

國立中山大學電機工程學系
National Sun Yat-Sen University Department of Electrical Engineering

碩士論文

以類神經網路為架構之語音辨識系統

The Speech Recognition System using Neural Networks

**研究 生：陳松琳
指導教授：陳遵立 老師**

中華民國九十一年六月

致謝

承蒙恩師陳遵立老師兩年來悉心指導，於研究方向、治學態度與待人處世上多方啟發與訓示，對於研究所需之硬體設備亦毫無保留的投入經費，引進最新之儀器與技術，使得本論文得以順利完成，特此致上十二萬分之謝忱。並感謝交通大學吳永春老師、中科院高一智老師、清雲技術學院吳英秦老師以及微星科技黃金請副總經理，惠予寶貴意見與匡正，使本論文更臻至完美，學生銘感於心。

另外感謝在研究所期間，學長連強、耿魁、仁裕、國光、偉德、學姐詠宜以及同窗好友偉智、仁偉、富存、國棟、証賀、盈州、育和、睿余、平峽與義隆，彼此互相的砥礪與協助，使得學生才能度過每回的研究低潮期，順利走過兩年的研究生活。當然還有充滿活力的學弟芳易、嘉宏、億晉與凱文，讓實驗室平常的休閒活動更加的多彩多姿，可說是為實驗室帶來更活躍的生命力。

最後，僅將此論文獻給養育我、栽培我的父母親及我的家人；還有我摯愛的老婆淑理和岳父岳母，由於你們的支持、鼓勵和體諒，使我得以順利完成學業與論文，希望你們可以永遠健康與快樂，謝謝你們。

陳松琳 於西子灣

2002.7.5.

學年度：90.

學期：2.

校院：國立中山大學.

系所：電機工程學系研究所.

論文名稱(中)：以類神經網路為架構之語音辨識系統.

論文名稱(英)：The Speech Recognition System using Neural Networks.

學位類別：碩士.

語文別：Chi.

學號：8931605.

提要開放使用：是.

頁數：78.

研究生(中)姓：陳.

研究生(中)名：松琳.

研究生(英)姓：Chen.

研究生(英)名：Sung-Lin.

指導教授(中)姓名：陳遵立.

指導教授(英)姓名：Tzuen-Lih Chen.

關鍵字(中)①：語音辨識.

關鍵字(中)②：類神經網路.

關鍵字(中)③：倒傳遞演算法.

關鍵字(英)①：Speech Recognition.

關鍵字(英)②：Neural Network.

關鍵字(英)③：Backpropagation Algorithm.

中文摘要

本論文以倒傳遞類神經網路(BPNN : Backpropagation Neural Network)為架構設計一非特定語者之中文數字語音辨識系統，辨識率可達 95%。當此系統應用於特定語者時，經適應修正後，更可使系統之辨識率高於 99%。為能使系統轉移到數位處理器(DSP)平台，針對類神經網路模型，提出了神經元移除法則，利用此法則可減去約 $\frac{1}{3}$ 數量之神經元，降低系統 20% ~ 40% 的記憶體需求，且系統之辨識率仍可達 85%。在 BPNN 網路模型的輸出架構中，提出以二進位編碼之方式取代傳統一對一之架構，以增加系統可辨字彙之數量。對於語音訊號端點偵測，也提出另一有效之搜尋法則，不論雜訊干擾存在與否，不需複雜之運算，可有效定位出有聲段所在之處。

Abstract

This paper describes an isolated-word and speaker-independent Mandarin digit speech recognition system based on Backpropagation Neural Networks(BPNN). The recognition rate will achieve up to 95%. When the system was applied to a new user with adaptive modification method, the recognition rate will be higher than 99%. In order to implement the speech recognition system on Digital Signal Processors (DSP) we use a neuron-cancellation rule in accordance with BPNN. The system will cancel about 1/3 neurons and reduce 20% ~ 40% memory size under the rule. However, the recognition rate can still achieve up to 85%. For the output structure of the BPNN, we present a binary-code to supersede the one-to-one model. In addition, we use a new ideal about endpoint detection algorithm for the recoding signals. It can avoid disturbance without complex computations.

目 錄

第一章 緒論	
1.1 前言	1
1.2 研究背景	3
1.3 研究動機與目標	4
第二章 分析框處理	
2.1 簡介	7
2.2 最常用的兩種分析框	9
2.3 固定寬度與變動寬度之分析框	11
第三章 端點偵測演算法	
3.1 簡介	15
3.2 時域端點偵測法相關參數簡介	16
3.3 端點偵測法	21
第四章 特徵參數擷取	
4.1 簡介	27
4.2 前置強波處理	28
4.3 倒頻譜參數	29
4.4 頻譜參數與倒頻譜參數之比較	34
第五章 動態時間校準演算法	
5.1 簡介	35
5.2 動態時間校準法	36
5.3 語音樣板資料庫的建立	46
第六章 倒傳遞類神經網路模型	
6.1 簡介	47
6.2 多層感知機之架構	48
6.3 倒傳遞演算法	51
6.4 辨識系統訓練方法	55
6.5 神經元移除法則	56

第七章 實驗方法	
7.1 前言	57
7.2 語音樣本資料庫	57
7.3 實驗設計	59
第八章 實驗結果	
8.1 端點偵測演算法之比較結果	67
8.2 特徵參數的選定	68
8.3 學習調整率 a 對系統學習效能的影響	69
8.4 非特定語者之中文數字辨識系統	69
8.5 倒傳遞類神經網路輸出架構之設計	70
8.6 神經元移除法則的應用	70
8.7 辨識演算法 DTW 與 BPNN 之比較	72
8.8 非特定語者之系統應用於特定語者之影響	72
第九章 結論與展望	
9.1 結論	75
9.2 展望	76
參考文獻	77

表目錄

表 6-1 類神經網路符號與術語定義表	50
表 7-1 多媒體控制實驗室中文數字語音樣本資料庫	57
表 7-2 中文數字語音樣本頻譜特徵參數資料庫	58
表 7-3 中文數字語音樣本倒頻譜特徵參數資料庫	58
表 7-4 端點偵測法對照表	60
表 7-5 二進位編碼輸出型之中文數字對照表	63
表 8-1 DTW 與 BPNN 特性測試結果	72

圖目錄

圖 1-1	語音辨識系統方塊圖	2
圖 2-1	分析框處理過程示意圖	7
圖 2-2	分析框寬度大小對語音訊號分析的影響	8
圖 2-3	A.矩形框波形圖；B.矩形框頻譜圖；C.漢明框波形圖；D.漢明框頻譜圖	9
圖 2-4	A.待分析訊號；B.處理後之結果；C.矩形框；D.處理後之頻譜圖	10
圖 2-5	A.待分析訊號；B.處理後之結果；C.漢明框；D.處理後之頻譜圖	11
圖 2-6	固定寬度分析框之處理示意圖	12
圖 2-7	變動分析框重疊寬度之處理示意圖	13
圖 2-8	變動分析框寬度之處理示意圖	14
圖 3-1	非連續語音訊號之端點偵測	15
圖 3-2	分析框能量與越零率參數估算示意圖	16
圖 3-3	能量參數應用實例	18
圖 3-4	消除雜訊影響後之能量參數應用實例	18
圖 3-5	越零率參數示意圖	19
圖 3-6	A.B.為無雜訊干擾；C.D.為受雜訊干擾之越零率參數應用實例	20
圖 3-7	能量曲線判別法示意圖	21
圖 3-8	能量曲線判別法之應用實例(無雜訊干擾)	22
圖 3-9	能量曲線判別法之應用實例(雜訊干擾)	22
圖 3-10	R-S 判別法之示意圖	24
圖 3-11	R-S 判別法之之應用實例(無雜訊干擾)	25
圖 3-12	R-S 判別法之之應用實例(雜訊干擾)	25
圖 4-1	前置強波器之頻譜圖	28
圖 4-2	前置強波器之應用實例(右圖為處理後之結果)	28
圖 4-3	並聯式數位帶通濾波器組	29
圖 4-4	以線性刻度所設計之帶通濾波器	30
圖 4-5	以線性刻度濾波器之應用實例(LFS 與 LFCC)	31

圖 4-6	Bark 及 Mel 與實際頻率之關係圖	32
圖 4-7	以梅爾刻度所設計之帶通濾波器	33
圖 4-8	以梅爾刻度濾波器之應用實例(MFS 與 MFCC)	33
圖 4-9	四種特徵參數對辨統系統的影響	34
圖 5-1	動態時間校準法之語音辨識系統	35
圖 5-2	最佳途徑尋找問題圖例	37
圖 5-3	自節點 F 到節點 A 的最佳途徑選擇	38
圖 5-4	動態時間軸校準函數示意圖	40
圖 5-5	符合連續性之搜尋途徑	41
圖 5-6	符合單調性之搜尋途徑	42
圖 5-7	局部限制途徑	42
圖 5-8	加入局部限制條件之搜尋途徑	43
圖 5-9	最佳時間校準函數之有效區域	44
圖 5-10	穿過不同節點數目之校準函數途徑範例	45
圖 6-1	倒傳遞類神經網路之語音辨識系統	47
圖 6-2	神經元模型	48
圖 6-3	單層感知機網路模型	49
圖 6-4	兩層感知機(MLP)之網路架構	49
圖 7-1	一對一輸出型之類神經網路架構	63
圖 7-2	二進位編碼輸出型之類神經網路架構	63
圖 7-3	實驗測試方塊圖	64
圖 8-1	訊號未受雜訊干擾之端點偵測結果	67
圖 8-2	訊號受雜訊干擾之端點偵測結果(SNR=20dB)	68
圖 8-3	不同特徵參數對系統辨識效能的影響	68
圖 8-4	採用不同學習調整率所得之學習過程	69
圖 8-5	非特定語者之中文數字辨識系統響應圖	70
圖 8-6	一對一與二進位編碼之輸出架構性能比較圖	71
圖 8-7	移除次要神經元之學習曲線	71
圖 8-8	非特定語者系統應用於特定語者之系統響應圖	73



第一章 緒論

1.1 前言

語音是人類用來交換訊息最自然的工具，因此設計一套能夠瞭解人類說話內容的語音辨識系統，一直是研究人員的理想與目標。然而縱使相關產業技術一日千里，微處理器的處理速度以指數形態成長，電腦在處理語意(semantics)這類抽象觀念，還是力有所未逮。因此，目前語音辨識(speech recognition)方面的研究主要還是在將語音訊號轉換成對應的文字(或模型)，而很少對這些文字所組成的句子之含意做分析。語音辨識的結果可應用在不同的領域，譬如在讀寫機(dictation)的應用方面，辨識所得到的文字便會被顯示在文書處理器的文件上，此類的應用提供電腦中文輸入的解決方案。而在語音控制應用上，系統則可依據辨識的結果做出相對應的動作，實際的例子便是汽車行動電話以語音方式撥號，可避免駕駛人分神，維持行車安全。

目前的語音辨識系統不論針對那一種語言，所使用的基本演算法都是大同小異。主要的差別在於針對各種語言的特性，這些演算法必須做適當的組合或修正，才能構成一套具有高辨識率的系統。舉例而言，中文字的發音是由聲韻母及聲調所構成的，因此在中文語音辨識時，除了聲韻母辨識之外，還要加上聲調的判別。而聲韻母辨識與聲調判別在圖樣識別(pattern recognition)上是屬於相同的問題，因此所使用的演算法是一樣的。反觀以詞(word)為單位的英文，其聲調並不具有辨義功能，所以其語音辨識系統並不需要做聲調判別的處理。

語音辨識系統的分類若按照其能夠辨識的字彙數量大小可分為：少量(數百字以下)、中量(數千字)、及大量(數萬字以上)字彙語音辨識器。少量字彙語音辨識器主要應用在語音指令控制上。中量字彙語音辨識器可應用於資料輸入或查詢系統。大量字彙語音辨識器則常見於讀寫機應用上。

若按照辨識對象來分類，語音辨識系統可分為：特定語者(speaker-dependent)及非特定語者(speaker-independent)語音辨識器[1]。特定對象語者辨識器是針對特定的使用者所設計的系統，設計的要求只是提高對這些使用者的語音辨識率，對其他使用者而言，系統不保證能夠提供可接受程度的辨識率。另一方面，非特定語者語音辨識器則是針對一般使用者所設計的，由於非特定語者語音辨識器的訓練語音收集較為困難，其辨識率一般都不如特定語者語音辨識器。為了改善這個問題，非特定語者語音辨識器大多具有語者訓練(speaker adaptation)功能。也就是說，使用者第一次使用系統時，先依據系統的要求讀入一段語音，經由訊號的分析，計算該使用者的調適參數，並建立檔案，之後該使用者在操作系統時，可指定對應檔案使用個人的調適參數以提昇辨識率。

另外一種分類方法是按照使用方式來區分有：非連續語音，及連續語音辨識系統。非連續語音辨識器的使用，說話者必須在字與字間稍做停頓，系統藉由停頓訊號來決定每個字音訊號的範圍，並加以辨識。字與字之間的停頓固然可以簡化系統在訊號分段上的困難，然而卻會造成使用上的不便，失去了改善人機介面的原意，連續語音辨識系統的發展便是要解決這項缺失。使用者可以自然的說話方式，將語音訊號讀入連續語音辨識系統，系統再依據前後連接語音及相關資料(如字典)來判斷說話的內容。因為連續語音辨識系統所要處理的可能情況比較複雜，所需要的運算量當然比非連續語音辨識系統要多出許多，而辨識率一般也較差。因此在這個領域中還有許多研究尚待進行。

語音辨識的困難除了每個人說話隨著其性別、年齡、地域等因素有其特定的發音方式之外，同一位使用者在不同心理或生理狀況所產生的語音訊號也會有所差異[2]。因此，直接使用比對的方式，所得到的辨識成功率是非常有限的，除此之外，語音訊號的變異性並非語音辨識上唯一的難題，另外如背景雜訊的影響，傳播媒介(如電話線路)的品質等等，都會嚴重影響到語音辨識的結果，由於這些問題，目前語音辨識的商業產品大部分還是以辦公室為主要操作環境。

早期語音辨識系統的辨識核心，都採用動態時間校準(Dynamic Time Warping : DTW)[3][4]的辨識方法，在訊號比對時便考慮到說話速度快慢的差異性，做適當的補償。此外便是類神經網路(Artificial Neural Network : ANN)[5]在語音辨識系統中的應用，逐漸的取代了傳統的動態時間校準法，且其相關的應用相當的多；直到隱藏式馬可夫模型(Hidden Markov Model : HMM)[6]的出現，因其採用統計方式來描述語音的特徵，經過多年的發展隨即成了語音辨識系統的主流，尤其是連續語音辨識系統幾乎是 HMM 的天下。然而在實際的日常電子用品上的應用，複雜的 HMM 却又遠不如 DTW 與 ANN 來的合適。

一般語音辨識系統的架構，如圖 1-1 所示。它的組成包括：聲音訊號的錄音、有聲段訊號的端點偵測、高頻訊號的強波處理、語音訊號的特徵參數擷取以及語音辨識系統的核心：圖樣辨識，最後即可獲得辨識之結果。所有語音辨識系統的處理流程皆可區分為此六大步驟，而端點偵測、強波處理與特徵參數擷取的過程中，還包括分析框處理。

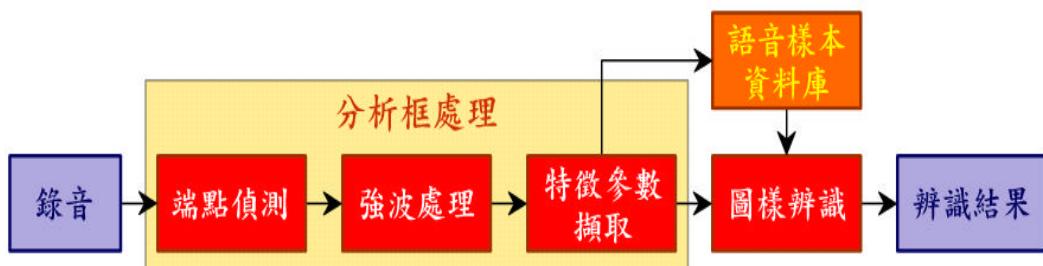


圖 1-1 語音辨識系統方塊圖

1.2 研究背景

1985 以前，語音辨識系統的研究，仍以理論研究為導向，直到個人電腦的快速發展，語音辨識理論才找到發揮的舞台，於是慢慢的將所有的語音辨識理論實際的應用到人們的日常生活中，為人們帶來更多的便利性，使得語音訊號處理領域，成了目前最為熱絡的研究學科。

科技的演進，使得電腦成了人們不可或缺的工具，為了讓電腦或電子產品操作更便利，於是學界與業界無不投入心力在操作界面的設計與改良上，期能設計出更人性化的電子產品。而最令人們感到簡便的操作方式，便是語音操控界面；從國外幾家電腦通信相關的大型企業(如：AT&T 及 IBM 等)每年招聘的語音研究人員數目，不難看出語音辨識的研究在未來幾年將扮演著很重要的角色，微軟(Microsoft)也曾對外公佈，語音操作界面將會是視窗系統將來的主流，並投以大量人力和資源專門從事此一研究；而電子大廠菲利普(Philip)更看好此一技術未來將為人們帶來革命性的演進，成立專屬研究部門，期能在此一領域中拔得頭籌。

目前語音辨識系統的實際應用，包括：語音操控玩具、語音撥號電話、語音密碼鎖、語音電腦操作界面、語音輸入法、語音查詢系統、語音訂票系統等等不勝枚舉。雖然語音辨識系統的應用相當的廣泛，且理論的發展也很完整，但仍有很多地方需要人們不斷的努力和研究，如：簡化辨識理論模型、增加辨識速度、提高系統辨識率、降低系統硬體上的需求。設計一套能夠瞭解人類說話內容的語音辨識系統，一直是研究人員最終的理想與目標。

可以預期語音辨識應用的範圍，會隨著相關技術的成熟而越來越廣泛。然而，也要瞭解距離研究人員最終的理想：非特定語者、連續音、無限量字彙語音辨識器的出現，其間還有許多尚待克服的困難；也期待國內有更多更聰明、年輕的學子，能夠一同加入這個尖端且實用的研究領域。

1.3 研究動機與目標

在語音辨識系統的領域中，主要的研究和應用仍以個人電腦平台為主流；直至近年，由於數位訊號處理器(Digital Signal Processor : DSP)的快速發展，為語音辨識系統找到了另一個發揮的舞台，此後才逐漸應用到隨身型電子用品上，如：語音撥號手機、個人行動數位助理(PDA)以及最新的產品電子書，皆已可窺見語音辨識系統的蹤跡。

數位訊號處理器(DSP)的應用，在實驗室已行之多年，特別是在電力電子領域中，如：馬達伺服控制系統、智慧型電池充電器以及並聯式不斷電系統中，已有多項成功之應用實例與經驗。除電力電子領域外，近兩年來實驗室的另一發展主軸即是多媒體控制的研究，包括語音訊號處理與影像訊號處理兩大領域，期以實驗室現有之技術與經驗，能在此兩大研究主題中開拓出實驗室新的應用主題。

而語音辨識系統的研究，即是實驗室在語音訊號處理領域中，初步規劃之探討主題與研究方向；同時分五個階段目標來進行：

- 1.資料之搜集與理論之研究：搜集語音辨識系統相關之資料，包括相關的演算法、實際的應用實例以及目前在此領域中的研究方向與發展趨勢等相關資料。在此階段另一重要工作，便是辨識演算法的研究與比較，進而找出適用於 DSP 上應用之辨識演算法則。
- 2.發展軟體的設計與語音樣本的收集：在進行理論的驗證時最需要的便是一套有效且方便的發展軟體，用以分析語音訊號的特性與進行相關參數的資料分析。而語音辨識系統中，另一個關鍵且必要的就是語音樣本的收集，錄製一套適用於實驗室所需之語音樣本，目前已完成之語音樣本資料庫為，21人(12男9女)的中文數字語音資料庫。
- 3.辨識演算法的驗證與系統性能評估：對於第 1 階段中所整理之辨識演算法，進行更進一步的理論驗證，並對於實現之系統進行性能的評估，包括：辨識速度、辨識率以及硬體需求(如記憶體)等等，期能分析出更適用於 DSP 上發展之辨識架構，以做為將系統從 PC 平台轉移到 DSP 平台時之參考依據。
- 4.DSP 平台語音辨識系統的開發：將第 3 階段所得之適用於 DSP 平台之語音辨識系統架構，嘗試轉移到 DSP 平台上發展，並評估其實際轉移後之系統性能優劣。
- 5.語音辨識系統的實際應用：克服了第 4 階段之平台轉移工作後，即可將語音辨識系統做實際的應用，如語音撥號器或聲控電子狗等。

為了能明確的評估辨識系統功能，故在本論文中，設計一非特定語者之中文數字語音辨識系統為主要研究題目，其次為將非特定語者之語音辨識系統，應用於特定語者的使用中，並加入適當的樣本重新訓練過程，進而提高系統的辨識率。此語音辨識系統架構適用於任何指令型態之語音辨識系統，且對於特定對象之使用者，可獲得更高的辨識率。

目前實驗室的進度已完成第 3 階段的目標，即將著手進行第 4 階段 DSP 平台的發展工作，使實驗室能按當初之規劃目標逐一往前邁進，假以時日，定能整合到實驗室智慧型電子狗的實際應用中。

本論文中共分為九章，詳細說明整個語音辨識系統中設計時所需之基本知識與應用理論，其安排如下：第一章為緒論；第二章為分析框處理；第三章為端點偵測演算法；第四章為特徵參數擷取；第五章為動態時間校準演算法；第六章為倒傳遞類神經網路；第七章為實驗方法；第八章為實驗結果；第九章則為結論與展望。

