，由圖中可以看出，經過預強調處理後，低頻部份的能量確實被壓抑住，而高頻的能量相對被加強。

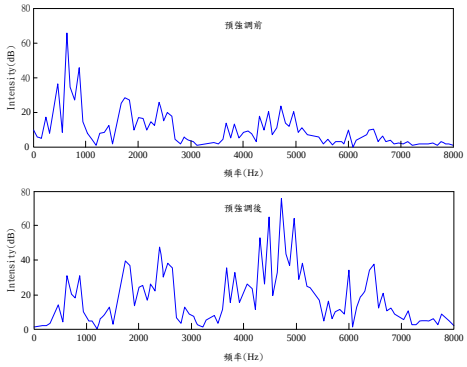

圖 3 預強調前後語音訊號之頻譜分佈

(2)音框化[8]

語音訊號處理中，通常假設聲音的特徵是緩慢變化的。為了處理語音訊號，會用一段固定時間的視窗套上去稱為加窗，也就是一次僅處理窗中的數據。而加上去的一段語音訊號即稱為音框化(Frame blocking)。若將視窗移動到下一個標記的時間點，即可得出下一個音框。圖 4 描述一個語音訊號加窗與音框化的情形。假設視窗的長度為 N，在時域中可以寫成

$$w(n) = \begin{cases} S_w(n), & 0 \le n \le N-1 \\ 0, & otherwise \end{cases}$$

$S_w(n)$ 表示為視窗的加權。

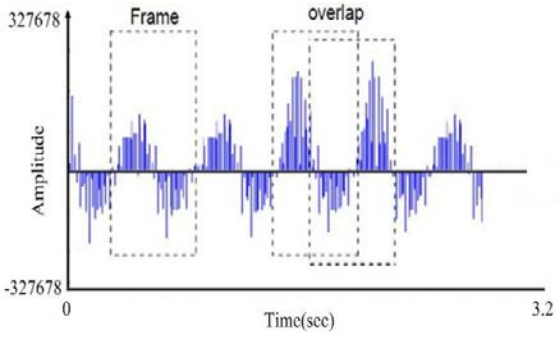

圖 4 語音訊號音框化

語音訊號加窗時，等於將窗乘在語音訊號中的某一時間點上，數學式表示如下：

$$f_x(n,m) = x(n)w(m-n)$$

m 為音框位置，當 n 在 $m-N+1 \le n \le m$ 之間，$f_x(n,m)$將不等於零，所以當 $x(n)$ 在 $m-N+1 \le n \le m$ 這一時段被取出，作為一個音框。

(3)漢明窗[8]

音框處理之後，離散傅立葉轉換成頻域訊號，每個音框大小是固定時間點切割，音框邊緣會形成不連續的訊號現象，使得音框經由離散傅立葉轉換之後產生了高頻雜訊。為了減低高頻雜訊，所以音框在離散傅立葉轉換前會乘上一個漢明窗，以增加音框邊緣間的連續性，讓各個音框在頻譜上的能量能更加集中。漢明窗其數學式如下：

$$h(n)\begin{cases}(1-\alpha)-\alpha\cos(\dfrac{2n\pi}{N-1}), & n=0,1,\ldots,N-1\\ 0 & otherwise\end{cases}$$

其中 α 為漢明窗的調整參數，不同的 α 值會產生不同的漢明窗。如圖 5 所示。

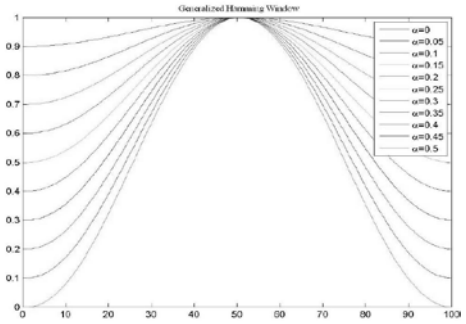

圖 5 漢明窗曲線圖

(4)離散傅立葉轉換[8]