第二章 分析框處理

2.1 簡介

當人們說話時，發音器官的位置及形狀會隨時間而改變，以發出各種不同的聲音，故發音系統為時變系統(time-varying system)；而其所產生的語音訊號則是屬非恆態訊號(non-stationary signal)，即訊號的統計特性是時變的。然而發音器官的位置及形狀在短暫時間內(約 20ms)不會有太大的變化，因此在分析語音訊號特性時，可將語音訊號分段處理，在每一時間區段內假設發音系統為一非時變系統(time-invariant system)；換言之，假設語音訊號為一種區段性恆態訊號(piece-wise stationary signal)。所謂框分析[7]，便是利用上述之發音特性，所採用的一種分段訊號處理方式，以獲取更有效、更方便的語音訊號表示法。

假設在時間軸上有一固定寬度的分析框存在，每次只針對分析框內的訊號區段加以處理，在做完處理後便將分析框移至下一個時間點，重覆相同的步驟直到訊號結束；再將每一分析框所得之參數值，按時間先後順序排列，即可觀察語音訊號中，該參數隨時間變化的特性，如圖 2-1 為語音訊號經框分析後所得的能量參數變化圖。

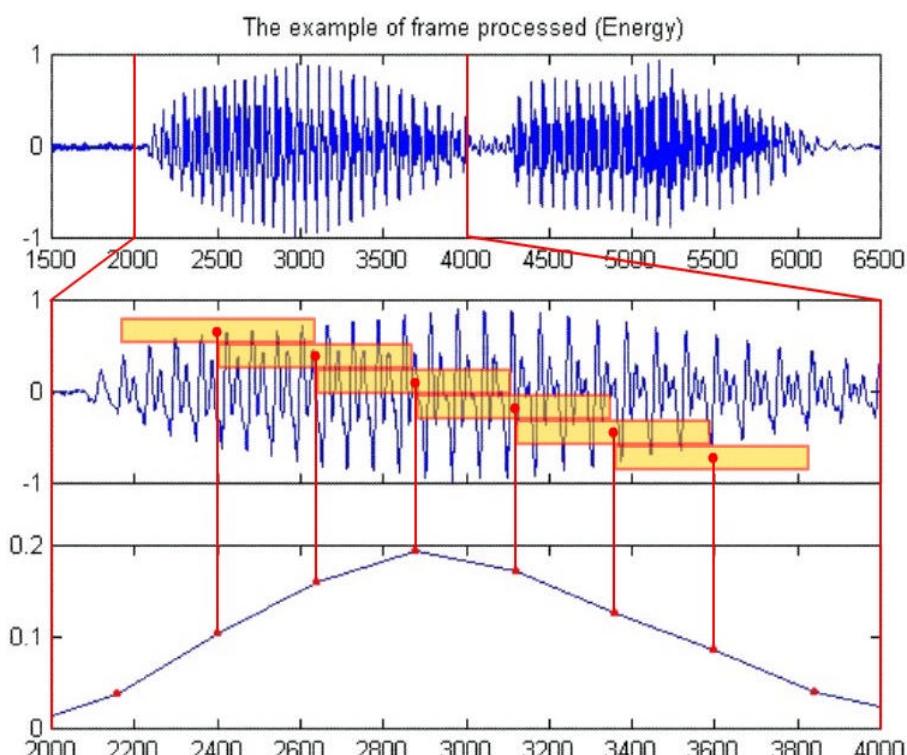


圖 2-1 分析框處理過程示意圖

為了要使處理之結果，能夠更清楚地呈現出語音訊號中，某項參數特性隨時間變化的情形，將分析框與分析框之間，做了重疊(Overlap)的處理。至於兩相鄰分析框的重疊多寡，一般多選擇分析框寬度的 $1/3$ 到 $1/2$ 之間(如圖 2-1 所示)。

在框分析處理中，分析框寬度的大小會影響分析的結果。當所使用的分析框寬度愈大，其所需的計算量會相對的減少，但分析框所得到的參數值差別也會變小，使得不易觀察到語音訊號隨時間變化的特性(如圖 2-2 中之 B 小圖所示)。然而，寬度太小的分析框所得到的結果，會因為所使用的取樣點數很少，容易受訊號突然變化的影響，較不具代表性且會使系統的計算量變大(如圖 2-2 中之 D 小圖所示)。故一般常用的分析框寬度約在 20ms 到 30ms 之間，此乃根據語音訊號特性所定之範圍。以 20ms 長的分析框及取樣頻率為 8 KHz 的系統為例，該分析框的寬度相當於 160 個取樣點(如圖 2-2 中之 C 小圖所示)。

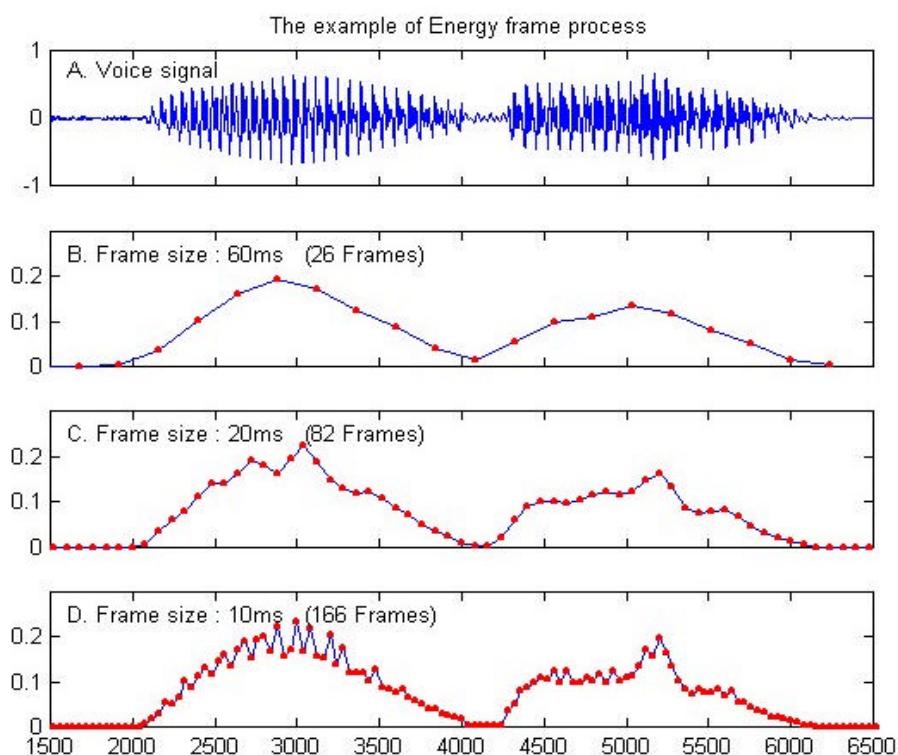


圖 2-2 分析框寬度大小對語音訊號分析的影響

2.2 最常用的兩種分析框：

就數學的觀點而言，所謂框分析相當於將訊號乘上分析框函數所得之結果；也就是從訊號中，擷取一小段訊號，並對於所擷取的訊號中之每個取樣點，給予不同的加權值。在此將介紹最常用的兩種分析框(如圖 2-3 所示)：矩形框(rectangular window)與漢明框(hamming window)。

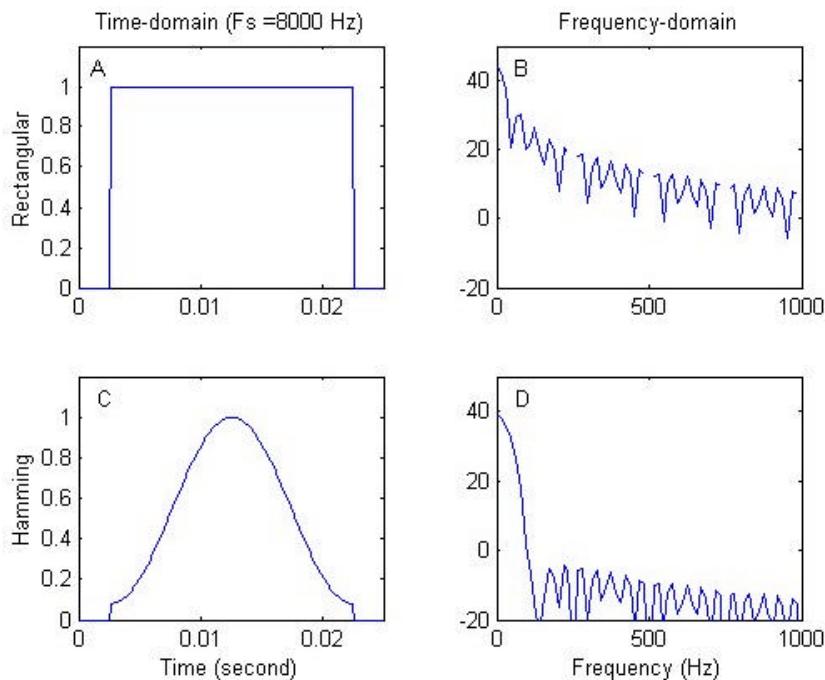


圖 2-3 A.矩形框波形圖；B.矩形框頻譜圖；C.漢明框波形圖；D.漢明框頻譜圖

2.2.1.矩形框(Rectangular window)

$$w_R(n) = \begin{cases} 1, & 0 \leq n \leq N-1 \\ 0, & \text{Otherwise} \end{cases} \quad (2-1)$$

(2-1)式為矩形框的數學函數，其中 N 為分析框的長度(為奇數)；圖 2-3 中之 A、B 小圖分別是此分析框的波形圖與頻譜圖。若一訊號經此分析框處理，則其結果可表示成：

$$\hat{f}(n) = w_R(n) f(n) \quad (2-2)$$

圖 2-4 為一 500Hz 之正弦波訊號，經 20ms 寬之矩形框處理後所得結果的波形圖與頻譜圖。此分析框在語音辨識系統中，僅於語音端點偵測時適用；在語音特徵擷取時，皆採用漢明框來做處理。

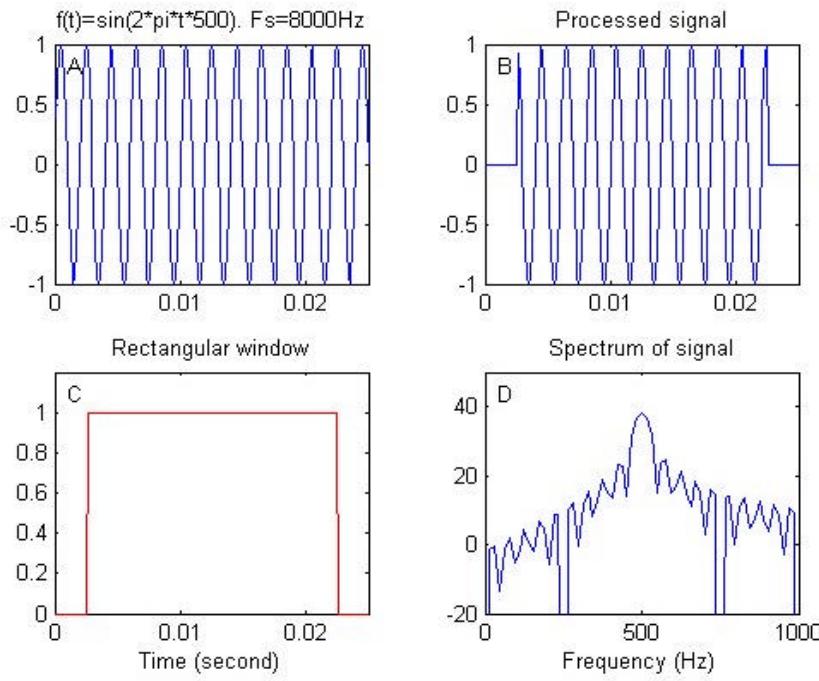


圖 2-4 A.待分析訊號；B.處理後之結果；C.矩形框；D.處理後之頻譜圖

2.2.2.漢明框(Hamming window)

$$w_H(n) = \begin{cases} 0.54 - 0.46\cos(2n\pi/(N-1)), & 0 \leq n \leq N-1 \\ 0, & \text{Otherwise} \end{cases} \quad (2-3)$$

(2-3)式為漢明框的數學函數，其中 N 為分析框的長度(為奇數)；圖 2-3 中之 C、D 小圖分別是此分析框的波形圖與頻譜圖。該函數在分析框中間份有較大的值，而兩端的值則非常接近零。若一訊號經此分析框處理，則其結果可表示成：

$$\hat{f}(n) = w_H(n)f(n) \quad (2-4)$$

圖 2-5 為一 500Hz 之正弦波訊號，經 20ms 寬之漢明框處理後所得結果的波形圖與頻譜圖。此分析框相當於對訊號區段中間部份的取樣值給予較大的加權(weighting)，而訊號區段內兩端的取樣值則較不易影響框分析的結果；這種加權方式再配合相鄰分析框的重疊設計，可使框分析後所得到的結果具有平滑的效果。

比較圖 2-3 中兩分析框的頻譜圖，可知漢明框函數頻譜的主頻帶(main lobe)寬度較寬，且邊緣頻帶(side lobe)被衰減較大，此兩項特性便是數位訊號處理中所謂的視窗效應(windowing effect)。在語音特徵擷取時，採用漢明框而不使用簡單且不須額外計算量的矩形框來做處理，也正是因為此一緣故。

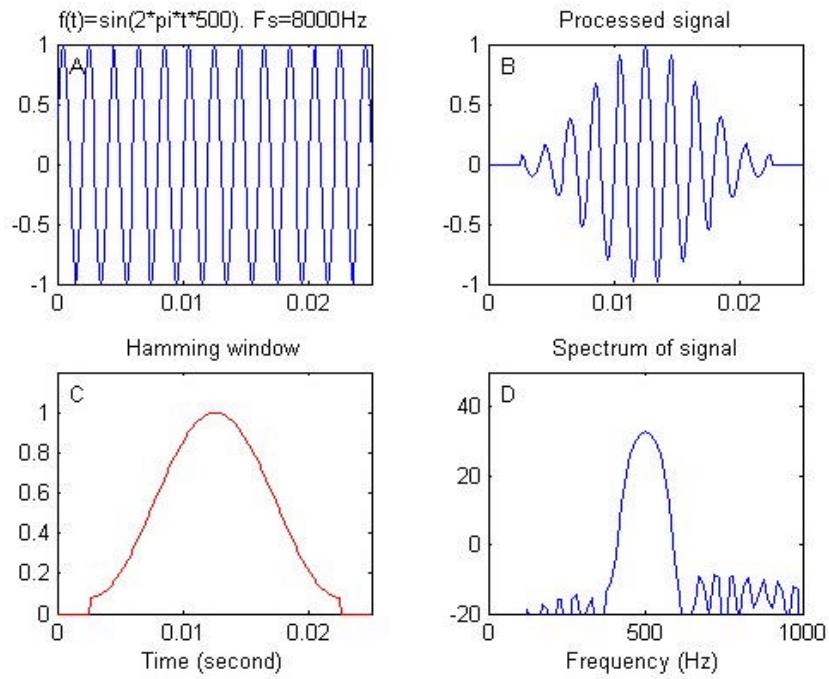


圖 2-5 A.待分析訊號；B.處理後之結果；C.漢明框；D.處理後之頻譜圖

2.3 固定寬度與變動寬度之分析框：

在實際的語音訊號處理中，相同的字音由於說話速度快慢不一，使得所錄得之語音訊號的長短也很少長度剛好一樣，如此一來在做語音特徵擷取時，便會得到不同數量的分析框，然而有時受限於系統架構的關係[8]，不論錄得之語音訊號長短為何，都只能取得固定數量的分析框，以供系統進行更進一步的處理(訓練或辨識)，所以若遇到此情形時，在框分析處理上就得做適當的修正。

在框分析時，若採用固定寬度分析框(fixed-size frame)處理，便會因訊號長度不同，而得到不同數量之分析框，此結果適用於輸入量不需固定之辨識架構，如動態時間軸校準法 DTW；對於輸入向量固定的類神經網路 ANN，便無法使用此類分析框，而須改用可以得到固定分析框數量的方法，也就是變動寬度分析框(dynamic-size frame)，以及變動重疊寬度法(dynamic-size overlap)。以下便針對此三種分析框處理方法，藉由範例做詳細的說明。

2.3.1. 固定寬度分析框(Fixed-size frame)

不論錄音訊號的長短為何，將分析框的寬度固定，重疊比率也固定，再逐一的去計算每個分析框的參數值，如此一來，當訊號長短不同時，所得到的分析框數量也不會相同。其數學關係式可表示成：

$$N_F = \text{Fix} \left[\frac{l_s - l_F}{l_F(1 - R_o)} \right] + 1 \quad (2-5)$$

其中， N_F 為分析框數量， l_s 為訊號長度(取樣點數)， l_F 為分析框寬度， R_o 為分析框重疊比率(%)，而 $\text{Fix}(x)$ 為數學函數，其運算結果為小於 x 的最大整數值。

固定寬度分析框的應用，多用於語音訊號的前置處理(如端點偵測)，或 DTW 所需的特徵向量或樣版(feature vector or pattern)。如圖 2-6 所示，為取樣頻率 8 kHz 且長短不同之語音訊號，經固定寬度分析框 40ms(320 sampled)，重疊比率為 50%(160 sampled)處理後所得之結果。從圖中可知，2-6-A 圖由於訊號較長，所以分析後可得到 14 個分析框參數；而 2-6-C 圖因訊號較短，僅得到 10 個分析框參數。在語音訊號處理上，若用到固定寬度分析框時，建議採用 20ms 到 30ms 寬的分析框為佳。

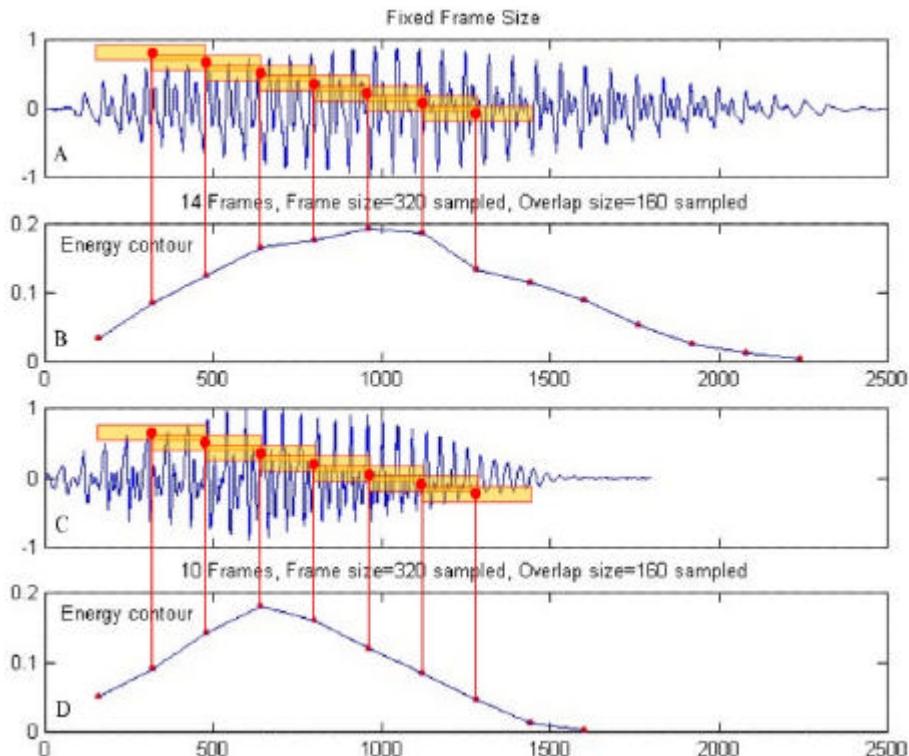


圖 2-6 固定寬度分析框之處理示意圖

2.3.2. 變動重疊寬度分析框(Dynamic-size overlap)

不論訊號長短，此法皆可得到所指定之固定數量的分析框參數值。將分析框的寬度固定，根據訊號長度與所指定之分析框數量，計算出重疊比率，再依此數據逐一的去計算每個分析框的參數值，如此一來，當訊號長短不同時，所得到的分析框數量仍會是一樣的。其數學式可表式成：

$$R_o = \frac{1}{l_F} \left\{ l_s - \text{Fix}\left[\frac{l_s - l_F}{(N_F - 1)}\right] \right\} \quad (2-6)$$

其中， R_o 為分析框重疊比率(%)， l_F 為分析框寬度， l_s 為訊號長度(取樣點數)， N_F 為分析框數量，而 $\text{Fix}(x)$ 為數學函數，其運算結果為小於 x 的最大整數值。

如圖 2-7 所示，為取樣頻率 8 kHz 且長短不同之語音訊號，經固定寬度之分析框 40ms(320 sampled)，重疊比率為變動的方式，處理後所得之結果。從圖中可知，兩訊號雖然訊號長短不同，但都可以得到相同數量(15 Frame)的分析框數值，2-7-A 圖所採用的重疊比率為 33%(79/240)，而 2-7-C 圖則為 54%(129/240)。