語音訊號在時域上變化會隨著時間不斷改變，因此語音訊號在時域上無法作有效的探討。在頻域上短時段語音訊號呈現周期性，一般經由離散傅立葉轉換(Discrete Fourier Transform, DFT)將語音訊號由時域轉換為頻域，在頻域中觀察語音訊號的特性，或是抽取出頻域中的特徵參數。假設一個語音訊號 $x(n)$ ，作離散傅立葉轉換，其數學式如下：

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

為了減少計算量通常真正演算時，會使用簡化過程，使得演算速度加快，叫做快速傅立葉轉換(fast Fourier transform, FFT)。$W_N=e^{-j\frac{2\pi}{N}}$,因為 N 為 2 的整數次方，即 $N=2M$,可以推得

$$X(k)=X_{even}(k)+X_{odd}(k)\cdot W_{2M}^{k} \qquad ,k=0,1,2,\ldots M-1$$

$X_{even}(\mathrm{k})$和 $X_{odd}(k)$數學式為：

$$\begin{cases}X_{even}(k)=\sum_{n=0}^{M-1}x(2n)W_M^{nk}\\ X_{odd}(k)=\sum_{n=0}^{M-1}x(2n+1)W_M^{nk}\end{cases}$$

(5)梅爾濾波器組[8]

梅爾濾波器，近似於三角形的遮蔽曲線，由彼此跨越相鄰頻帶的三角形濾波器組成。對頻譜進行平滑化，並消除諧波的作用，凸顯原先語音的共振峰。在頻域中以梅爾刻度(Mel scale)劃分頻帶。梅爾刻度使得所有三角形濾波器的中心頻率 1kHz 以下為等間隔，1kHz 以上為對數間隔。圖 6 展示由三角形濾波器組成的梅爾濾波器組。

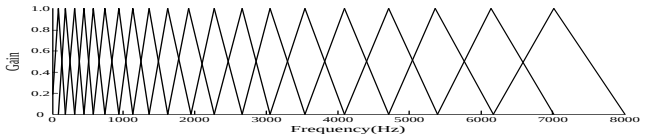

圖 6 三角形濾波器

將各頻譜能量乘上三角形濾波器後，累加起來，結果就是通過這個濾波器的能量，然後取對數值，得到對數能量。數學式為：

$$H(m) = \log\{ \sum_{k=f_{m-1}}^{f_{m+1}} | X(k) |^2 \ B_m(k) \}$$

(6)離散餘弦轉換

之後對全部濾波器輸出的對數能量，作離散餘弦轉換(discrete cosine transform, DCT)，得到梅爾倒頻譜參數。

3.動態時軸校正

動態時軸校正(Dynamic Time Warping，DTW)為傳統的語音辨識方法之一。兩段不同語音訊號，通常長度不同的情況下，採用最佳化控制的動態規劃(Dynamic Programming，DP)方式做比對計算。比對兩個向量之間最佳路徑的最短距離。2D平面上橫軸表示測試語音，縱軸表示參考語音，交叉點代表從兩個語音中各取出一個音框的特徵參數做比對，呈現一個座標平面。如圖 7 所示，橫軸長度為N個音框，縱軸長度為M個音框，在座標平面上做比對就會連成一條比對路徑，讓座標軸做合理的伸縮調整，通常是讓縱軸上的參考語音做伸縮調整。從第k-1 點到下一個第k點，稱為一個轉移(Transition)，完成一條比對路徑，包含許多轉移與距離計算，累積的距離最小代表兩段語音最相似[2]、[3]、[14]、[16]、[17]。

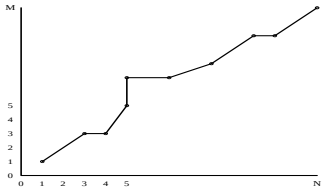

圖 7 動態時軸校正演算法比對座標圖

語音訊號在時序情況下做比對過程，時軸代表說話的快慢。橫軸與縱軸都是漸增，不可能有遞減的現象。因此比對路徑在座標平面上的斜率不應該大於 2 倍也不會小於 1/2，稱為全程路徑限制(Global path constraint)，在座標平面上範圍為一菱形。如圖 8 所示。