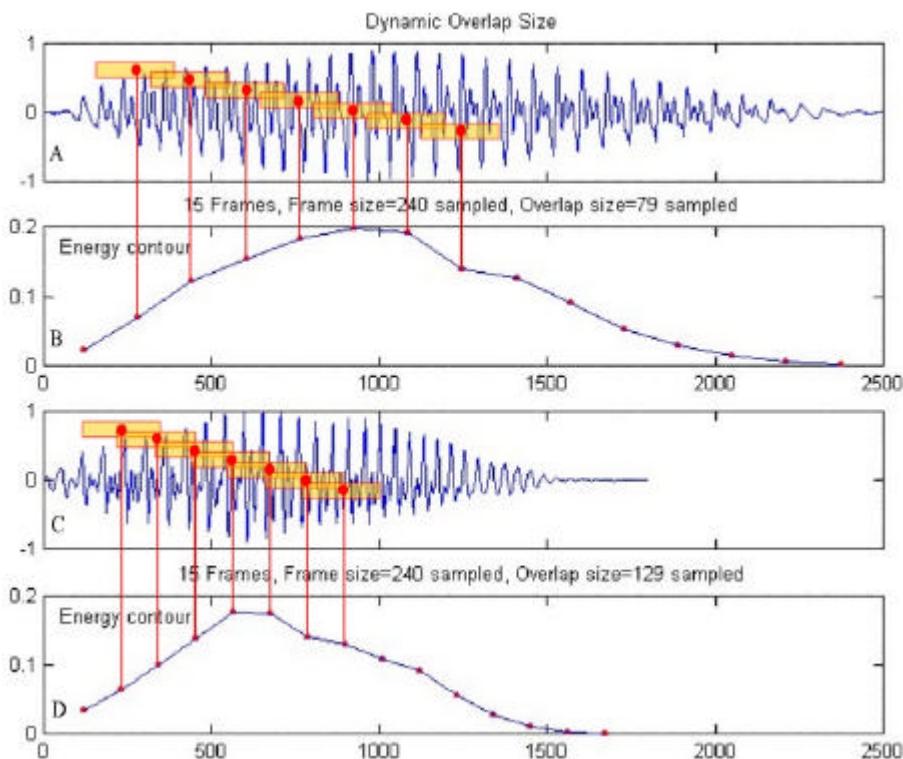


圖 2-7 變動分析框重疊寬度之處理示意圖

2.3.3. 變動寬度分析框(Dynamic-size frame)

此法亦可得到所須之固定數量的分析框參數值。指定分析框重疊比率，根據訊號長度與所指定之分析框數量，計算出分析框寬度大小，再依此數據逐一的去計算每個分析框的參數值，如此一來，當訊號長短不同時，所得到的分析框數量仍會是一樣的。其數學式可表示成：

$$l_F = \text{Fix}\left[\frac{l_s}{(N_F - 1)(1 - R_o) + 1}\right] \quad (2-7)$$

其中， l_F 為分析框寬度， l_s 為訊號長度(取樣點數)， N_F 為分析框數量， R_o 為分析框重疊比率(%)，而 Fix(x)為數學函數，其運算結果為小於 x 的最大整數值。

如圖 2-8 所示，為取樣頻率 8 kHz 且長短不同之語音訊號，經重疊比率固定為 50%，而分析框寬度變動的情況下，處理後所得之結果。從圖中可知，兩訊號雖然訊號長短不同，但都可以得到相同數量(15 Frame)的分析框數值，2-8-A 圖所採用的分析框寬度為 39ms，而 2-8-C 圖則為 28ms。

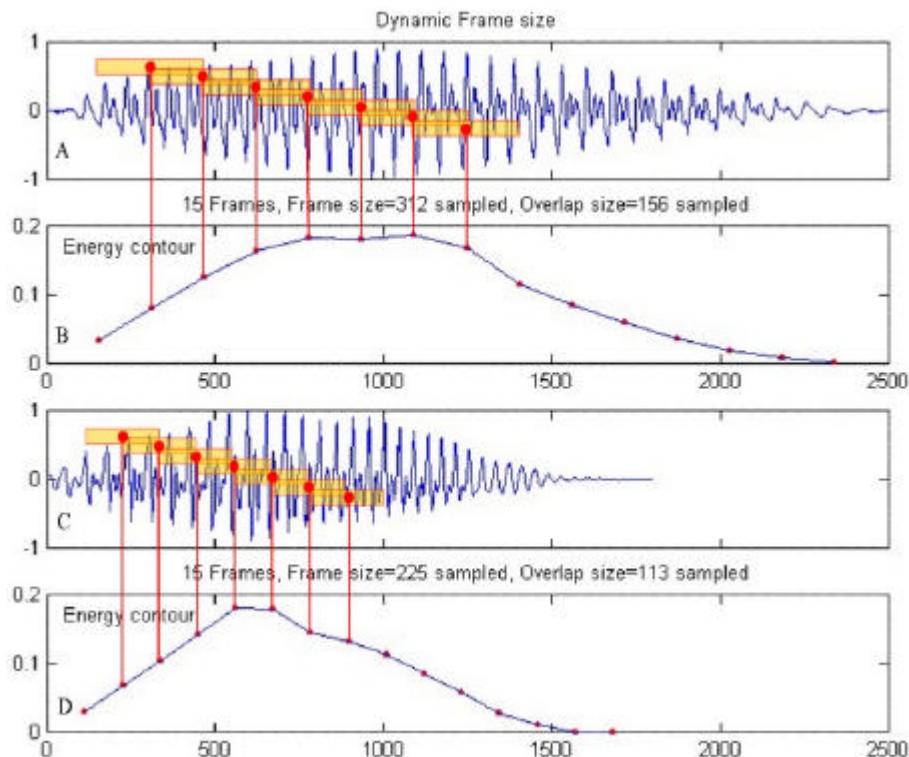


圖 2-8 變動分析框寬度之處理示意圖

第三章 端點偵測演算法

3.1 簡介

在非連續字音辨識(Isolated Word Recognition : IWR)系統中，語音訊號必須先經處理，以判斷訊號中那些區段是有聲段(speech segment)，那些是屬於無聲段(silence segment)或背景雜訊，接著再針對有聲段做更進一步的處理，此過程即稱之為語音訊號端點偵測(End Point Detection)；如圖 3-1 中所示，在兩垂直線之間即為有效之語音訊號。在理想的錄音環境下，即沒雜訊的干擾，語音訊號要做精確的端點偵測是件很簡單的事情，然而大部份的情況並非如此的理想，使得端點偵測的精確度變的相當的低！

語音訊號端點偵測的主要目的，如上所述是要從所錄下的訊號中，將所需或有效的聲音訊號部份擷取出來(即移除靜音段或背景雜訊)，以做更進一步的處理。在非連續字音辨識系統中，最重要的前置處理工作，便是利用「最精確」的端點偵測法，擷取訊號中有效的語音訊號，以建立有效的語音樣本；而所謂「最精確」的端點偵測，其定義為：該端點偵測擷取之訊號所提供的樣本，可以使系統辨識精確度達到最高的情況下，即稱之為「最精確」的端點偵測法。

語音訊號端點偵測的演算法有很多種，根據其判斷時所採用的參數，大致上可區分為三大類型：(1)時域端點偵測法；(2)頻域端點偵測法；(3)混合參數端點偵測法等，其中時域端點偵測法是最簡單也最常被應用的一種方法，但其最大的缺點是對雜訊的免疫力較低；而頻域端點偵測法以及混合參數端點偵測法，兩者的精確度較高，對抗雜訊的能力也較強，唯其所需的計算量較煩雜；故在此我們僅針對時域端點偵測法做詳細的說明。

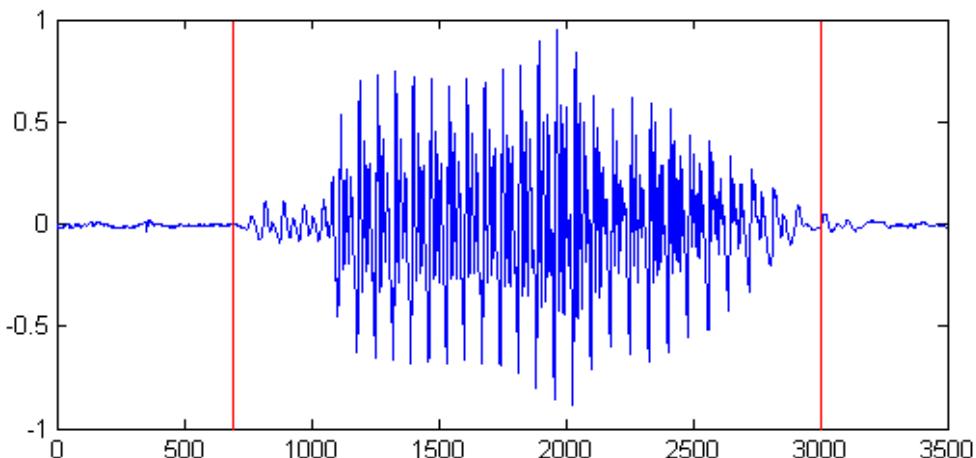


圖 3-1 非連續語音訊號之端點偵測

3.2 時域端點偵測法相關參數簡介

在時域端點偵測法中，因為語音訊號的能量一般都較背景雜訊大，所以偵測有效語音訊號段最直接的方法，便是依據訊號區段的能量大小，來做為判斷的依據。然而，有些字音在字首或字尾有子音或摩擦音的存在，因其能量太小不易被偵測出來，如此一來便會影響到系統的辨識率，因此在端點偵測時，除了依據能量參數外，也常使用訊號的越零率參數，以找出更精確的語音訊號的端點。

在背景雜訊不是很大的錄音環境下所錄得的語音訊號，利用能量和越零率兩個參數的配合，即可有效的找到語音訊號的端點；然而越零率參數最大的缺點，便是容易受雜訊的干擾，所以若想獲得較強健性的端點偵測演算法，便得採用頻域端點偵測法[9]，利用梅爾頻譜(Mel Frequency Spectral)能量參數，此參數即使在背景雜訊環境中所錄得之語音訊號，亦可有效的找出語音訊號的端點，當然其所付出的代價便是需要額外的轉換，以及龐大的計算量，故在此不再多做說明。

在時域端點偵測法中，用以做為端點偵測判斷依據的參數為：能量參數(Energy)以及越零率參數(Zero-Crossing Rate : ZCR)[10]。可藉由這兩個參數之曲線圖，再配合適當的臨界值之設定，即可找出錄音訊號中，有聲段訊號的端點所在之處。

如第二章所述，在做語音訊號分析時，為了能有效的求出語音訊號的特性，所以需要做框分析處理，接著再從每個分析框中求出所需要的參數值(如圖 3-2 所示)，然而在端點偵測過程中，分析框的重疊計算是沒有必要的(此做法在第三節中將會證明)，而且僅需採用矩形分析框(分析框寬度：10ms ~ 30ms 為宜)即可，如此便可獲得所需的資訊以及減少很多不必要的計算量。以下便針對此兩個參數如何應用做詳細的說明。

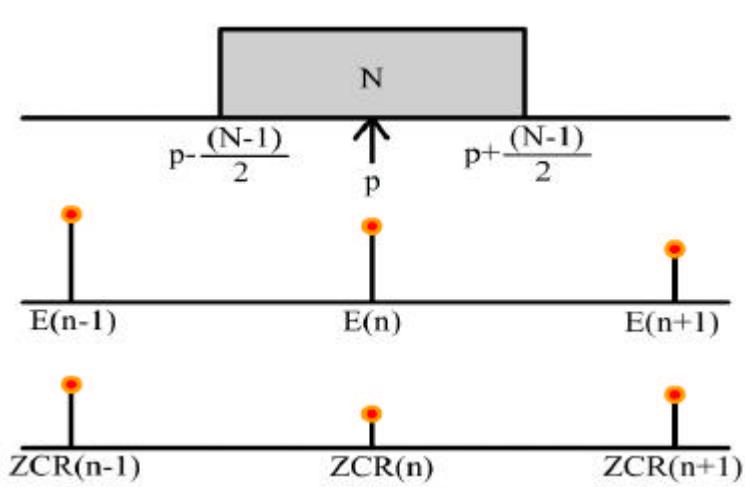


圖 3-2 分析框能量與越零率參數估算示意圖

3.2.1 能量參數(Energy)

A. 平方和參數：以某一分析框內訊號取樣值的平方之和，來做為該分析框的能量參數估算值；其數學式可表示成：

$$E(n) = \sum_{i=-(N-1)/2}^{(N-1)/2} [s(p+i)]^2 \quad (3-1)$$

其中， $E(n)$ 為第 n 個分析框的能量估算值， p 為該分析框的中心點在錄音訊號中的位置， N 為分析框寬度；因為採用矩形分析框，所以矩形框函數在式中予以省略，請參考圖 3-2 之示意圖。圖 3-3 中的 A2 與 B2 為利用此能量估算法的應用實例。

B. 均方和參數：由於平方和參數值，會隨著分析框寬度大小的不同而有所差異，為了減去分析框寬度的影響，所以一般都會做平均化(或稱之為正規化)的處理，即：

$$E(n) = \frac{1}{N} \sum_{i=-(N-1)/2}^{(N-1)/2} [s(p+i)]^2 \quad (3-2)$$

圖 3-3 中的 A3 與 B3 為利用此能量估算法的應用實例，分別與 A2 與 B2 比較，相當於各別對 A2 與 B2 做正規化處理(Normalization)後所得之結果。

C. 絶對值和參數：由於前面介紹的兩種能量估算方法都須要做平方的計算，為了減少這些計算量，有時也會用分析框中訊號的振幅(即取絕對值)之和來做為能量估算值，其數學式可表示成：

$$E(n) = \frac{1}{N} \sum_{i=-(N-1)/2}^{(N-1)/2} |s(p+i)| \quad (3-3)$$

圖 3-3 中的 A4 與 B4 為利用此能量估算法的應用實例；式中 $(1/N)$ 為正規化係數，在實際的應用中為了減少計算量，有時會不考慮此係數。

從圖 3-3 中可以得知，對於沒有背景雜訊的語音訊號，三種能量估算法所得的結果是沒有多大的差異(見圖 3-3 中 A2 A4)，但就計算量而言，採用絕對值和的方式較佔優勢；然而實際的錄音環境是無法避免雜訊的干擾，所以在這種情況下，就得採用平方和或均方和的方法，才能有效的將有聲段與背景雜訊區隔開來，其主要原因仍因為平方的結果，會使得訊號雜訊比(SNR)獲得平方的效果所致(見圖 3-3 中 B2 B4)，若採用絕對值和的方法，則語音訊號處理過程對雜訊干擾的免疫力就會變的很低。

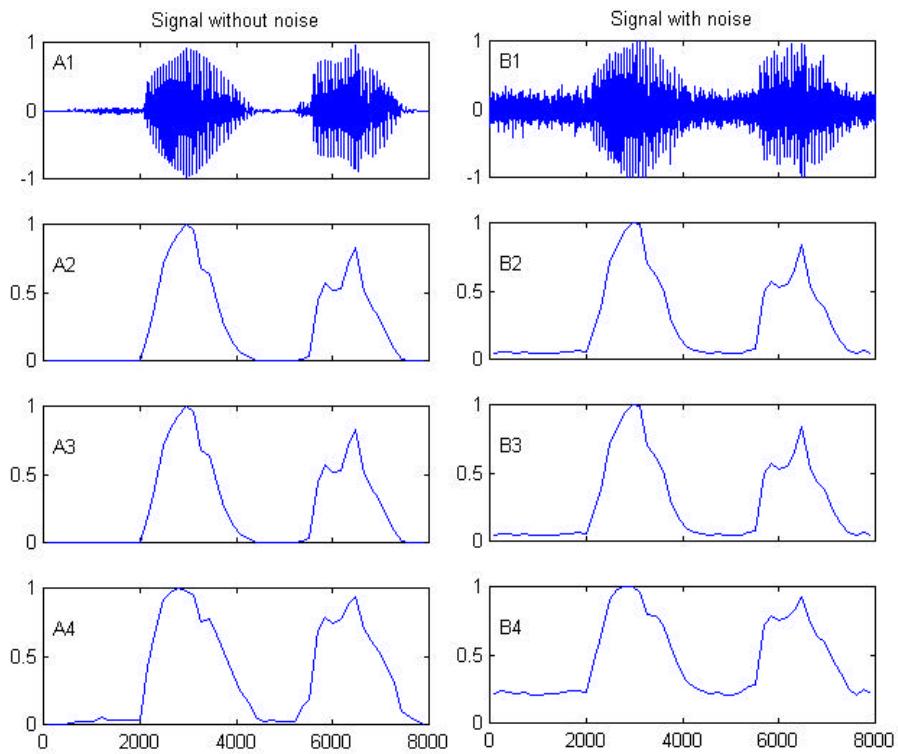


圖 3-3 能量參數應用實例

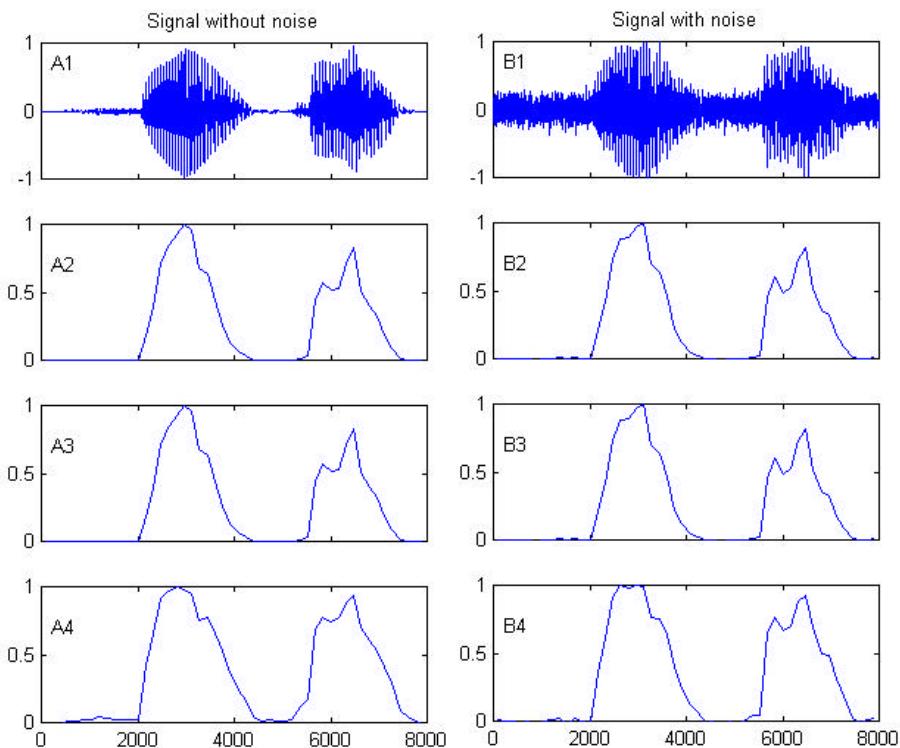


圖 3-4 消除雜訊影響後之能量參數應用實例

絕對值和的方法雖然易受雜訊干擾，但不需平方計算的優勢，使其仍舊是在做即時化系統(real-time system)時的最佳選擇；所以為了克服雜訊對它的影響，在此加入一雜訊消除法(noise cancellation)，使能量絕對值和的方法在含有雜訊的錄音環境下亦能適用。

將錄音訊號經框分析處理後，以最前面3~5個分析框(50ms~100ms)的能量估算值之平均值，當做背景雜訊能量，再將每個分析框的能量估測值減去雜訊能量估測值，即可得到所需的能量估測值，其數學式可表示為：

$$E(n) = E(n) - \frac{1}{N} \sum_{i=1}^N E(i) \quad (3-4)$$

其中， $n=1,2,\dots,N_F$ ，而 N 為背景雜訊能量之分析框數量；如圖 3-4 所示為加入雜訊消除法後所得之能量估測曲線圖。可清楚的發現，加入雜訊消除法後，雜訊對能量絕對值和的方法，影響已不再那麼明顯了。故在下一節聲音訊號端點偵測時，將採用能量絕對值和的方法，再佐以雜訊消除法，以加快端點偵測的速度。

3.2.2 越零率參數(Zero-crossing rate)

越零率參數主要應用於語音訊號端點偵測中，用於判斷氣音段或摩擦音段之訊號。在氣音或摩擦音段，由於訊號的頻率較一般聲音段來的高，所以可藉由此特性來找出錄音訊號中的氣音或摩擦音段所在之處；而要找出此特性最簡單的方法，就是直接去計算在分析框內訊號穿過零準位的次數，此即越零率參數(如圖 3-5 所示，ZCR(n)=15)。

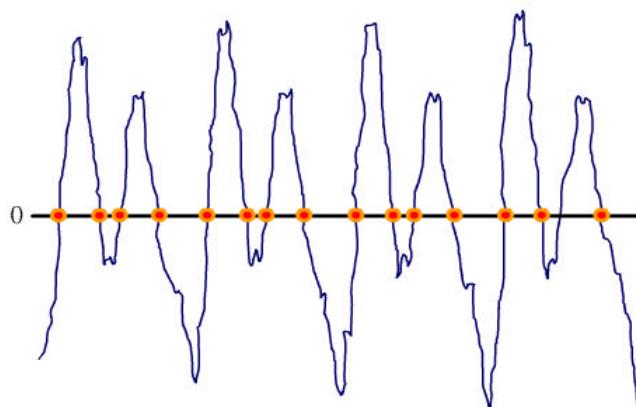


圖 3-5 越零率參數示意圖

越零率參數之數學式可表示成：

$$ZCR(n) = \sum_{i=-(N-1)/2}^{(N-1)/2-1} u[Sgn(-s(p+i) \times s(p+i+1))] \quad (3-5)$$

其中 $u(x)$ 為步階函數， $Sgn(x)$ 為符號函數，其定義如下：

$$u(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (3-6)$$

$$Sgn(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases} \quad (3-7)$$

在(3-5)式中， $s(p+i) \times s(p+i+1)$ 的意思是：目前訊號取樣值與下一個取樣值相乘，若兩者為同號時，其結果必為正數，反之若為異號則其結果必為負數，亦即表示訊號跨越過零準位線，所以(3-5)式可求得實際的越零率參數值。