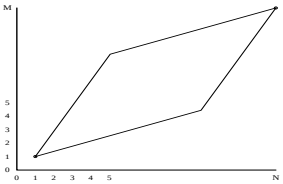

圖 8 全程路徑限制

在範圍內的某一點上考慮最佳路徑時，定義一個區域路徑限制(Local path constraint)，如圖 9，表示在點(n,m)時，可能到達這個點的路徑來自於(n-1,m)、(n-1,m-1)、(n,m-1)三個點，p、q、r 表示為三個點轉移到(n,m)的加權值。定義一個點之最佳累積距離計算公式如下：

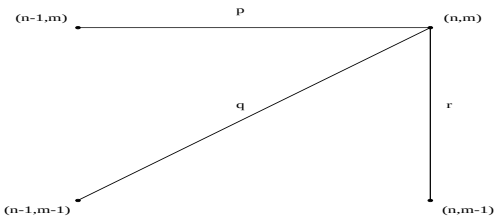

圖 9 區域路徑限制

$$D(n,m)=\min\begin{Bmatrix}D(n-1,m)+p\cdot d(\mathrm{n,m})\\ D(n-1,m-1)+q\cdot d(\mathrm{n,m})\\ D(n,m-1)+r\cdot d(\mathrm{n,m})\end{Bmatrix}$$

相異度(Distortion)比較法，衡量測試語音與參考語音之間相似程度。求取方法是將測試語音與參考語音的特徵值做兩者相減求取平方和之後再加總。其最小相異度參數辨識結果，如下所示：

$$dist[i]=\sum_{n=0}^{d}(x[n]-r[n])^2 \qquad ,1\le i\le 10$$

4.實驗驗證

驗證滑板車語音辨識應用設計，採用 GoldWave 軟體做語音樣本錄製，以單聲道 8000Hz 取樣頻率，取樣時間為 0.6 秒，即時聲音錄製，語音頻譜特徵值使用倫敦大學心理語言學實驗室的 SFSWIN 軟體做分析。[19]、[20]系統部分以 Images SI 公司語音訓練模組 SR-07 及語音介面模組 SRI-03，連接本實驗室自製 IO 板、馬達控制器與電動滑板車馬達驅動器，控制負載端直流馬達轉動。

語音控制為設定 4 個語音指令，分別為：語音指令「1」為全速運轉，語音指令「2」為加速/啟動，語音指令「3」為減速，語音指令「4」為停止。本文系統部分，將 SR-07 主板和 SRI-03 語音介面模組連接，再連接本實驗室設計 IO 板與馬達控制器(整合為馬達核心控制板)，圖 14 為馬達語音控制整合系統。再連接電動滑板車驅動器，達到語音控制效果。

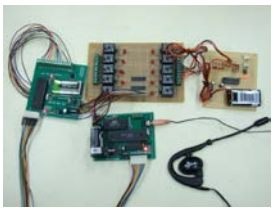

圖 14 語音控制整合系統

由於我們將調速手把油門訊號線連結到馬達控制器上面，調速控制改由語音指令控制。因此轉動電動滑板車調速手把，直流馬達不會轉動。首先開啟 SR-07 主板電源、SRI-03 語音介面模組電源與電動滑板車電源，系統通電後，使用麥克風做語音訊號輸入，HM2007 語音 IC 做語音訊號辨識，語音訊號傳至馬達控制器，再將語音指令下達給直流馬達做運轉。其直流馬達電壓波形量測，擷取馬達波形圖來探討電動滑板車直流馬達運轉情形。圖 15 為直流馬達從靜止狀態啟動的波形圖，縱軸為電壓刻度大小，一格刻度表示為 10v。顯示直流馬達電壓維持在 24v 左右。接著，透過語音指令「2」，直流馬達從啟動後加速一段波形。圖 16 為直流馬達啟動後第二段加速波形圖。接著語音指令「2」，第三段加速如圖 17。接著語音指令「2」，為第四段加速如圖 18。接著語音指令「2」，如圖 19 第五段加速波形圖。接著語音指令「2」，第六段加速如圖 20。

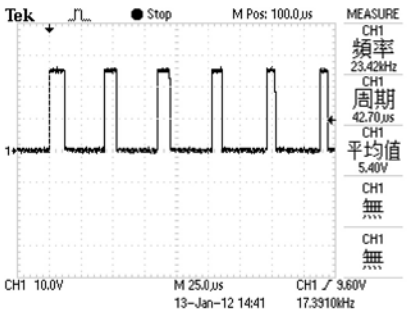

圖 15 馬達啟動波形

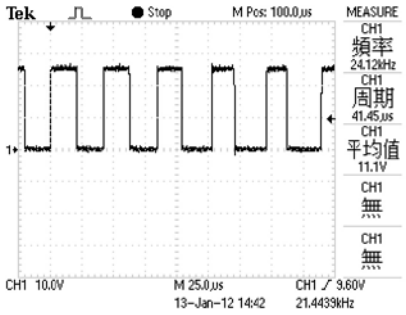

圖 16 馬達第二段加速波形

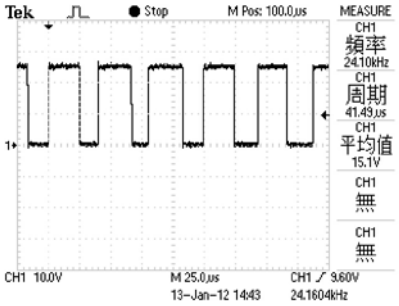

圖 17 馬達第三段加速波形

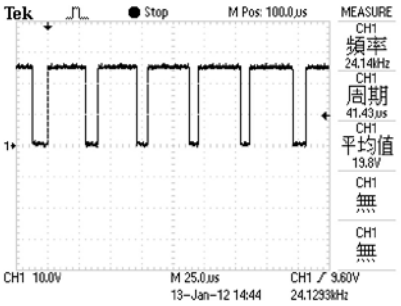

圖 18 馬達第四段加速波形

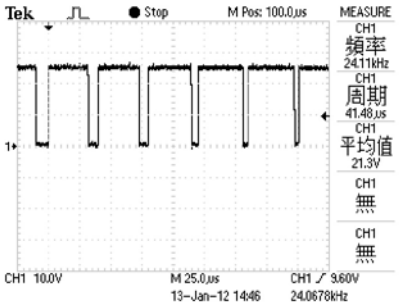

圖 19 馬達第五段加速波形

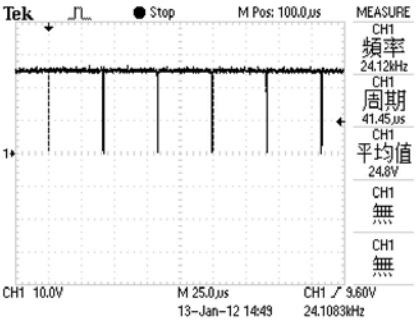

圖 20 馬達第六段加速波形(全速運轉波形)

第六段加速波形圖，接近語音指令「1」全速運轉波形圖。所以也為語音指令「1」波形圖。接著，我們再透過語音指令「3」減速訊號，驗證直流馬達是否能夠從全速轉動一段一段地減速到停止的情形。