圖 3-6 為越零率之應用實例，從圖 3-6 中的 A 和 B 圖可發現，有聲段的越零率的確比氣音或摩擦音段來的低，所以可藉由此參數有效的找出氣音或摩擦音段。然而在實際的應用上，如前一節所遭遇的問題：雜訊干擾；越零率最大的缺點就是抗雜訊能力非常的低，此乃因雜訊亦多屬高頻訊號，所以在雜訊干擾的情況下，利用此參數來判斷聲音訊號時，其效能是非常差的，如圖 3-6 中的 C 和 D 圖所示，即可發現幾乎無從找出何處為聲音訊號段。

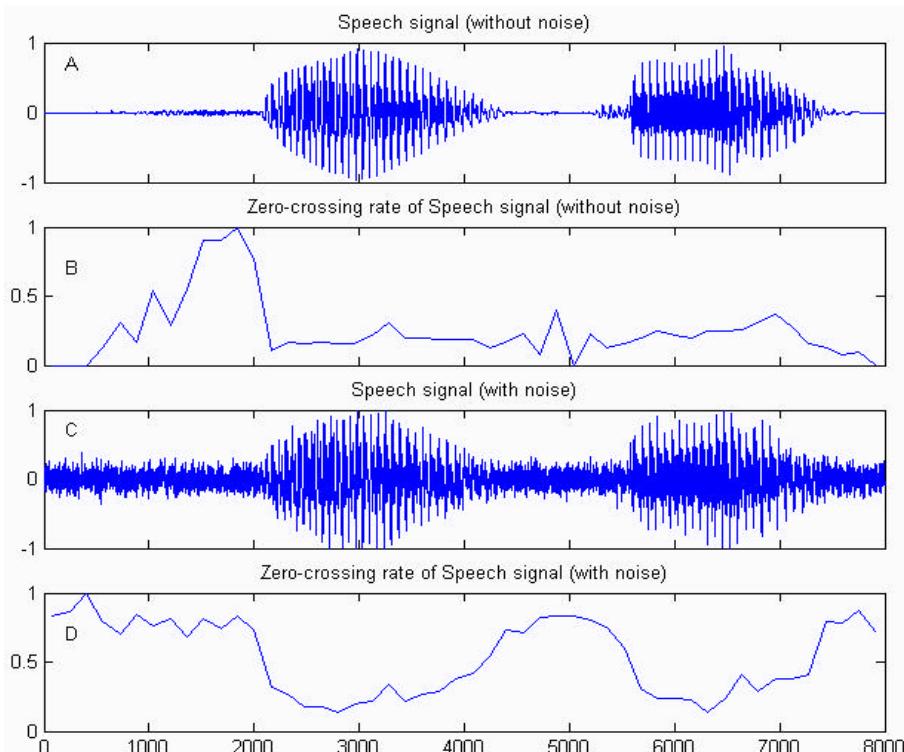


圖 3-6 A.B.為無雜訊干擾；C.D.為受雜訊干擾之越零率參數應用實例

若所設計之系統的傳輸媒介是電話線的話，由於電話線的有效傳送頻率約為 3k~4k Hz，也就是取樣頻率大約在 8k Hz 左右，因此氣音或摩擦音段的訊號將很難傳送出去，故一般在設計此類系統時，通常只採用能量參數，而不考慮越零率參數的影響。但若所設計的系統是針對辦公室或實驗室時，則將越零率參數考慮進去，將可更有效的找出有聲段訊號的精確位置。

3.3 端點偵測法

從錄音訊號中，求得能量曲線圖(Energy Contour)與越零率曲線圖(ZCR Contour)後，即可利用兩曲線所呈現的數據找出語音訊號之端點，以下將針對最容易的端點偵測法：能量曲線判別法(Energy contour method)，以及最被廣泛使用的：R-S 端點偵測法(Rabiner & Sambur method)做詳細的說明。

3.3.1 能量曲線判別法(Energy contour method)[11]

利用在本章前一節中所介紹之能量參數曲線圖，透過門檻值的設定，即可找出有聲段語音訊號的端點位置(如圖 3-7 所示)；經過試驗，能量參數門檻值可取最大能量值的 5%~10% 之間，即可有效找出有聲段訊號端點，同時為了降低雜訊干擾，所以將雜訊的影響也考慮到門檻值設定式中：

$$Thd = C \times \max[E(n)] + \frac{1}{N} \sum_{i=1}^N E(i) \quad (3-8)$$

其中，C 為常數值 5%~10% 之間，最右項為(3-4)式中之雜訊能量估測值。圖 3-8 及圖 3-9 為利用能量曲線判別法的應用實例：C=7.5%，N=5。其中圖 3-8-B 為在做分析框處理時，沒有採用重疊方法所得的結果，而圖 3-8-C 與圖 3-8-D 則分別採用 25% 以及 50% 重疊處理所得之結果，所以在做端點偵測時分析框的重疊處理過程並不會增加端點偵測的精確度，因此可省略此一程序。

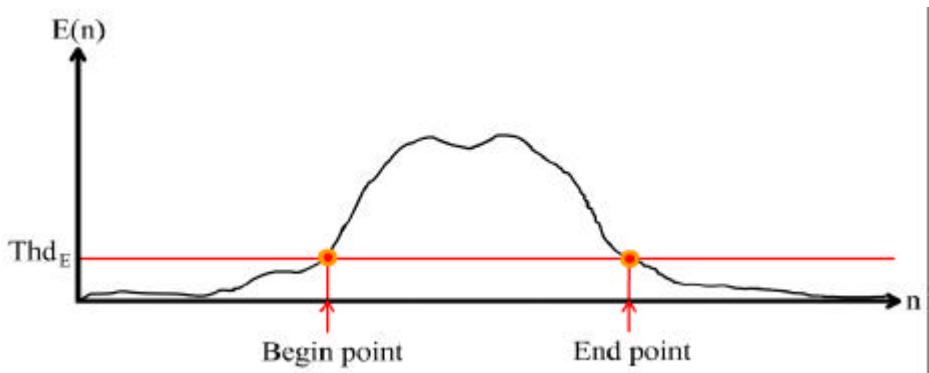


圖 3-7 能量曲線判別法示意圖

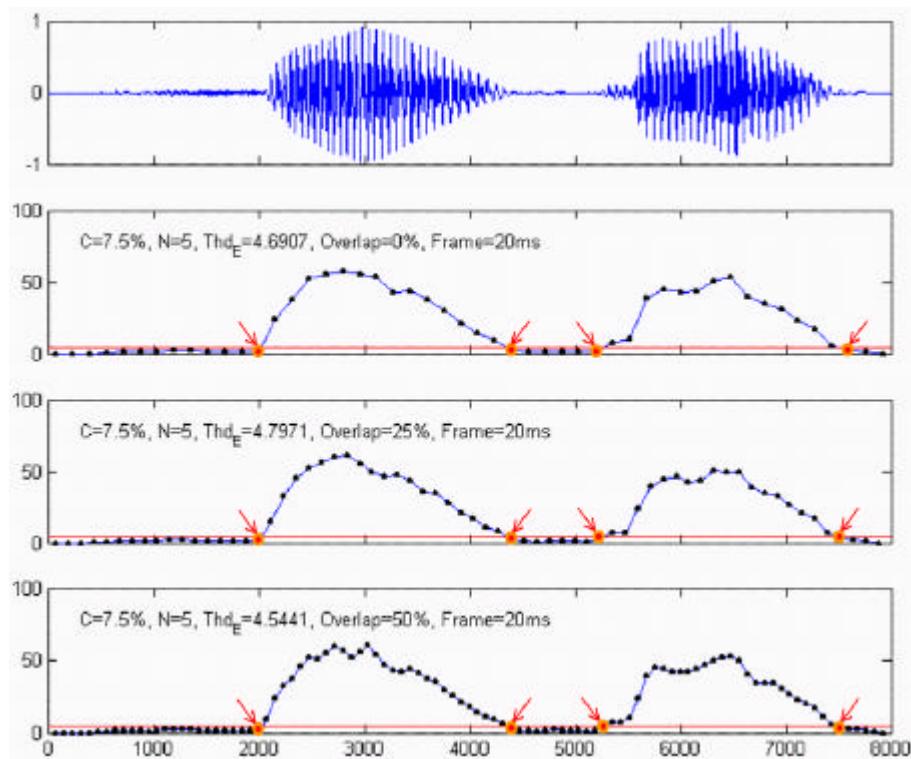


圖 3-8 能量曲線判別法之應用實例(無雜訊干擾)

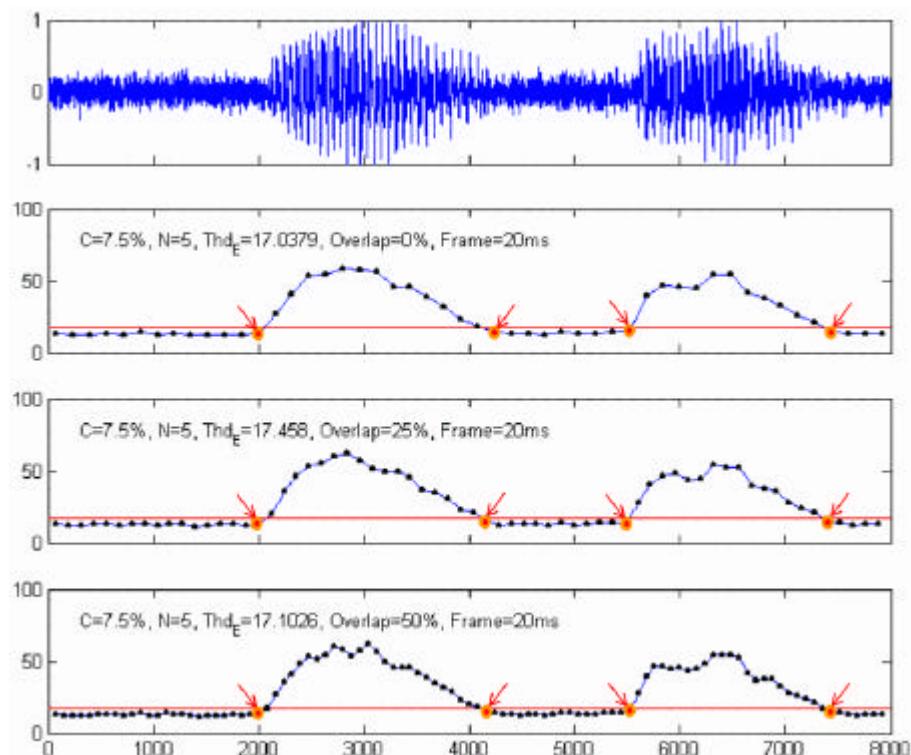


圖 3-9 能量曲線判別法之應用實例(雜訊干擾)

利用能量曲線做端點偵測的過程中，為了降低有聲段的誤判率，所以根據人們說話的特性，對於所偵測的端點加入以下兩個限制條件：

1. 有聲段長度須大 20ms，方可視為有效之語音訊號。
2. 兩有聲段之間的間隔須大於 5ms，此為非連續字音辨識系統中，對說話字音間隔所加入的限制條件。

加入以上兩個限制條件，透過能量曲線圖即可更有效的找出有聲段訊號的端點所在；然而此判別法最大的缺點是：無法找出氣音或摩擦音段，如圖 3-8 中的第一個字音(數字：4)，其前段之氣音部份，便無法偵測出來；所以當氣音或摩擦音段對設計者而言很重要的話，那就得採用 R-S 端點偵測法。

3.3.2 R-S 端點偵測法(Rabiner & Sambur method)

R-S 端點偵測法是由 Rabiner 以及 Sambur 於 1981 年所提出的語音訊號端點偵測法[12]，此法主要是利用訊號的能量曲線與越零率曲線，來判斷有聲段訊號的端點所在之處，在本章中所介紹之 R-S 端點偵測法是參考該論文中之基本概念，再經過簡化後的結果，若使用者有興趣請參考原始論文中之說明。

首先，根據語音訊號之特性定出能量曲線門檻值(Thd_E)以及越零率曲線之門檻值(Thd_Z)：

$$Thd_E = C_E \times \max[E(n)] + \frac{1}{N} \sum_{i=1}^N E(i) \quad (3-9)$$

$$Thd_Z = C_Z \times \max[ZCR(n)] \quad (3-10)$$

其中(3-9)與(3-8)式是一樣的，其中 C_E 為能量曲線常數介於 5% ~ 10% 之間，而 C_Z 為越零率曲線常數值介於 20% ~ 40% 之間。

參考圖 3-10 之 R-S 判別法示意圖，利用能量門檻值 Thd_E 所畫的水平線，記錄其與能量曲線之交點位置：P1 與 P2，分別以這兩個分析框位置做為有聲段的初步起始點與結束點之所在；因為 P1、P2 為初步的端點偵測結果，事實上它們之間所包含的訊號可能僅是有聲段的一部份，無法有效的找到氣音或摩擦音的部份，為了得到準確的端點位置，所以利用越零率曲線圖並畫上越零率門檻值 Thd_Z 的水平線，在此圖中自 P1 及 P2 向兩邊延伸，以找出真正的語音訊號之端點位置。因為由 P1 向左搜尋有聲段的起始點位置的步驟，與由 P2 向右搜尋結束點位置的步驟是一樣的，所以在此只說明如何自 P1 延伸找到有聲段的起始端點位置。

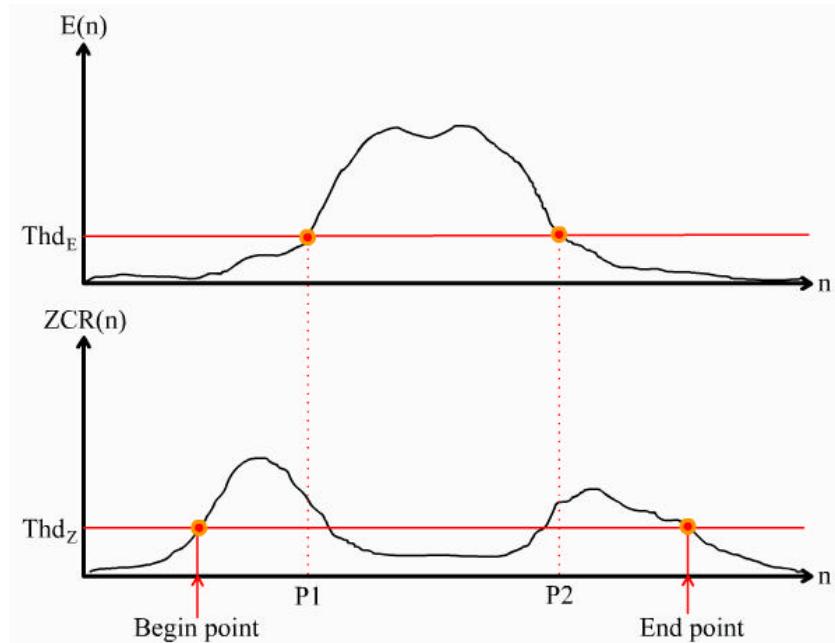


圖 3-10 R-S 判別法之示意圖

在找到 P1 的位置後，對照到越零率曲線圖，若此時該分析框的越零率值低於門檻值 Thd_Z ，則 P1 即視為該字音的起點位置；若高於門檻值 Thd_Z 時，就從 P1 向左尋找，逐一比較前一個分析框的越零率是否大於門檻值 Thd_Z ，當找到第一個穿零率小於 Thd_Z 之分析框時，則該分析框所對應的時間指標即是所偵測到的有聲段起始點。而有聲段的終點搜尋方式亦相同唯方向相反。

為了降低此偵測法對有聲段的誤判率，得像能量曲線端點偵測法中一樣，得加入一些限制條件：

1. 有聲段長度須大 20ms，方可視為有效之語音訊號。
2. 兩有聲段之間的間隔須大於 5ms。
3. 在越零率曲線搜尋時，從 P1 往前(或 P2 往後)搜尋的距離不可超過 150ms，若超過 150ms 後，越零率值仍大於門檻值 Thd_Z 時，則將之視為雜訊干擾，仍然以 P1 為有聲段之起點。

加入以上三個限制條件，透過能量曲線圖以及越零率曲線圖，便可有效的找出有聲段訊號的端點所在，同時亦可找出氣音或摩擦音段，比較圖 3-8 與圖 3-11 可清楚發現，第一個字音(數字：4)，其前段之氣音部份，已可偵測出來；此特色即為 R-S 端點偵測法與能量端點偵法之最大差異。

此法雖可有效找出氣音或摩擦音段，然而受限於越零率參數最薄弱的特性：雜訊免疫力低所致；使得 R-S 端點偵測法對於受雜訊干擾之訊號的端點判別能力變的很差，如圖 3-12 所示，即可清楚發現此一缺點。當雜訊干擾嚴重時，利用 R-S 端點偵測法所得到的結果，與採用能量端點偵法所偵測到結果是一樣的。

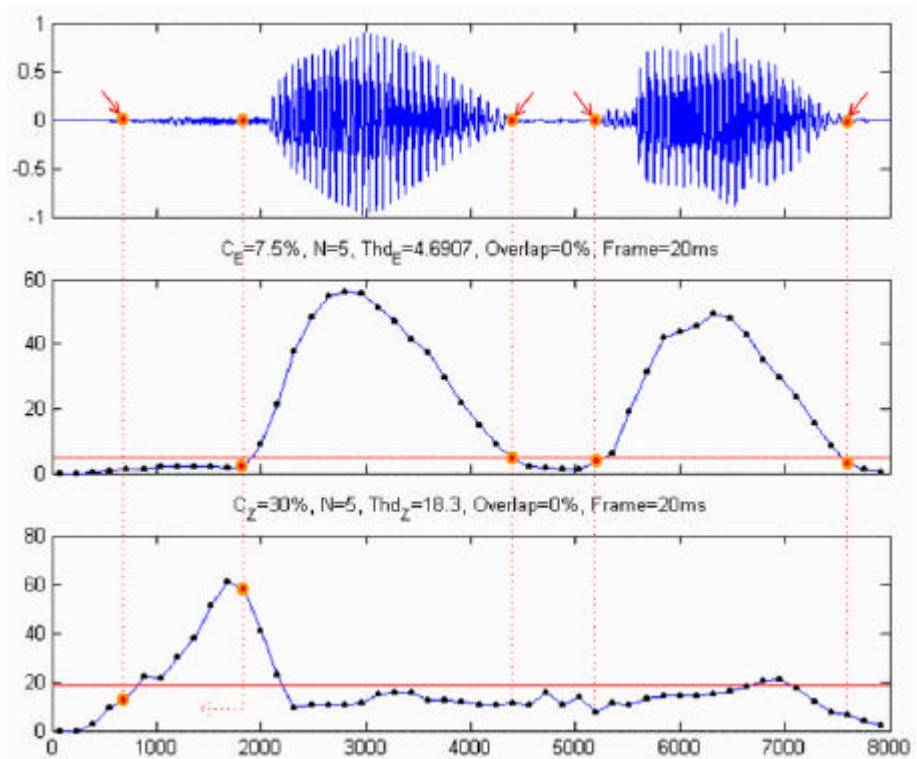


圖 3-11 R-S 判別法之應用實例(無雜訊干擾)

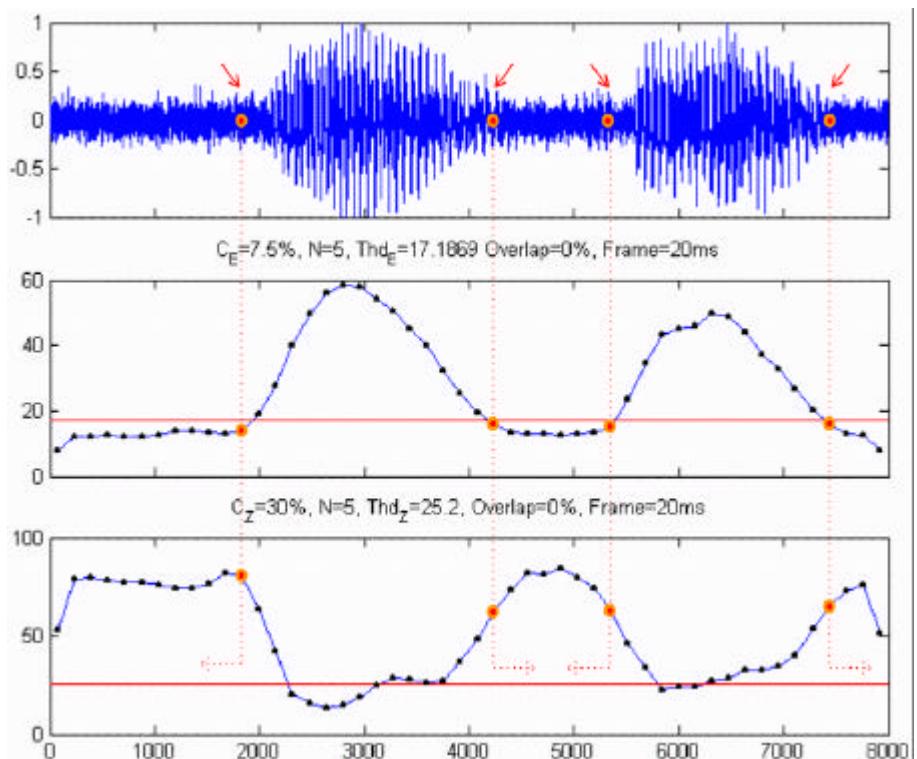


圖 3-12 R-S 判別法之應用實例(雜訊干擾)

第四章 特徵參數擷取

4.1 簡介

每個人說話時，隨著其性別、年齡、地域等因素都會有其特定的發音方式，即使是同一個人，在不同心理狀況或生理狀態下所產生的語音訊號也會有所差異；因此若直接採用語音訊號的波形來從事比對的工作，不僅資料的處理量很大，同時所得到的辨識率也是非常有限的；因此在從事語音訊號處理時，得先求得較適當語音訊號特徵參數。從每個分析框中，計算其所對應的語音訊號特徵，並將他們組成所謂的特徵向量(feature vector)，經處理後，原本的語音訊號便可以此特徵向量來取代，以做為系統辨識的依據，此過程即為語音訊號特徵擷取(feature extraction)。

在語音辨識上常使用的語音特徵可分為兩大類；一為頻譜特徵(spectral features)，另一為倒頻譜特徵(cepstral features)。頻譜特徵係指語音訊號在各頻帶的平均能量分佈；倒頻譜特徵則是指語音訊號的倒頻譜係數。雖然頻譜特徵似乎更接近人類聽覺系統所處理的訊息，但根據目前研究的結果顯示，使用倒頻譜特徵往往可得到較高的語音辨識率，因此倒頻譜特徵已成為目前最常被使用的語音特徵。在本章中將介紹兩種以數位濾波器組所求得的特徵參數：線性倒頻譜係數(Linear-Frequency Cepstrum Coefficient : LFCC)與梅爾倒頻譜係數(Mel-Frequency Cepstrum Coefficient : MFCC)。

在做語音訊號特徵參數求取之前，須對語音訊號做前置強波處理，主要是濁音段語音訊號的頻譜，會隨著頻率的增加而衰減，如此一來將會造成辨識時，高頻部份的比重較輕的現象，所以經端點偵測所得之有效語音訊號，需要再做前置強波處理工作(preemphasis)，其詳細過程將在第二節做說明。

於語音訊號特徵參數擷取過程中，為了能有效的求出語音訊號的特性，所以需要做框分析處理，同時為有效呈現出短時矩內訊號特性的變化，在分析時得採用漢明分析框而且還得做分析框重疊處理的步驟，如此才能有效的求得語音訊號的特徵，關於此過程以及倒頻譜參數之詳細說明，將在第三節做解說。

4.2 前置強波處理(Preemphasis)

前置強波處理主要的目的[13]，就是為了補償語音訊號中濁音訊號會隨著頻率的增加而衰減的現象，所以在特徵參數擷取之前，先將所測得之有效語音訊號段經一高通濾波器處理，其轉換函數可表示成：

$$H(z) = 1 - az^{-1} \quad (4-1)$$

其中， a 濾波器參數($0.9 < a < 1.0$)，一般取 $a=0.95$ ；若將上式轉換成數列的型式來表示，則可表示成：

$$\hat{s}(n) = s(n) - as(n-1) \quad (4-2)$$

其中， $\hat{s}(n)$ 為處理後之訊號， $s(n)$ 為原訊號；前置強波處理的過程，可視為將原本之語音訊號經一高通濾波器處理所得之結果。圖 4-1 為前置強波處理器(濾波器)的頻率響應圖，左圖為 $a=0.95$ ，右圖為 $a=0.9$ 。圖 4-2 為實際之應用實例，從圖中可清楚的發現處理後之頻譜分佈，在高頻的部份已明顯被強化了(放大)。

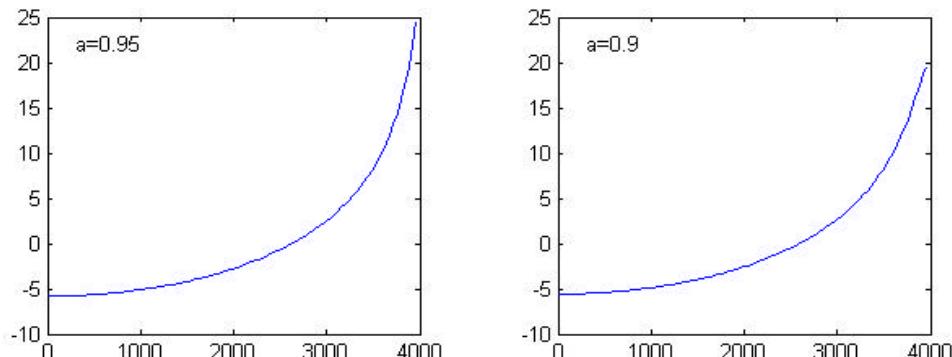


圖 4-1 前置強波器之頻譜圖

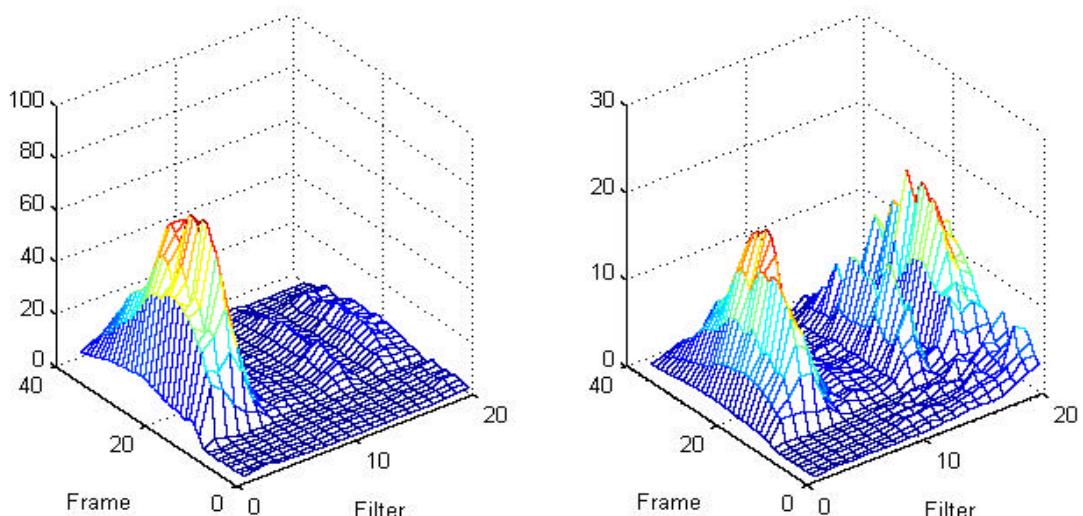


圖 4-2 前置強波器之應用實例(右圖為處理後之結果)

4.3 倒頻譜參數(Cepstrum Coefficient)

對任何語音辨識系統而言，語音訊號特徵參數的選擇是非常重要的，可用來表示語音訊號的特徵參數有很多種[14]，而在本節中僅就：線性倒頻譜係數(Linear-Frequency Cepstrum Coefficient : LFCC) 以及梅爾倒頻譜係數(Mel-Frequency Cepstrum Coefficient : MFCC)，做詳細的說明，此兩語音特徵參數都是利用數位濾波器組(多個帶通濾波器)來處理語音訊號，再將每個濾波器的頻譜能量值做參數轉換，所得到之語音訊號特徵參數。

由於人類聽覺系統對語音訊號之相位(phase)關係並不敏感，因此語音訊號的頻譜特徵皆指語音訊號在各頻帶的能量分佈而言；故可依據頻譜特徵隨時間的變化情形來進行語音辨識。擷取語音訊號的頻譜特徵最直接的方法，便是將訊號送入一組並聯式數位帶通濾波器組(digital band-pass filter bank)中，如圖4-3所示。各帶通濾波器將訊號的頻譜分割成數個部分，在各個濾波器的輸出端所觀察到的結果，便是訊號存在該頻帶的組成成份。

根據帶通濾波器，過濾頻帶寬度的分割方式的差異，可將數位帶通濾波器組分為：(1)均勻分割(uniform partition)，線性倒頻譜係數(LFCC)即是屬於此類；(2)非均勻分割(nonuniform partition)，梅爾倒頻譜係數(MFCC)即是屬於此類。以下將針對此兩特徵參數做詳細的說明。

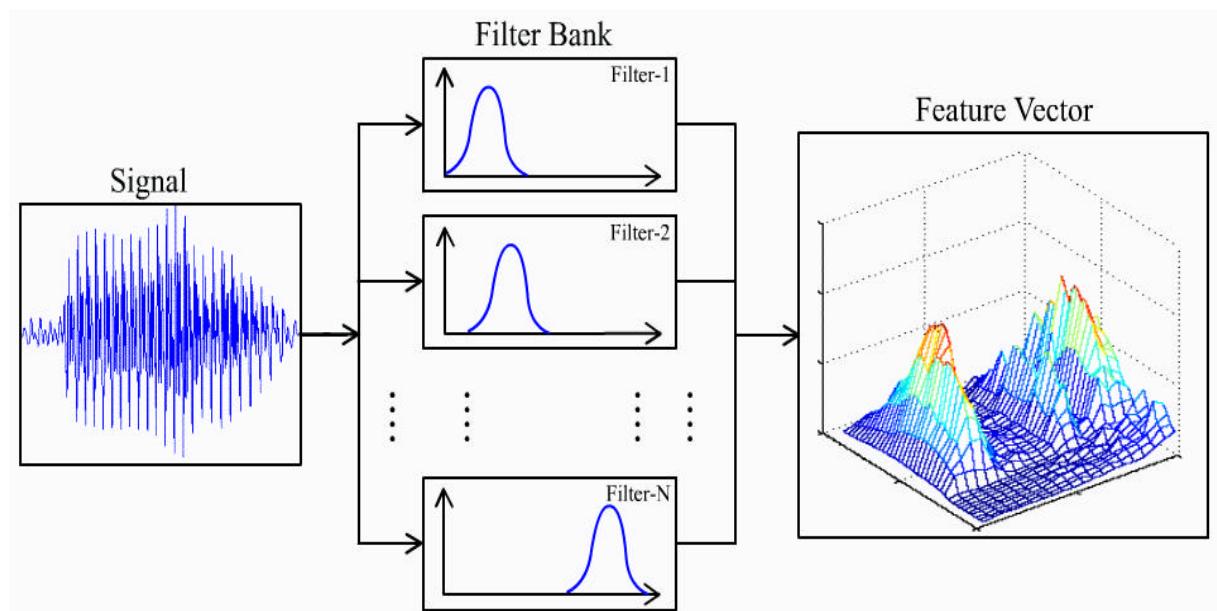


圖 4-3 並聯式數位帶通濾波器組

4.3.1 線性倒頻譜係數

利用均勻分割之數位帶通濾波器組的主要概念是：將人類聽覺系統的感知頻率與實際頻率間的關係視為線性關係，即人類聽覺對聲音頻率反應的感知度全都視為一樣，所以我們將有效的頻率範圍內之頻寬均勻的分割，以線性刻度來設計帶通濾波器組(如圖 4-4 所示)，再將每個分析框之訊號經此濾波器處理，處理後所獲得的每個頻帶能量值，即為該分析框之語音訊號特徵參數，或稱之為線性頻譜參數(Linear Frequency Spectrum : LFS)，如圖 4-5 之左下圖所示。

直接將線性頻譜參數做為語音辨識系統之特徵參數，即可獲得辨識效果，然而在此情形下，系統所需之參數量將會很大，以分割成 20 個帶通濾波器之系統而言，每個分析框就需要記錄 20 個參數，除此之外為了加大參數間之差異性，所以我們將線性頻譜參數再做倒頻的轉換計算，所獲得的參數稱之為線性倒頻譜係數 LFCC，其轉換公式如下：

$$LFCC(i) = \sum_{k=0}^{N-1} LFS(k) \times \cos\left(\frac{ik\mathbf{P}}{N}\right) \quad i = 1, 2, \dots, M \quad (4-3)$$

其中， $LFCC(i)$ 為第 i 個倒頻譜係數， N 為帶通濾波器之個數， $LFS(k)$ 為第 k 個線性頻譜參數之能量值。一般的語音辨識系統採用 $N=20$ ， $M=12$ 。如圖 4-5 右下圖所示即為線性倒頻譜係數的應用實例。

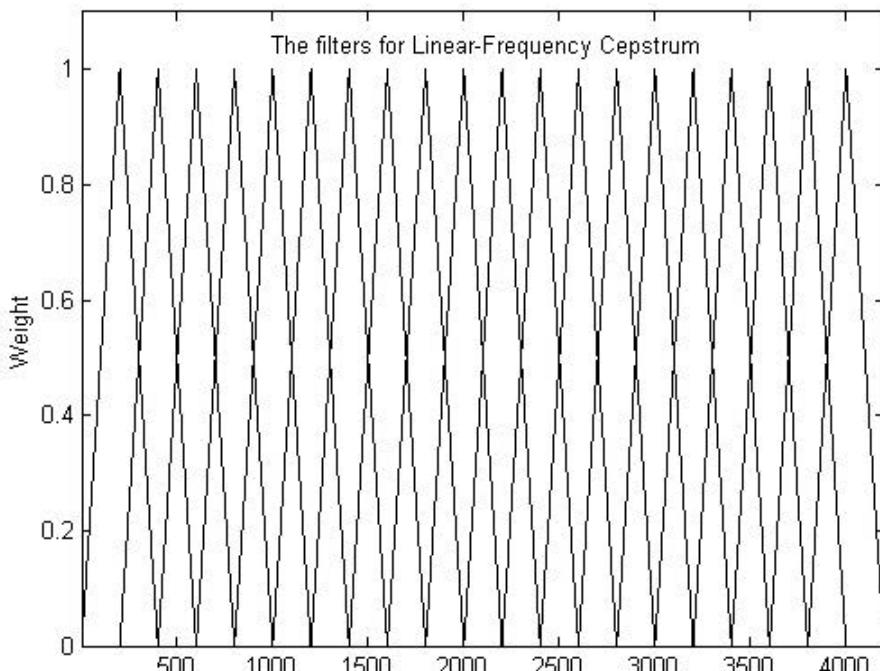


圖 4-4 以線性刻度所設計之帶通濾波器

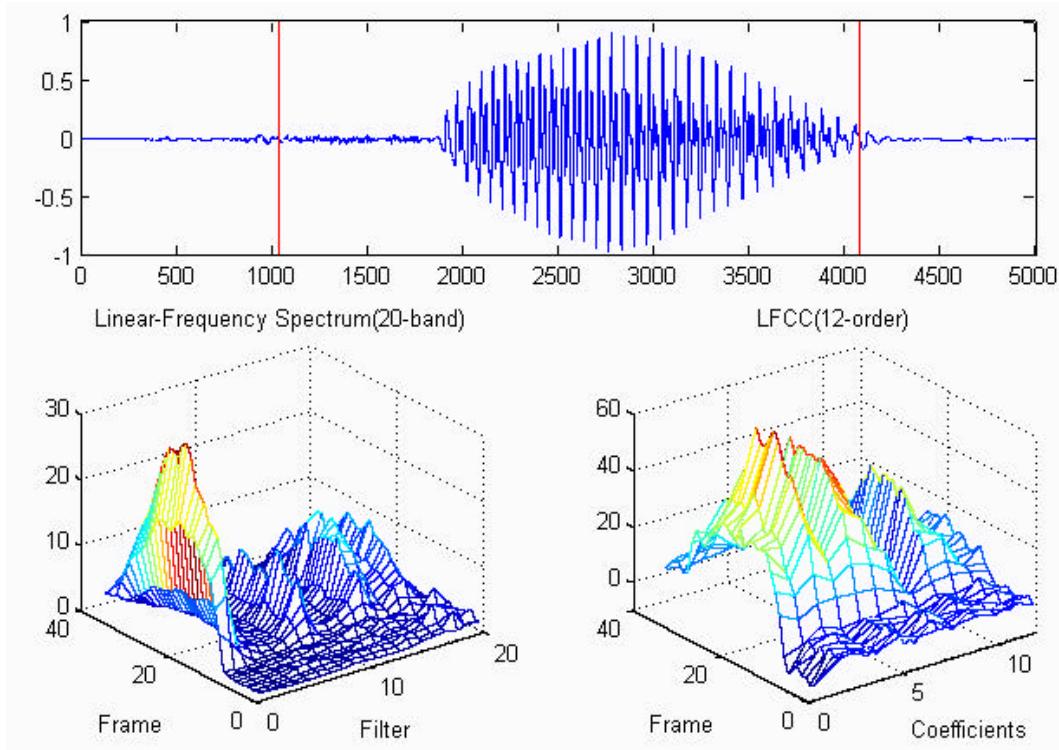


圖 4-5 以線性刻度濾波器之應用實例(LFS 與 LFCC)

4.3.2 梅爾倒頻譜係數

根據人耳聽覺特性的實驗結果顯示，當人耳在傾聽不同基頻頻率的單音(tone，即只含單一基頻的聲音)時，所感知到的頻率(perceptual frequency)並非是呈線性關係；感知頻率與實際頻率間的關係可表示成：

$$Bark = 13 \tan^{-1}(0.00076f) + 3.5 \tan^{-1}\left(\frac{f^2}{7500^2}\right) \quad (4-3)$$

其中，Bark 為感知頻率刻度的單位， f 為訊號的實際頻率，而 $\tan^{-1}(x)$ 的運算結果是以徑度為單位。圖 4-6 左圖所示為 Bark 刻度與實際頻率的關係。

在語音辨識系統領域裏，最常使用的是一種較簡單且近似感知頻率刻度的轉換公式，稱之為梅爾頻率刻度(Mel-scale frequency)，其公式如下：

$$Mel = 2595 \times \log\left(1 + \frac{f}{700}\right) \quad (4-4)$$

$$f = 700 \times (10^{\frac{Mel}{2595}} - 1) \quad (4-5)$$

(4-4)式為頻率刻度轉換成梅爾頻率刻度的公式，而(4-5)式則為(4-4)式之反轉換公式。

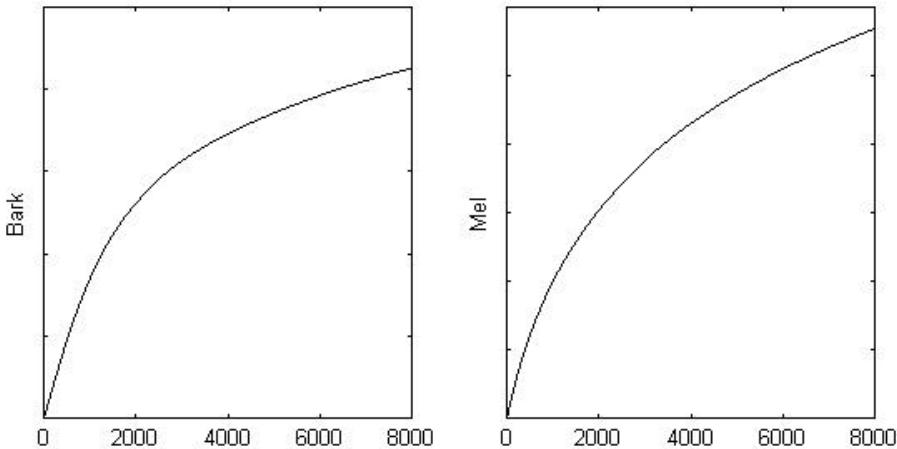


圖 4-6 Bark 及 Mel 與實際頻率之關係圖

圖 4-6 右圖即為梅爾刻度與實際頻率的關係；在頻率小於 1k Hz 時，Mel 的刻度與實際頻率刻度具有線性關係；而在頻率大於 1k Hz 時，Mel 刻度與實際頻率刻度間則呈現對數關係。比較圖 4-6 中之左右兩圖可發現，Bark 與 Mel 兩者與實際頻率的關係很相似，且 Mel 的轉換計算較簡單，故一般在語音辨識上都採用梅爾刻度。

根據以上所述之人類聽覺特性來設計數位帶通濾波器組，則在低頻區之帶通濾波器較密集，且頻寬較窄；隨著頻率的增加，在高頻區之帶通濾波器較為疏散，且頻寬較寬，如圖 4-7 所示即為根據梅爾刻度頻率所設計之三角型數位帶通濾波器，其乃將有效頻率 4k Hz 分割成 20 個頻帶。再以這 20 個帶通濾波器來處理每個分析框內之訊號，即可得到該分析框內聲音訊號的頻譜能量值參數，此特徵參數稱之為梅爾頻譜參數 (Mel-Frequency Spectrum : MFS)，如圖 4-8 左下圖所示為此特徵參數之應用實例。

如同在線性頻譜參數一樣，可直接將梅爾頻譜參數做為語音辨識系統之特徵參數，然而系統所需之參數量亦將會很大，所以也可將梅爾頻譜參數做倒頻的轉換計算，而所獲得的參數稱之為梅爾倒頻譜係數 MFCC，其轉換公式如下：

$$MFCC(i) = \sum_{k=1}^N MFS(k) \times \cos\left(\frac{(k-0.5)iP}{N}\right) \quad i=1,2,\dots,M \quad (4-6)$$

其中，MFCC(i)為第 i 個倒頻譜係數，N 為帶通濾波器之個數，MFS(k)為第 k 個梅爾頻譜參數之能量值。一般的語音辨識系統採用 N=20，M=12。如圖 4-8 右下圖所示即為梅爾倒頻譜係數的應用實例。