語音指令「3」，第一段減速。圖 21 為直流馬達從全速轉動情形下第一段減速波形圖。第二段減速，圖 22。依此類推到圖 26。

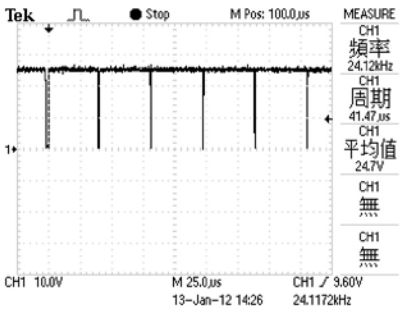

圖 21 馬達第一段減速波形

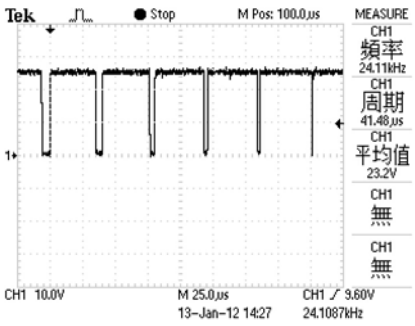

圖 22 馬達第二段減速波形

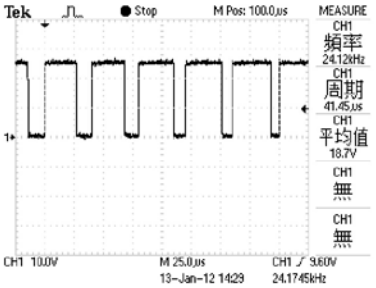

圖 23 馬達第三段減速波形

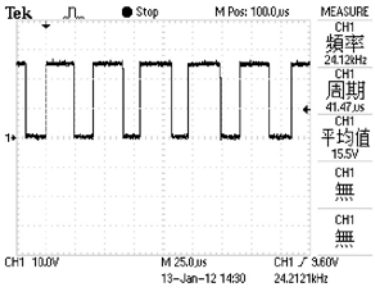

圖 24 馬達第四段減速波形

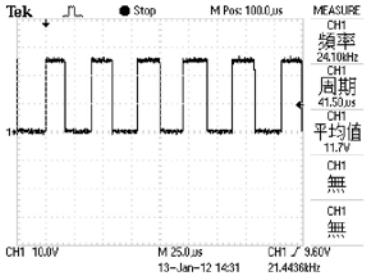

圖 25 馬達第五段減速波形

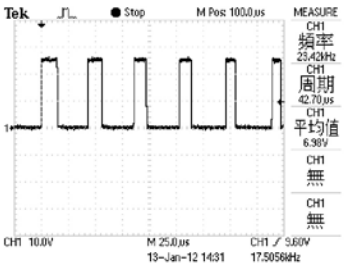

圖 26 馬達第六段減速波形(啟動波形)

接著語音指令「3」，第七段減速。圖 27 為直流馬達第七段減速波形圖。此階段馬達波形近似馬達停止波形。

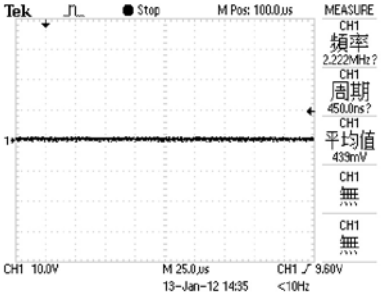

圖 27 馬達第七段減速波形(停止波形)

以上順序的結果，可以了解本文設計的直流馬達加速語音指令「2」，每一段的加速語音指令能夠讓馬達運轉速度每一段提升上去，到達馬達全速運轉；相對的，馬達減速語音指令「3」也能夠將速度每一段的減少下去直到馬達停止。

特定語者電動滑板車實驗

實驗人員：本人。

經由梅爾倒頻譜係數擷取特徵值如表 1。

表 1 實驗一的梅爾倒頻譜係數特徵值

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	-1.66	1.46	0.20	-0.18	-0.72	-0.36	0.005	-0.63	-0.24	0.001
2	0.41	-1.83	-3.93	-2.28	-1.69	-1.29	0.66	1.30	0.15	0.65
3	-2.49	0.54	-0.52	-0.15	-0.63	-0.22	-0.33	-0.29	0.17	0.38
4	-3.25	1.08	0.68	-1.07	-1.11	0.33	0.11	-0.50	0.32	0.31
測試	0.36	4.24	0.44	-0.58	-0.75	-0.45	-0.93	-0.11	0.46	0.17

計算方式：

採用最小相異度比較法：(數值越小辨識結果越接近)

$$dist[i]=\sum_{n=0}^{d}(x[n]-r[n])^2 \qquad ,1\le i\le 10$$

計算結果：

$$dist_{測試\text{-}1}=13.70 \qquad dist_{測試\text{-}2}=65.27 \qquad dist_{測試\text{-}3}=23.51 \qquad dist_{測試\text{-}4}=25.55$$