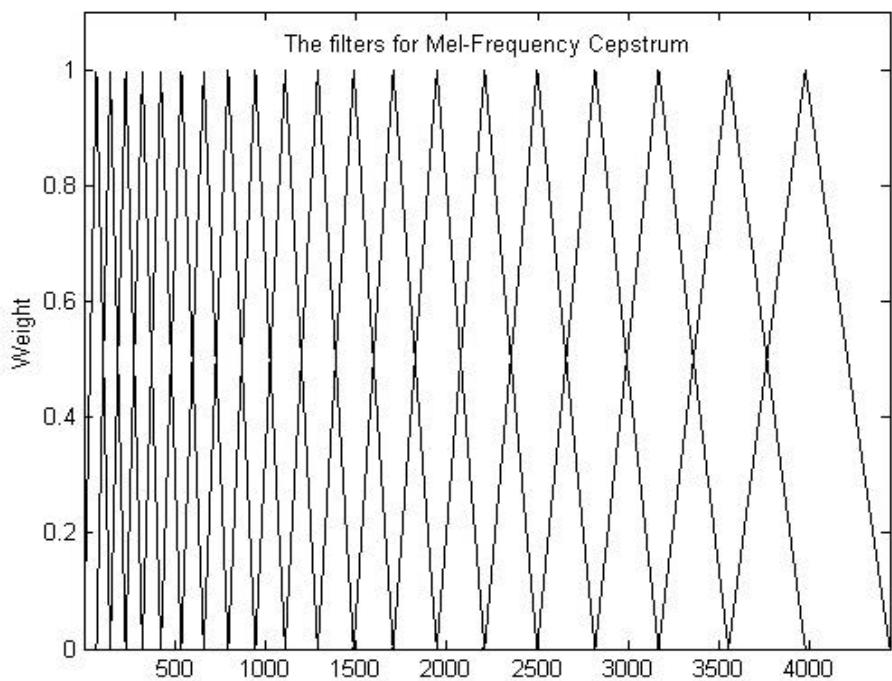


圖 4-7 以梅爾刻度所設計之帶通濾波器

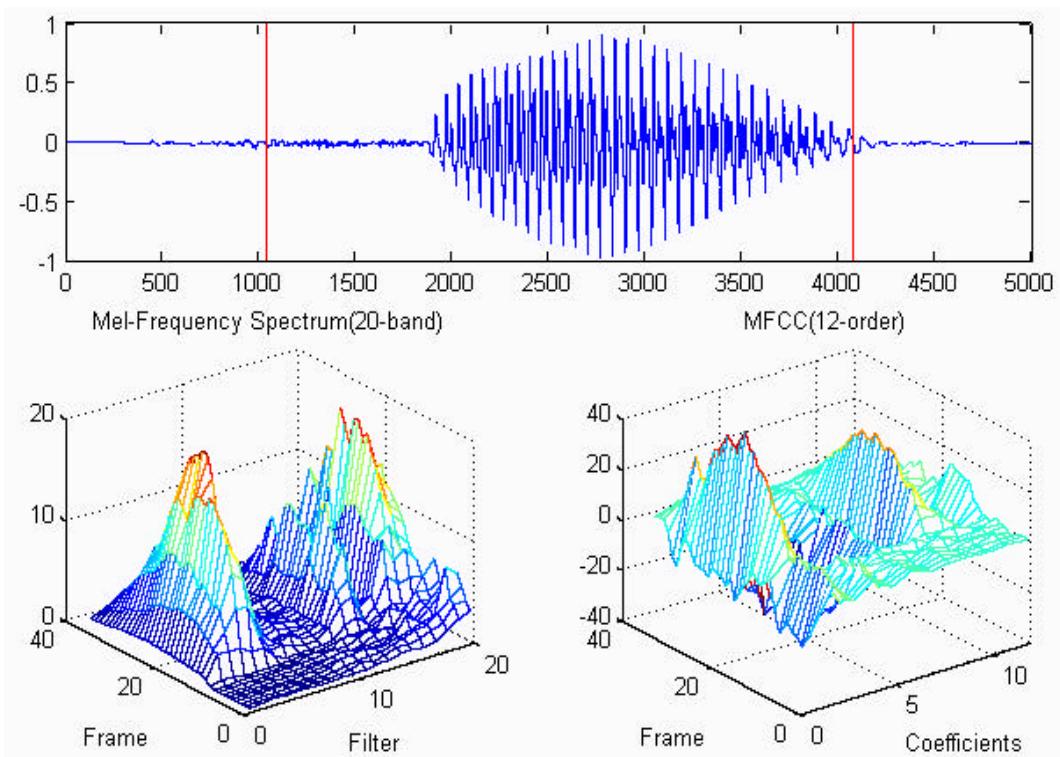


圖 4-8 以梅爾刻度濾波器之應用實例(MFS 與 MFCC)

4.4 頻譜參數與倒頻譜參數之比較

第三節中詳細介紹濾波器組求得語音特徵參數的方法，在所介紹的四種特徵參數中可分為兩類：(1)頻譜參數，包括：LFS 與 MFS；(2)倒頻譜參數，包括：LFCC 與 MFCC。從實驗中可發現：採用倒頻譜(cepstrum)參數的辨識系統，其收斂速度比採用頻譜(spectrum)參數的系統來的快。

如圖 4-9 所示為以類神經網路為架構之非特定語者(speaker independent)中文數字語音辨識系統，分別採用四種不同的特徵參數所做的測試；其中用了 15 組(9 男，3 女)訓練樣本，6 組(3 男，3 女)測試樣本，每個中文字音，每人錄三次；從測試的結果可發現：不論是均勻分割或非均勻分割之濾波器組，倒頻譜參數的收斂速度較快，且所需的參數量也較少。故此後在本文中將以倒頻譜參數 MFCC 做為系統的特徵參數。

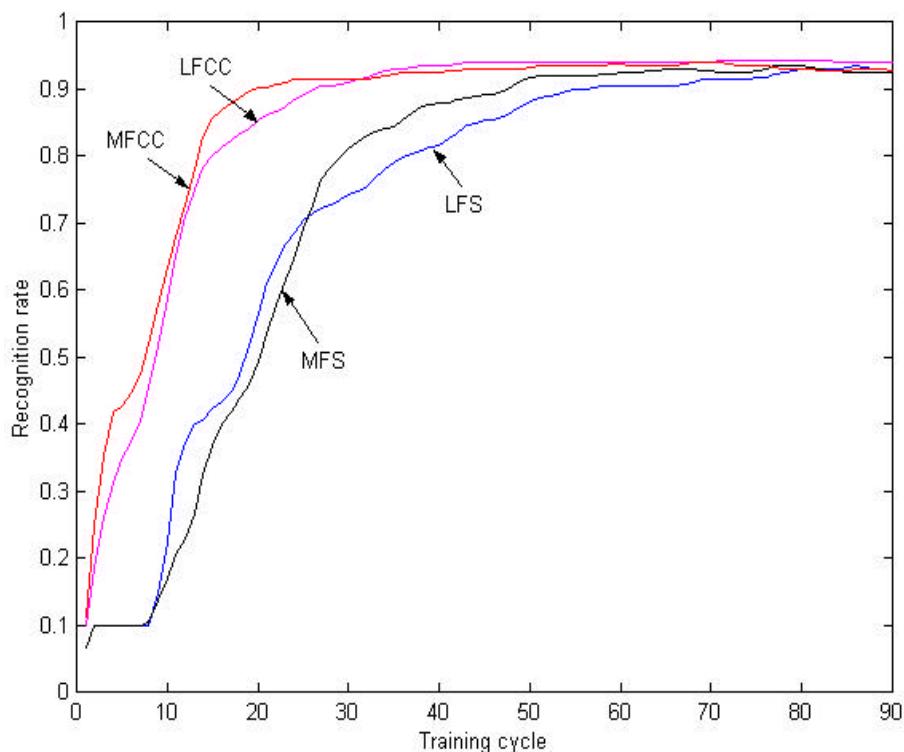


圖 4-9 四種特徵參數對辨統系統的影響

第五章 動態時間較準演算法

5.1 簡介

如圖 5-1 所示為一完整之語音辨識系統，首先錄音訊號經端點偵測，擷取出有聲段之訊號，再經一前置強波處理，將高頻訊號之能量放大，接著求得該有聲段之特徵向量參數(或稱之為圖樣)，以做為系統辨識時之比對依據(或建立所需之語音資料庫樣本)，最後再把所擷取之特徵向量透過動態時間校準法進行圖樣比對，即可獲得語音辨識之結果。關於語音辨識中，圖樣比對或辨識的方法有很多，而在本文中僅探討兩種辨識演算法，第一個為本章所要介紹的主題：傳統的動態時間校準法(Dynamic Time Warping : DTW)[15]，以及下一章所要介紹的方法：倒傳遞類神經網路(Backpropagation Neural Network : BPNN)。

語音訊號經特徵參數擷取之後被轉換為特徵向量；在進行語音辨識時，可將輸入之語音特徵向量視為一個圖樣(pattern)，若字音的語音資料庫樣板圖樣(template patterns)已事先建立，則語音辨識的問題便只是輸入之語音的圖樣與語音資料庫中，各樣板圖樣間的比對工作而已，而辨識的結果便是選擇與輸入語音圖樣最接近的樣板所對應的字音。然而由於講話速度及說話者的習慣等因素，即使是同樣的字音也會有訊號(或特徵向量)長短不一的現象。因此使用這種圖樣比對的語音辨識方法所遭遇最大的困難，便是如何校準兩圖樣的時間對應關係，以計算其相近程度。

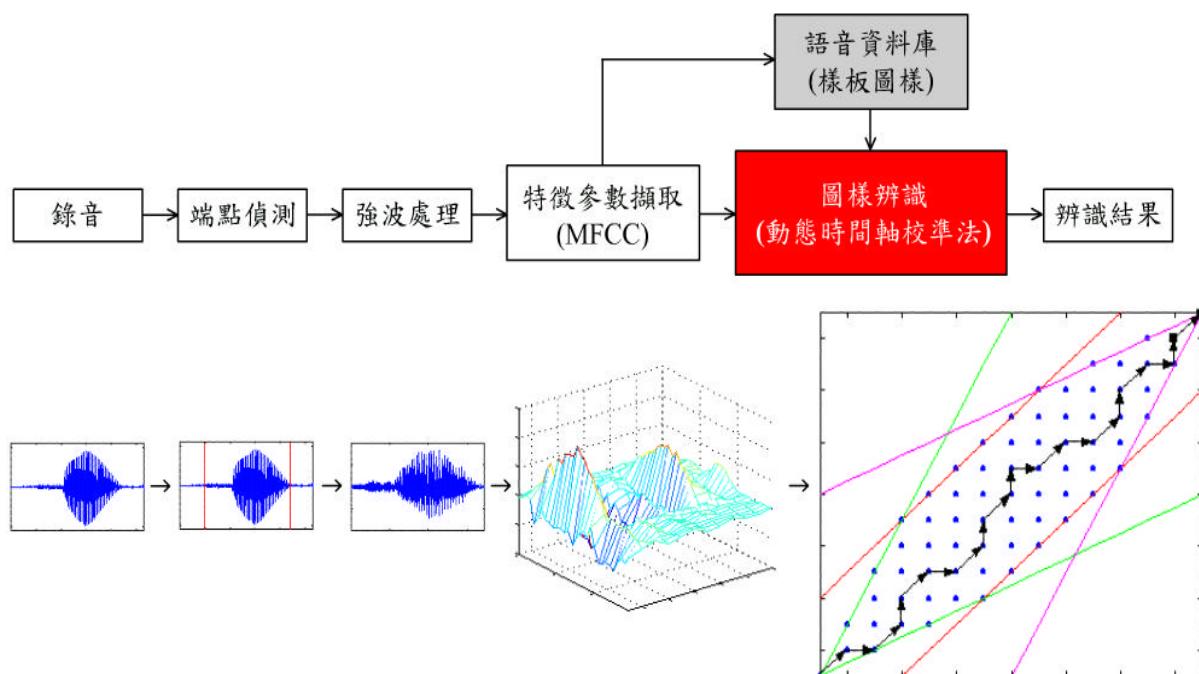


圖 5-1 動態時間校準法之語音辨識系統

假設現有兩組比對的特徵向量 $R = (u_1, u_2, \dots, u_m)$ 及 $T = (v_1, v_2, \dots, v_n)$ ，其中 M 及 N 為這兩組向量的長度，而 u_i 及 v_i 為對應於第 i 個時間指標的特徵向量。一般直覺的時間校準方法便是按照向量的長短比例做線性對應。例如 $R = (u_1, u_2, u_3)$ 且 $T = (v_1, v_2, v_3, v_4, v_5, v_6)$ ，則其線性對應關係可對應成： $u_1 \Leftrightarrow v_1, v_2$ 、 $u_2 \Leftrightarrow v_3, v_4$ 與 $u_3 \Leftrightarrow v_5, v_6$ ；則兩特徵向量的距離(或圖樣的相似程度)可計算如下：

$$d(R, T) = d(u_1, v_1) + d(u_1, v_2) + d(u_2, v_3) + d(u_2, v_4) + d(u_3, v_5) + d(u_3, v_6) \quad (5-1)$$

其中 $d(u_i, v_j)$ 為兩對應特徵向量的距離。

然而這種線性的時間對應關係，並不符合語音訊號的特性。例如當說話速度放緩時，母音段的發音時間易被拉得較子音段為長，在此情況下若是使用上述的線性時間刻度校準，所得到的比對結果必然無法反應圖樣間真正的相似程度；動態時間校準便是針對上述問題所發展出來的一種非線性時間校準技術，該技術是直接應用最佳化理論(optimization theory)中的動態規劃(dynamic programming)。在隱藏式馬可夫模型 HMM 與類神經網路 ANN 的語音辨識法大量使用之前，動態時間校準法為主要的語音辨識方法。

下一節中，將介紹如何使用動態規劃的方法，找出以格狀連接的兩個點間最短的途徑及其距離，以及將兩語音的特徵圖樣之時間校準問題轉換為兩點最短途徑的搜尋，如此一來動態規劃的方法便可直接應用到語音圖樣時間刻度的校準及相似性的計算。最後在第三節中將討論如何建立特定語者(Speaker Dependent : SD)辨識字音的語音資料庫樣板圖樣，以及利用動態時間校準法所進行的語音辨識系統性能測試。

5.2 動態時間校準法(Dynamic Time Warping)

在討論使用動態規劃以解決語音特徵向量的時間校準問題之前，在此先介紹如何使用該方法解決尋找最佳途徑的問題，即可發現使用動態規劃方法解決這兩類問題的相似性，進而發展出所謂動態時間校準的語音辨識法。

5.2.1 動態規劃(Dynamical Programming)[16]

首先定義最佳途徑尋找的問題如下：在圖 5-2 中的格狀網路，每一個交會點稱為節點(node)，分別以英文字母做為各節點的代碼。由線段直接連接的兩個節點稱為相鄰節點，舉例而言 A 與 B 為相鄰節點，但是 A 和 F 則否。連接兩相鄰節點的線段上所列之數字代表穿過該線段所需付出之成本(cost)。假設前進路徑只允許向上及向右，則從節點 A 前進至節點 P 付出最少成本的最佳途徑為何？

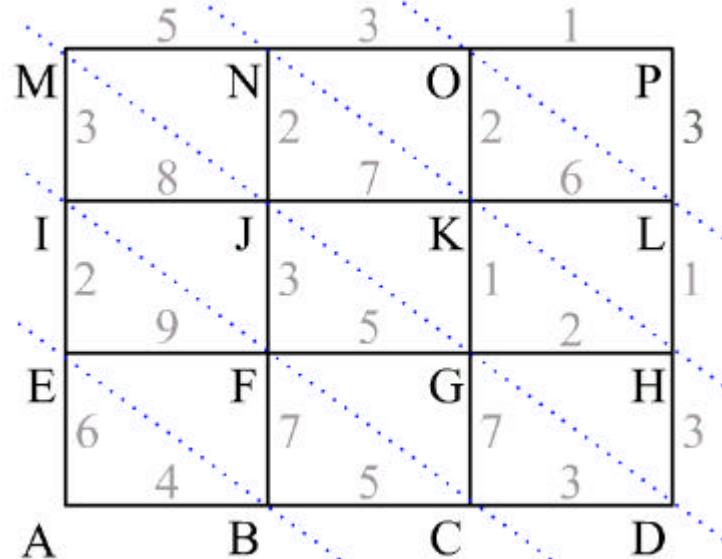


圖 5-2 最佳途徑尋找問題圖例

如果直接使用土法煉鋼法來解決上述問題，則可列出所有可能的途徑，並計算對應於各個途徑所需之成本，再選擇其最少者即可。在這個例子中，從 A 到 P 可能的途徑包括：

- S1: $A \rightarrow B \rightarrow F \rightarrow G \rightarrow K \rightarrow L \rightarrow P$
- S2: $A \rightarrow B \rightarrow C \rightarrow G \rightarrow K \rightarrow L \rightarrow P$
- S3: $A \rightarrow E \rightarrow F \rightarrow J \rightarrow K \rightarrow O \rightarrow P \dots$

若全列出來，則可發現有 20 個可能途徑需要計算及比較，當介於 A 與 P 兩點間的可能途徑數目不是很多的情形，這種方法尚為可行；但是隨著網路大小的增加，可能途徑的尋找變得非常複雜、繁瑣。依據公式，如果格狀網路的外圍每邊邊長的分割線段數為 n ，則所有可能的途徑數目可經由公式求得：

$$S = \frac{(2n)!}{(n!)(n!)} \quad (5-2)$$

其中， S 為可能途徑數目 (search path)；由(5-2)式中可知，當網路稍大時，可能的途徑數變得相當的多，所以逐一列舉計算的方法便顯得相當沒有效率。而動態規劃提供解決這類問題的一種快速演算法，其乃根據 Bellman 所提出的最佳原則 (principle of optimality)：

“策略(policy)是由一連串的決定(decisions)所構成。一個最佳策略有如下特性：不論前面一連串的決定與初始狀態為何，剩餘的決定對經由前面一連串決定所抵達的目前狀態而言亦構成一最佳策略”

為配合介紹語音特徵的時間校準方法，在應用上述原則來解決最佳途徑問題時，假設途徑的選擇是由終點 P 往前倒推回到起始點 A。在圖 5-2 中，假設經由一連串的最佳途徑選擇，最後抵達節點 F。根據最佳原則，如果我們再選擇自節點 F 到起始點 A 的最佳途徑，亦即 $F \rightarrow B \rightarrow A$ （所付出之成本為 11 較途徑 $F \rightarrow E \rightarrow A$ 的 15 為少，如圖 5-3 所示），則這整個途徑便是從終點 P 到起始點 A 的最佳途徑。

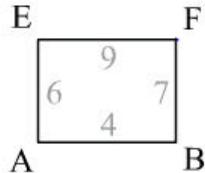


圖 5-3 自節點 F 到節點 A 的最佳途徑選擇

然而，由於事先並不知道前面的最佳途徑的選擇會抵達那些中途節點，因此在動態規劃中，可使用下列步驟來尋找終點與起始點間的最佳途徑：

步驟 1：在圖 5-2 中，將位於同一斜線（即虛線）的節點歸屬於同一層（level），並依據各層與起始點的距離之遠近，給予各層編號；以圖 5-2 為例，則各層之節點如下為：

- L1 : B, E
- L2 : C, F, I
- L3 : D, G, J, M
- L4 : H, K, N
- L5 : L, O
- L6 : P

再加上前進方向的限制條件，可發現所有可能途徑會依序穿過各層的某一節點，關於限制條件將在 5.2.3 小節中做詳細的說明。

步驟 2：依照各層的號碼順序，計算每一層中各節點到達起始點 A 所須付出的最少成本。這項最少成本的計算，可利用前一層的結果來推算，以簡化運算數量，其方法如下：假設節點 Y 為第 $n+1$ 層的節點，而節點 X 為節點 Y 在第 n 層中的某一個相鄰節點。則由起始點 A 到節點 X 的最少成本若為 $f(A, X)$ ，而由節點 X 到節點 Y 的成本以 $d(X, Y)$ 表示，則由節點 Y 到達起始點 A 所付的最少成本為

$$f(A, Y) = \min_X (f(A, X) + d(X, Y)) \quad (5-3)$$

這項計算所得的最少成本，將其標註在節點 Y 上。經此項計算，可逐層得到從各節點抵達起始點 A 的最少成本。為說明這項結果，以第 3 層的節點 J 為例，在第 2 層與節點 J 相鄰的節點有 I 及 F，而從

I 及 F 到達 A 所付的最少成本分別為 8 及 11。從圖 5-2 中，可知道從節點 J 到達節點 I 及節點 F 的成本分別為 8 及 3，所以(5-3)式中的計算便是從(8+3=11)及(11+3=14)中選取最小值，而這個最小值便是節點 J 到節點 A 的最少成本。依此類推，計算到第六層的終點 P 時，便可得到從節點 A 到達節點 P 所付出的最少成本為何。

在計算(5-3)式的最小值時，同時記錄在上一層中，那一個 Y 的相鄰節點產生這項最小值，亦即：

$$\mathbf{J}_n(Y) = \arg \min_X [\mathbf{f}_{n-1}(A, X) + d(X, Y)] \quad (5-4)$$

在節點 J 到節點 A 最少成本計算的例子中，可得到 $\mathbf{J}_3(J) = I$ 。

步驟 3：再由 $\mathbf{J}_n(Y)$ 的記錄中(若以圖 5-2 為例， $n = 6, 5, \dots, 2, 1$)，逐層地倒推回去，便可求得最佳途徑。

使用上述之演算法，僅需逐層計算各節點與起始點的最少成本並記錄前一層節點的代碼；其計算量恰為節點數目減掉 1，在本例中為 15。當格狀網路的外圍每邊邊長的分割線段數為 n 時，這種方法與前述的土法練鋼法在計算量上的比值可表示成：

$$S = \frac{(n^2 - 1)}{(2n)! / (n! n!)} \quad (5-5)$$

以 $n=5$ 為例，土法練鋼的搜尋途徑數量($S=252$)是上述搜尋法($S=24$)的 10.5 倍。

5.2.2 動態時間校準(Dynamic Time Warping : DTW)

在上一小節中已討論如何使用動態規劃的方法以快速計算兩點間的最佳途徑；接著將使用相同的方法來解決語音圖樣比對過程中，時間校準的問題。假設有兩組屬於同一字音的語音訊號，由於說話速度、習慣等因素，造成訊號長度的不一致，而從這兩組訊號擷取的特徵向量數目也必然會有所差異。可將這兩組特徵向量的時間校準問題描述如下：令這兩組向量分別為 $R = (u_1, u_2, \dots, u_M)$ 及 $T = (v_1, v_2, \dots, v_N)$ ，其中 M 及 N 為這兩組向量的長度，而 u_i 及 v_i 代表特徵參數的第 i 個特徵向量，也就是第 i 個分析框所得之特徵參數，由於兩組特徵向量屬於同一語音，因此在經適當的時間校準後，兩組向量的距離必然十分相近。若以數學式表示，其時間軸校準函數，可以數列之方式表示成：

$$P = c(1), c(2), \dots, c(k), \dots, c(K) \quad (5-6)$$

其中， $c(k) = [u(k), v(k)]$ ， $k = 1, 2, \dots, K$ (參考圖 5-4)。

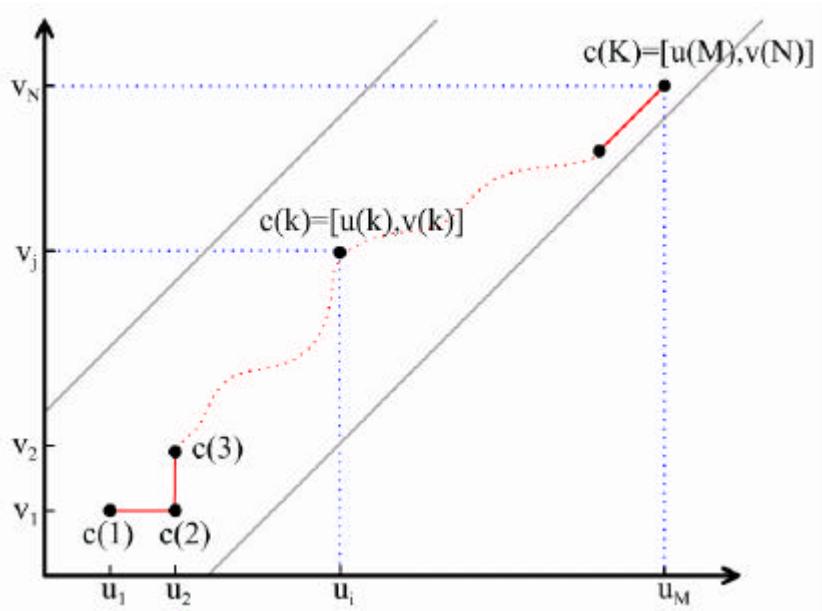


圖 5-4 動態時間軸校準函數示意圖

經動態時間校準後，兩特徵向量間的距離函數可表示成：

$$D(P) = \sum_{k=1}^K d(c(k)) = \sum_{k=1}^K d(u(k), v(k)) \quad (5-7)$$

而最佳之動態時間軸校準函數，為使兩特徵向量間之距離值為最小之途徑，故可將其定義為：

$$D(R, T) = \min_P [D(P)] = \min_P [\sum_{k=1}^K d(c(k))] \quad (5-8)$$

若將 R 及 T 兩組特徵向量的時間指標分別做為 X 及 Y 軸的座標，可得到如圖 5-4 之平面圖。令起始時間指標為 1，而兩組特徵向量的結束時間指標分別為 M 及 N ，在圖 5-4 中，從座標點 $(1,1)$ 到座標點 (M,N) 間的各個途徑，代表不同的時間校準函數，沿著某一條途徑上各點之座標，便代表兩組向量間的一種時間對應關係，依照不同的時間對應關係，可計算出兩向量間之距離，而對應於最佳時間校準函數的途徑，會使得兩向量之距離為最小。從以上的描述，可發現兩組特徵向量的時間校準問題與前一節中的最佳途徑搜尋是屬於同一類型的問題，故可利用動態規劃之方法來求解動態時間校準法之最佳解。

在最佳途徑尋找過程中，必須逐層計算從起始點到各點之最少成本途徑，而在時間校準問題上，成本便相當於對應向量間的距離，所以需要計算從起始點 $(1,1)$ 到任一點 (i,j) 的最佳途徑，而此途徑即代表兩向量的時間對應關係，再計算出 u_1, u_2, \dots, u_i 及 v_1, v_2, \dots, v_j 間的距離，即可得知在動態時間校準後兩特徵向量間的相差距離。

5.2.3 校準函數的限制條件

校準函數可視為語音訊號的時間刻度變動模型，所以此函數必然會符合實際語音訊號的特性，而一般語音訊號的特性包括：連續性(continuity)、單調性(monotonic)以及轉音速度上的限制等等；而這些特性都可針對校準函數加入對等的數學式條件來達成，如此一來可免去很多不必要或不可能存在搜尋路徑；而這些數學式條件包括：

(1)邊界條件(Boundary condition)：

$$\begin{cases} u(1) = u_1, & u(K) = u_M \\ v(1) = v_1, & v(K) = v_N \end{cases} \quad (5-9)$$

(2)最大相鄰框數條件(Adjustment window condition)：

$$|u(k) - v(k)| \leq g \quad (5-10)$$

其中， g 是兩比對樣本間可容許放大或壓縮的最大分析框數量。

(3)連續性限制條件(Continuity condition)：

$$\begin{cases} u(k) - u(k-1) \leq 1 \\ v(k) - v(k-1) \leq 1 \end{cases} \quad (5-11)$$

加入以上兩限制條件的結果，相當於限制校準函數中的搜尋途徑必須符合以下之規則：

$$c(k-1) = \begin{cases} [u(k), v(k)-1] \\ [u(k)-1, v(k)-1] \\ [u(k-1), v(k)] \end{cases} \quad (5-12)$$

圖 5-5 為此搜尋規則之說明圖。也就是說為了符合語音訊號的連續性，所以可允許經過 I 之節點為 F、E 以及 H，其餘的節點皆是不允許直接跳躍到 I 點，如 D 到 I 即為違法之搜尋途徑。

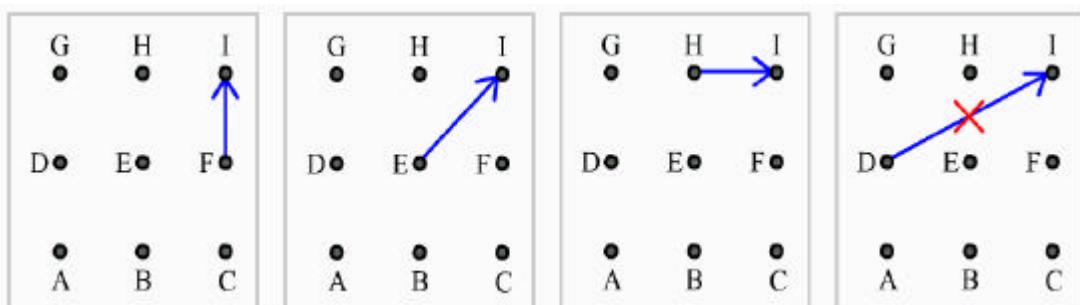


圖 5-5 符合連續性之搜尋途徑

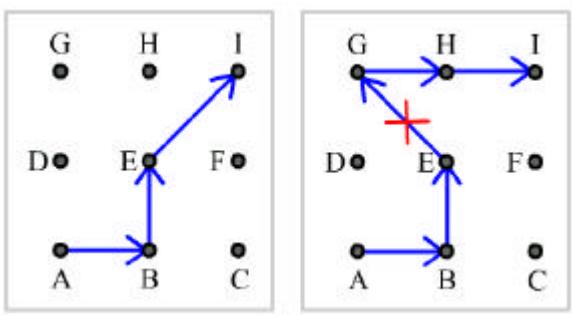


圖 5-6 符合單調性之搜尋途徑

(4)單調性限制條件(Monotonic condition)：

$$\begin{cases} u(k-1) \leq u(k) \\ v(k-1) \leq v(k) \end{cases} \quad (5-13)$$

也就是說在途徑搜尋的過程中，不允許發生逆向搜尋的情況；如圖 5-6 所示，左小圖為合法之搜尋途徑，右小圖則為非法搜尋途徑。

(5)局部限制條件(Local condition)：

為了避免校準函數發生局部過度壓縮或放大的現象，所以對於局部的搜尋途徑加入了額外的限制條件。為了有效的表示局部搜尋途徑，在此先定義搜尋途徑表示符號：

$$P_i \rightarrow (\mathbf{a}_1^i, \mathbf{b}_1^i), (\mathbf{a}_2^i, \mathbf{b}_2^i), \dots, (\mathbf{a}_L^i, \mathbf{b}_L^i) \quad (5-14)$$

其中， (\mathbf{a}, \mathbf{b}) 代表到達下一節點的相對位移點數。圖 5-7 即為所要介紹之局部限制條件途徑，其有效途徑有三個，利用(5-14)式之表示符號可表示成：

$$\begin{aligned} P_1 &\rightarrow (1,0), (1,1) \\ P_2 &\rightarrow (1,1) \\ P_3 &\rightarrow (0,1), (1,1) \end{aligned} \quad (5-15)$$

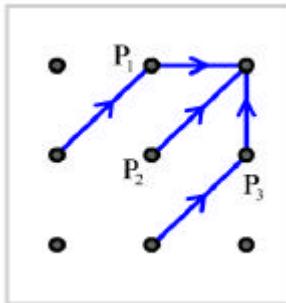


圖 5-7 局部限制途徑

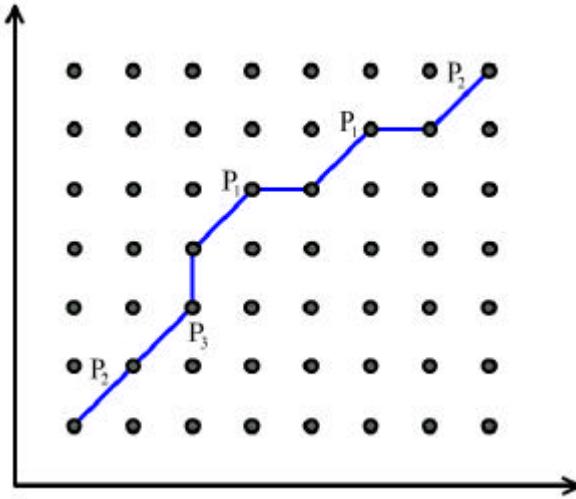


圖 5-8 加入局部限制條件之搜尋途徑

圖 5-8 為加入局部限制條件後所得之校準函數途徑圖，利用以上所介紹之路徑表示法，則該搜尋途徑可表示成：

$$P \rightarrow P_2 P_1 P_3 P_2 \quad (5-16)$$

(6)整體途徑限制條件(Global path condition)：

由於局部限制條件的關係，使得在搜尋平面上某些區域是不可能包含在最佳校準函數的途徑中，所以可根據局部限制條件中，所允許的最大壓縮率以及最大放大率來定出，最佳校準函數之途徑可能存在的有效區域。而最大放大率與壓縮率之公式別為：

$$E_{\max} = \operatorname{Max}_i \left[\left(\sum_{l=1}^L \mathbf{b}_l^i \right) / \left(\sum_{l=1}^L \mathbf{a}_l^i \right) \right] \quad (5-17)$$

$$E_{\min} = \operatorname{Min}_i \left[\left(\sum_{l=1}^L \mathbf{b}_l^i \right) / \left(\sum_{l=1}^L \mathbf{a}_l^i \right) \right] \quad (5-18)$$

若以圖 5-7 之局部限制途徑為例，則 $E_{\max}=1/E_{\min}=2$ 。根據(17)與(18)式，即可更進一步定出有效之區域邊界：

$$1 + E_{\min} \times [u(k) - 1] \leq v(k) \leq 1 + E_{\max} \times [u(k) - 1] \quad (5-19)$$

$$M + E_{\max} \times [u(k) - N] \leq v(k) \leq M + E_{\min} \times [u(k) - N] \quad (5-20)$$

若系統所採用的局部限制條件為圖 5-7 所介紹之途徑，再加上邊界條件與最大相鄰框數等兩條件，則兩語音特徵參數樣本的最佳時間校準函數之有效區，可縮小為圖 5-9 所示之灰色區域中。

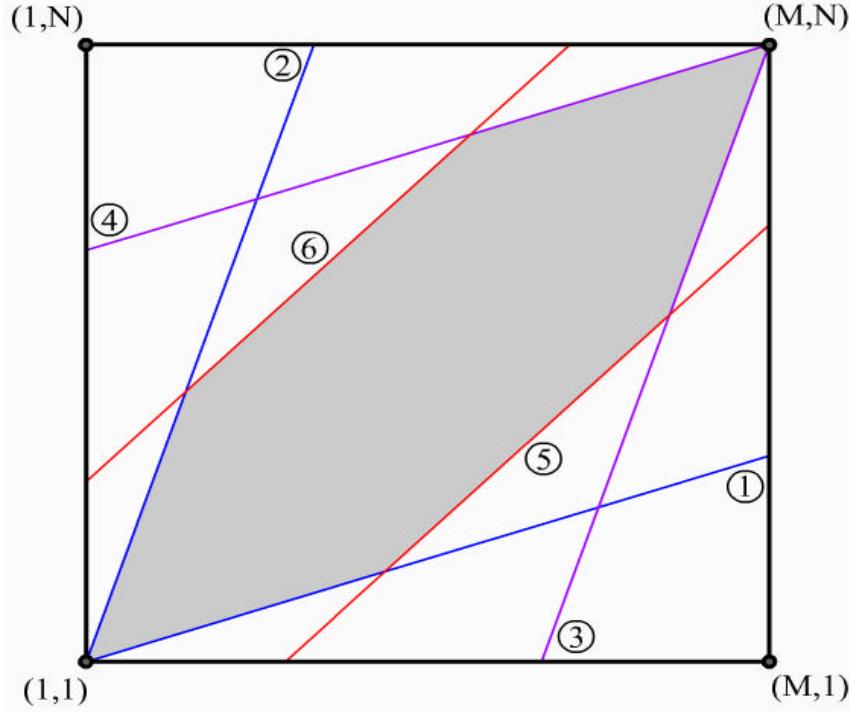


圖 5-9 最佳時間校準函數之有效區域

其中邊界線 1 與 2 由(5-19)式而得，邊界線 3 與 4 由(5-20)式而來，邊界線 5 和 6 則為(5-10)式而得，各邊界線之實際數學式如下：

$$1 + \frac{1}{2}[u(k) - 1] \leq v(k) \quad (5-21)$$

$$v(k) \leq 1 + 2[u(k) - 1] \quad (5-22)$$

$$M + 2[u(k) - N] \leq v(k) \quad (5-23)$$

$$v(k) \leq M + \frac{1}{2}[u(k) - N] \quad (5-24)$$

$$u(k) - v(k) \leq g, \quad u(k) \geq v(k) \quad (5-25)$$

$$v(k) - u(k) \leq g, \quad u(k) < v(k) \quad (5-26)$$

以上所介紹的六個限制條件中，(5-21)和(5-22)是針對有效搜尋範圍所加入之條件，而(5-23)和(5-24)則是對於相鄰節點可搜尋之方向所加入之限制條件。雖然前面四個限制條件已能使得校準函數符合語音的特性，然而為了更進一步的找到最佳之校準函數途徑，以避免局部發生過度壓縮或放大的錯誤，所以再加入兩個額外的限制條件(5-25)和(5-26)，其中條件(5-25)乃針對局部校準途徑所加入之限制條件，條件(5-26)則為從局部限制條件中所推演出來的整體有效搜尋途徑之限制範圍。

5.2.4 最佳校準函數的選擇

從前一小節中的討論，可以明確的知道最佳校準函數的可能搜尋途徑區域(如圖 5-9 所示)，則從起始點(1,1)到中途點(i,j)的最短距離可以透過遞迴公式求得：

$$D(1,1) = d(u_1, v_1) \quad (5-27)$$

$$D(i, j) = \text{Min}[D(i-1, j), D(i, j-1), D(i-1, j-1)] + d(u_i, v_j) \quad (5-28)$$

其中， $i=1,2,\dots,M$ ， $j=1,2,\dots,N$ 。然而在連續性限制條件中，若校準函數途徑選擇向右上方前進之途徑時，會造成各個可能途徑所穿過的座標點數不一定相同的現象，使得沿右上方前進的途徑，其距離計算的次數會比較少；舉例而言，在圖 5-10 中，左小圖之途徑穿過 12 個座標點，而右小圖途徑只穿過 10 個座標點，這種現象會使得沿右上方前進的途徑較容易成為最佳途徑；亦即會較可能得到線性的時間校準函數；為避免這種情況發生，一般都對各前進方向路徑予以不同的加權，所以(5-28)式須加以修正：

$$D(i, j) = \text{Min} \left\{ \begin{array}{l} D(i-1, j) + d(u_i, v_j) \\ D(i, j-1) + d(u_i, v_j) \\ D(i-1, j-1) + 2d(u_i, v_j) \end{array} \right\} \quad (5-29)$$

計算 $D(i, j)$ 後並記錄之，亦即記下沿該途徑抵達 (i, j) 的前一點之座標值，依此類推，逐層計算抵達各座標點的最佳途徑，直到終點 (M, N) 。再由 $D(M, N)$ 的記錄遵守局部限制條件倒推回始點，即可得到最佳時間校準函數。

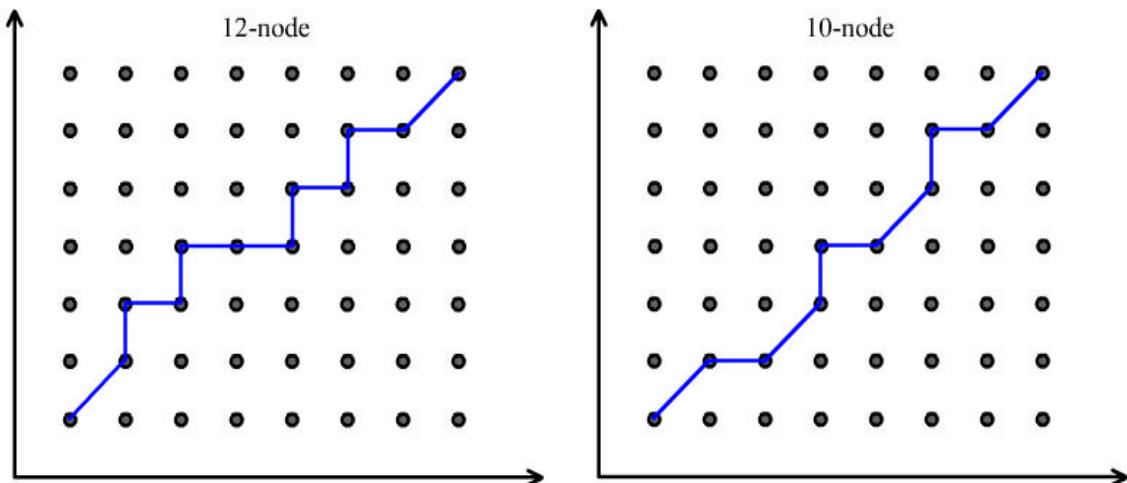


圖 5-10 穿過不同節點數目之校準函數途徑範例

以上所介紹的方法，便稱為動態時間校準法。利用這種方法，可計算兩組長短不同的語音特徵向量之相似程度(或距離)；假設已經事先建好各辨識語音的樣板圖樣資料庫，則所謂的語音辨識，便是使用動態時間校準方法，計算其特徵向量與各樣板圖樣資料庫中的語音特徵向量彼此間之距離，再選擇最接近者，即可得到語音辨識之結果。在下一節中，將介紹如何自一組訓練語音訊號中產生各辨識語音的樣板圖樣特徵向量。

5.3 語音樣板資料庫的建立

使用動態時間校準的方法進行語音辨識，須事前建好各辨識字彙的樣板特徵向量，也就是所謂的語音樣板資料庫。由所收集的語音訊號產生各辨識字彙的樣板特徵向量，稱之為樣板訓練(Training)，而樣板的訓練可分為特定語者及不特定語者兩類，這是針對訓練後所得到的樣板是否僅能代表某些說話者的語音特性，或對一般說話者也能適用來做分類，在這節中僅就特定語者之語音樣板訓練做詳細的說明。

針對特定語者的樣板訓練，可要求使用者對各辨識字彙進行錄音，再由錄音訊號中，擷取其特徵向量，直接做為各字彙的比對樣板；此種訓練方式非常簡單，然而所得到的樣板較不可靠，因其易受錄音過程的影響，而造成辨識率的不高；最主要的因素是這種方法產生的樣板無法表現同類語音訊號的多變性。要克服此上問題最簡單的方法，便是對各辨識字彙進行多次錄音以產生多個比對樣板，因此同類語音的特性(如說話速度)可出現在不同樣板上，此方法雖然克服了樣板可信度的問題，但如此一來會使系統在進行辨識時增加特徵向量的比對次數，造成系統辨識時間變長。

為避免增加比對次數並提高各字彙單一樣板的代表性，可使用平均的方法來產生比對樣板；其做法為要求使用者對同一字彙做多次錄音，再由錄音訊號中，先取出兩組錄音訊號並擷取其特徵向量，接著計算兩特徵向量的平均結果，計算好平均結果後，再取同一字彙錄音訊號，並與前一次計算的結果，再做一次平均化的處理，依此類推，直到所有的錄音訊號皆處理完畢，則最後的結果即為所求之語音樣板。

假設兩組向量分別為 $R_1 = (u_1^1, u_2^1, \dots, u_{M_1}^1)$ 及 $R_2 = (u_1^2, u_2^2, \dots, u_{M_2}^2)$ ，首先對其進行動態時間校準處理，以得到兩特徵向量的時間校準函數 $P(k) = (R_1(k), R_2(k))$ ，其中 $k = 1, 2, \dots, K$ ， K 為經校準後共同時間的結束點。接著再由下式逐一計算出樣板 $R = (u_1, u_2, \dots, u_K)$ 在每一時間點的特徵向量：

$$u_k = \frac{u^1(k) + u^2(k)}{2} \quad (5-30)$$

經上式處理後所求得之特徵向量，即可做為該字彙之語音資料庫樣板。

第六章 倒傳遞類神經網路模型

6.1 簡介

在前一章中，已經介紹利用傳統的動態時間校準法(DTW)做為辨識架構之語音辨識系統；在本章中將詳細的介紹另一種廣被應用於很多領域的演算法，也就是倒傳遞類神經網路(Backpropagation Neural Network : BPNN)。在類神經網路的領域中，經過學者們不斷的研究與改良，目前被提出來的架構已相當的多，而在此有關類神經網路的深入探討並非本書的主要目的；為了能將倒傳遞類神經網路應用到語音辨識系統中，故在此採用多層感知機(Multiple-Layer Perceptron : MLP)做為系統的辨識架構，並以倒傳遞演算法(Backpropagation Algorithm : BPA)做為訓練法則，此一系統稱之為倒傳遞類神經網路模型。