根據實驗數據，最小數值為 $dist_{測試\text{-}1}=13.70$，此為「1」基本語音辨識。測試語音為發音「1」語音訊號，語音辨識實驗結果正確。根據實驗數據，最小數值為 $dist_{測試\text{-}1}$=13.70，此為「1」基本語音辨識。測試語音為發音「1」語音訊號，語音辨識實驗結果正確。

5.結論

經過多次實驗驗證，本文電動滑板車語音辨識應用設計使用動態時軸校正分析，其特定語者之少量詞彙其辨識率可達 86%。此結果表示本文的設計語音訊號能夠近似正確地控制電動滑板車動力運轉，達到安全與便利的駕駛成效，進而可以成為中低收入肢體障礙人士便利駕駛的交通工具，本文電動滑板車語音辨識設計經過測試與實驗驗證，雖然特定語者辨識率可以達到 86%，但是系統建置實際上還未算建構完整，將來還有可以改良與修正的地方。例如：語音辨識方面，提高非特定語者辨識度、縮短語音辨識及時反應時間等等。實體系統架構方面，可以設計安全帽內建麥克風，以降低環境噪音的干擾。實驗室將來可以將語音辨識技術發展結合四輪電動車，或是增加語音訊號轉向控制，以提升駕駛的安全性與方便性。

6.參考文獻

[1] P.K.Sharma, B.R.Lakshmikantha, K.S.Sundar, “Real time control of DC motor drive using speech recognition”, *Power Electronics (IICPE) 2010 India International Conference on*, 28-30 Jan.2011, pp.1-5.

[2] Xuewen.Luo, Ing.Yann.Soon, Chai.Kiat.Yeo, “An auditory model for robust speech recognition”, *Audio Language and Image Processing (ICALIP) 2008 International Conference on*, 7-9 July.2008, pp.1105-1109.

[3] T.B.Amin, “Speech Recognition using Dynamic Time Wraping”, *Advances in Space Technologies 2nd International Conference on 2008*, pp.74-79.

[4] 郭自強，輕型電動車及其蓄電池，*第六屆全國輕型電動車技術研討會，中國電工技術學會*，2007。

[5] 孫立群，電動自行車維修從入門到精通，人民郵電出版社，2007。

[6] 廖任秀，陳桂蘭，基於 PROTEUS 電動滑板車控制器的設計，*金華職業技術學院學報*，2008，Vol.6。

[7] 陳家新，電動滑板車控制原理及設計中若干問題的研究，*電機電器技術*，2001，Vol.5，pp.38-42。

[8] 王小川，語音訊號處理，全華科技圖書公司，2005。

[9] 謝秀琴，數位語音訊號基本原理，全華科技圖書公司，1996。

[10] 黃啟祥，結合高斯混合及支撐向量機模型之語者確認研究，中央大學電機工程研究所碩士論文，2009。

[11] 林士翔，數據擬合與分群方法於強健語音特徵擷取，台灣師範大學資訊教育研究所碩士論文，2007。

[12] 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，臺灣師範大學資訊工程研究所碩士論文，2005。

[13] 古鴻炎，黃國勛，行動裝置上語音命令辨識系統之製作，*2008 民生電子研討會*，pp.741-746，2008。

[14] 張恆誌，使用動態時間校正演算法於國語數字語者辨識系統之研究，義守大學電子工程研究所碩士論文，2011。

[15] 李政益，特定語者特定中文語音指令雙模辨識技術，清雲科技大學電子工程研究所碩士論文，2005。

[16] Zhang Jing, Zhang Min, “Speech recognition system based improved DTW algorithm “, *CMCE 2010 International Conference on 24-26* Aug.2010, Vol.5, pp.320-323.

[17] 張成，蔣皓石，林嘉宇，基於 16 位單片機的語音電子門鎖系統，*電子技術應用*，2005，Vol.7，pp.22-25。