如圖 6-1 所示為一完整之語音辨識系統，首先錄音訊號經端點偵測，擷取出有聲段之訊號，再經一前置強波處理，將高頻訊號之能量放大，接著求得該有聲段之特徵向量參數(或稱之為圖樣)，以做為訓練類神經網路參數時使用，系統經適當的訓練，直到參數收斂到所設定的要求，即可將訓練後的類神經網路系統，拿來做語音辨識的工作；只要把新的語音訊號，擷取其特徵向量後，直接傳入已訓練完成之系統中，再由其輸出之數據，即可獲知語音辨識之結果。

關於類神經網路中多層感知機(MLP)之架構將在第二節中說明；而用來訓練多層感知機的倒傳遞訓練演算法(BPA)，將安排在第三節中介紹；最後第四節則用以說明如何利用倒傳遞演算法來訓練多層感知機，以及其他重要的經驗法則。

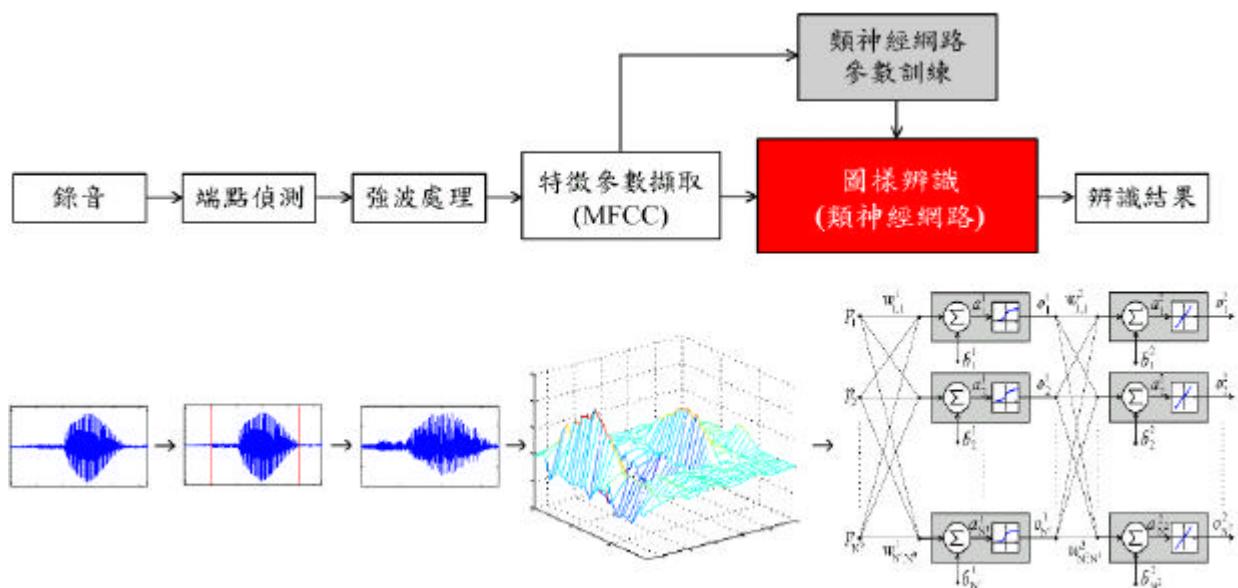


圖 6-1 倒傳遞類神經網路之語音辨識系統

6-2 多層感知機之架構(Multi-Layer Perceptron : MLP)

感知機(perceptron)是最早被提出來的類神經網路模型，然而此模型卻無法解決互斥或(exclusive OR : XOR)的問題[17]，其中 Minsky(人工智慧的創始人之一)更以專書討論：無隱藏層的感知機模型無法解決互斥或的問題，即使勉強加入隱藏層，卻苦無學習演算法可以訓練網路參數，以決定適當的連結加權值，僅能用人為的方式，以嘗試錯誤的方法，巧妙地設定其值。直到倒傳遞演算法被提出來後，被應用到多層感知機架構中，不僅可以解決互斥或的問題，而且也找到了有效的訓練法則，使得倒傳遞類神經網路被廣泛的應用到很多領域中。

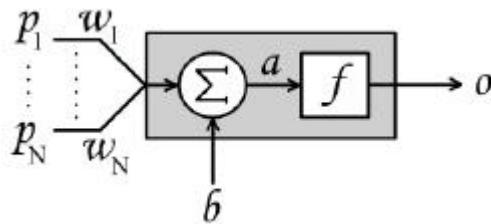


圖 6-2 神經元模型

為了能將倒傳遞類神經網路應用到語音辨識系統上[18]，所以在本節中先介紹多層感知機的網路架構模型[19]，關於倒傳遞演算法則留到下一節再詳細的說明。如圖 6-2 所示為類神經網路架構中最基本的單位：神經元(neuron)。神經元的輸出值可表示成：

$$o = f(a) = f\left(\sum_{i=1}^N w_i p_i + b\right) \quad (6-1)$$

其中， o 代表輸出值， $f(\cdot)$ 為轉換函數(transfer function)， s 為集成函數值(summation function)， w 為連結加權值， b 為偏移量， p 為輸入值。而早期所謂的感知機，便是由數個神經元所組合而成的網路架構，如圖 6-3 所示即為單層感知機模型，為了方便表示，所以以矩陣的方式來表示此網路模型：

$$O = f(WP + B) \quad (6-2)$$

其中， W 為連結加權值矩陣， P 為輸入向量， B 為偏移向量， O 為網路輸出向量。單層感知機網路模型，可應用於簡單的線性可分割應用場合，然而其無法克服互斥或問題的缺點，使得在應用上受到很多的限制。但經學者們不斷的研究後，提出加入隱藏層(hidden layer)的方法，使得加入隱藏層的感知機網路模型終於可以解決互斥或的問題，此種網路模型稱之為多層感知機模型(Multi-Layer Perceptron : MLP)，如圖 6-4 所示為加入一層隱藏層的感知機模型。

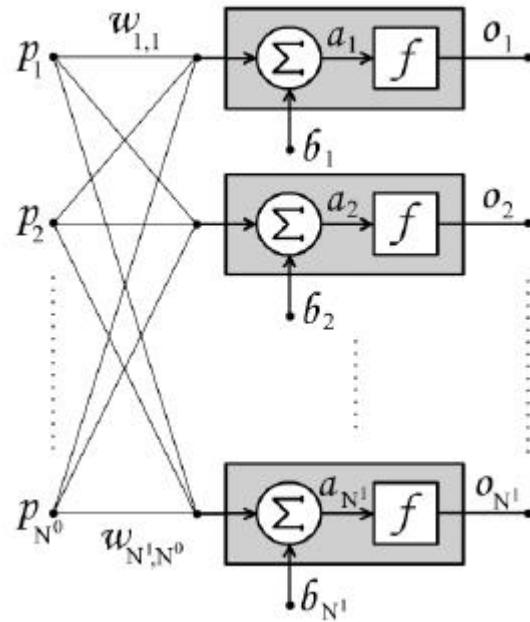


圖 6-3 單層感知機網路模型

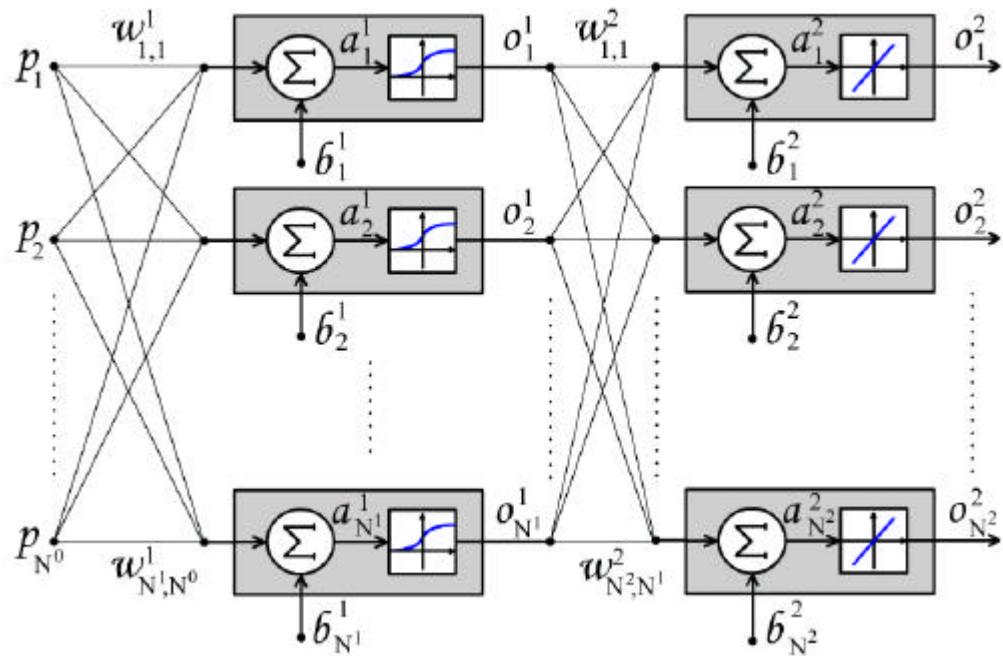


圖 6-4 兩層感知機(MLP)之網路架構

在類神經網路的相關書籍中，最令人困惑的莫過於數學符號表示法，以及網路架構的定義方式，為了方便往後的說明，表 1 列出文中關於網路架構與數學符號的定義或術語。

表 6-1 類神經網路符號與術語定義表

符號或術語	說 明
$N^0-N^1-N^2$	網路架構表示式。例：25-7-10，表示該網路為兩層網路模型之架構 ^{*1} ，由 25 個輸入點，一個含有 7 個神經元的隱藏層，以及 10 輸出點所組合而成。
P_i	第 i 個輸入參數。
a_i^l	第 l 層第 i 個神經元中集成函數的輸出值。其中輸入層視為第 0 層，從隱藏層開始算，直到輸出層(含)。
o_i^l	第 l 層第 i 個神經元的輸出值。其中輸入層視為第 0 層，從隱藏層開始算，直到輸出層(含)。
f^l	第 l 層中神經元所採用的轉換函數 ^{*2} 。
$w_{i,j}^l$	從第 $l-1$ 層的第 j 個神經元，連結到第 l 層的第 i 個神經元的加權參數值。
b_i^l	第 l 層中第 i 個神經元的集總參數偏移量。
*1：因為輸入是直接傳入隱層藏，故在文中並未把輸入層視為網路架構中的一層。所以所謂的兩層類神經網路架構，即一個隱藏層與一個輸出層(因此兩層中皆以神經元組成)。	
*2：為了使網路可應用倒傳遞演算法(最大梯度陡降法)，所以轉換函數得採用可徵分之數學函數。	

在語音辨識系統中，單層感知機是無法達成系統辨識的功能，所以得採用多層感知機的網路模型，加入隱藏層的網路模型，使得網路可以呈現輸入訊號間的交互影響，然而太多隱藏層的網路模型，不僅需要耗費很長的訓練時間，而且對辨識率亦無太多的改善，故在本書中將以兩層之感知機網路模型做為語音辨識系統之架構，如圖 6-4 所示，則整個辨識系統的數學式可表示成：

$$O^2 = f^2(a^2) = f^2(W^2 O^1 + B^2) = f^2(W^2 f^1(W^1 P + B^1) + B^2) \quad (6-3)$$

此即兩層之感知機網路模型輸出數學式，亦為文中所採用之網路架構。

接下來的問題是如何去決定各層之間，連結加權值的參數設定，在語音辨識系統中的類神經網路，是屬於監督式學習，所以可利用倒傳遞訓練演算法來訓練此系統，待系統輸出結果符合所求，即表示該網路連結加權參數已收斂到所設定之區域內，故可將此一訓練後之類神經網路做為系統的辨識器來使用。所以下一節中，將詳細的說明倒傳遞演算法的原理。

6.3 倒傳遞演算法(Back Propagation Algorithm : BPA)

多層架構網路模型之輸出與輸入間的轉換關係是非線性特性，使得輸出誤差與連結加權參數間的關係變得相當複雜，且兩者間微分關係並非是顯函數關係，所以最小均方法(Least Mean Square : LMS)無法應用到此類網路模型的訓練過程中；直到倒傳遞訓練演算法(Backpropagation Algorithm: BPA)被提出後，終於可有效解決多層網路架構缺乏學習演算法的問題，使得倒傳遞類神經網路(Backpropagation Neural Network : BPNN)成了目前應用最廣泛的網路架構模型。以下將說明倒傳遞演算法如何利用訓練樣本 P 與輸出向量 T，來修正連結加權值 W，而達到訓練的目的。以下將說明：A.倒傳遞演算法之原理；B.學習調整率之設定。

6.3.1 倒傳遞演算法之原理[20]

在倒傳遞網路中，第 n 層的第 j 個神經元之輸出值，為第 n-1 層中，所有神經元輸出值之和的非線性函數：

$$O^n = f^n(a^n) = f^n(W^n O^{n-1} + B^n) \quad (6-4)$$

其中，n=1,3,..,M，M 為架構中網路層之總數量。而監督式學習的主要目的即是要降低網路輸出值 O 與目標輸出值 T 之間的差距，所以在此先定義性能指標參數(performance index)：

$$F(x) = E[e^2] = E[(t - o)^2] \quad (6-5)$$

此即均方誤差函數(Mean Square Error)。其中，x 為神經元連結加權參數與偏移量參數，E[.] 為期望值函數，e 為輸出誤差值，t 為目標輸出值，o 為網路輸出值。為使上式更一般化，所以將其以向量的型式表示：

$$F(x) = E[e^T e] = E[(T - O)^T (T - O)] \quad (6-6)$$

再就訓練的觀點而言，可將上式誤差平方之期望值，近似成第 k 次訓練後的誤差平方：

$$\hat{F}(x) = e^T(k) e(k) = (T(k) - O(k))^T (T(k) - O(k)) \quad (6-7)$$

如此一來即可利用最大梯度下降法(steepest descent algorithm)[21][22]，做為網路訓練之依據，所以可得網路連結加權參數與偏移量參數之修正式為：

$$w_{ij}^n(k+1) = w_{ij}^n(k) - a \frac{\partial \hat{F}}{\partial w_{ij}^n} \quad (6-8)$$

$$b_i^n(k+1) = b_i^n(k) - a \frac{\partial \hat{F}}{\partial b_i^n} \quad (6-9)$$

其中，a 為學習調整率(learning rate)，此值的設定將決定系統學習速度快慢。

(6-8)與(6-9)式中的偏微分項對於單層線性網路而言，其計算是相當簡單的，然而對於多層網路而言，誤差值與連結加權值之間的關係並非為顯函數(explicit function)，使得該項偏微分之計算變的相當不易。不過誤差值為隱藏層中連結加權參數的間接函數，故可利用微積分中的連鎖律(chain rule)來求得式中偏微分項之值；利用連鎖律，則(6-8)與(6-9)式中的偏微分項可表示成：

$$\frac{\partial \hat{F}}{\partial w_{ij}^n} = \frac{\partial \hat{F}}{\partial a_i^n} \cdot \frac{\partial a_i^n}{\partial w_{ij}^n} \quad (6-10)$$

$$\frac{\partial \hat{F}}{\partial b_i^n} = \frac{\partial \hat{F}}{\partial a_i^n} \cdot \frac{\partial a_i^n}{\partial b_i^n} \quad (6-11)$$

在以上兩式中，等號右側第二項分別為連結加權參數與偏移量參數的顯函數，所以可直接求得：

$$a_i^n = \sum_{j=1}^{N^{n-1}} w_{ij}^n o_j^{n-1} + b_i^n \quad (6-12)$$

故可得： $\frac{\partial a_i^n}{\partial w_{ij}^n} = o_j^{n-1}$ 以及 $\frac{\partial a_i^n}{\partial b_i^n} = 1$ 。接著定義一靈敏度參數(sensitivity parameter)：

$$s_i^n \equiv \frac{\partial \hat{F}}{\partial a_i^n} \quad (6-13)$$

即第 n 層中第 i 個神經元的集成函數運算之值的變化，對系統性能指標的影響程度。將(6-10) (6-13)式之結果代回(6-8)與(6-9)式則可得：

$$w_{ij}^n(k+1) = w_{ij}^n(k) - \mathbf{as}_i^n o_j^{n-1} \quad (6-14)$$

$$b_i^n(k+1) = b_i^n(k) - \mathbf{as}_i^n \quad (6-15)$$

為了方便應用，故將其改用矩陣的方式來表示：

$$W^n(k+1) = W^n(k) - \mathbf{as}^n \cdot (O^{n-1})^T \quad (6-16)$$

$$b^n(k+1) = b^n(k) - \mathbf{as}^n \quad (6-17)$$

其中靈敏度參數為：

$$S^n \equiv \frac{\partial \hat{F}}{\partial a^n} = \left[\frac{\partial \hat{F}}{\partial a_1^n} \quad \frac{\partial \hat{F}}{\partial a_2^n} \quad \dots \quad \frac{\partial \hat{F}}{\partial a_{N^n}^n} \right]^T \quad (6-18)$$

以上即為倒傳遞訓練演算法學習過程中，參數調整式的詳細推導；從(6-16)與(6-17)式可知，剩下的工作便是如何計算靈敏度參數值 S^n 。

為求得靈敏度參數值 S^n ，利用 Jacobian 矩陣來推導其疊代關係，Jacobian 矩陣為：

$$\frac{\partial a^{n+1}}{\partial a^n} = \begin{bmatrix} \frac{\partial a_1^{n+1}}{\partial a_1^n} & \frac{\partial a_1^{n+1}}{\partial a_2^n} & \cdots & \frac{\partial a_1^{n+1}}{\partial a_{N^n}^n} \\ \frac{\partial a_2^{n+1}}{\partial a_1^n} & \frac{\partial a_2^{n+1}}{\partial a_2^n} & \cdots & \frac{\partial a_2^{n+1}}{\partial a_{N^n}^n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial a_{N^{n+1}}^{n+1}}{\partial a_1^n} & \frac{\partial a_{N^{n+1}}^{n+1}}{\partial a_2^n} & \cdots & \frac{\partial a_{N^{n+1}}^{n+1}}{\partial a_{N^n}^n} \end{bmatrix} \quad (6-19)$$

式中的第 i 列 j 行項的值為：

$$\frac{\partial a_i^{n+1}}{\partial a_j^n} = \frac{\partial \left[\sum_{l=1}^{N^n} w_{il}^{n+1} o_l^n + b_i^{n+1} \right]}{\partial a_j^n} = w_{ij}^{n+1} \cdot \frac{\partial o_i^n}{\partial a_j^n} = w_{ij}^{n+1} \cdot \frac{\partial f^n(a_j^n)}{\partial a_j^n} = w_{ij}^{n+1} \cdot \dot{f}^n(a_j^n) \quad (6-20)$$

其中 $\dot{f}^n(a_j^n) = \frac{\partial f^n(a_j^n)}{\partial a_j^n}$ ，所以 Jacobian 矩陣可重寫成：

$$\begin{aligned} \frac{\partial a^{n+1}}{\partial a^n} &= \begin{bmatrix} w_{11}^{n+1} \dot{f}^n(a_1^n) & w_{12}^{n+1} \dot{f}^n(a_2^n) & \cdots & w_{1N^n}^{n+1} \dot{f}^n(a_{N^n}^n) \\ w_{21}^{n+1} \dot{f}^n(a_1^n) & w_{22}^{n+1} \dot{f}^n(a_2^n) & \cdots & w_{2N^n}^{n+1} \dot{f}^n(a_{N^n}^n) \\ \vdots & \vdots & \ddots & \vdots \\ w_{N^{n+1}1}^{n+1} \dot{f}^n(a_1^n) & w_{N^{n+1}2}^{n+1} \dot{f}^n(a_2^n) & \cdots & w_{N^{n+1}N^n}^{n+1} \dot{f}^n(a_{N^n}^n) \end{bmatrix} \\ &= \begin{bmatrix} w_{11}^{n+1} & w_{12}^{n+1} & \cdots & w_{1N^n}^{n+1} \\ w_{21}^{n+1} & w_{22}^{n+1} & \cdots & w_{2N^n}^{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N^{n+1}1}^{n+1} & w_{N^{n+1}2}^{n+1} & \cdots & w_{N^{n+1}N^n}^{n+1} \end{bmatrix} \begin{bmatrix} \dot{f}^n(a_1^n) & 0 & \cdots & 0 \\ 0 & \dot{f}^n(a_2^n) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \dot{f}^n(a_{N^n}^n) \end{bmatrix} \\ &= W^{n+1} \dot{F}^n(a^n) \end{aligned} \quad (6-21)$$

利用連鎖律將(6-13)式展開，並把(6-21)式之結果代入，即可獲得靈敏度參數 S^n 與 S^{n+1} 之間的疊代關係式：

$$S^n \equiv \frac{\partial \hat{F}}{\partial a^n} = \left[\frac{\partial a^{n+1}}{\partial a^n} \right]^T \cdot \frac{\partial \hat{F}}{\partial a^{n+1}} = \dot{F}^n(a^n) \cdot (W^{n+1})^T \cdot S^{n+1} \quad (6-22)$$

(6-22)式即為靈敏度參數的疊代公式，也是倒傳遞演算法(BPA)名字的由來。而其開始進行疊代的第一個靈敏度參數為最後一層，即網路輸出層的靈敏度參數 S^M 之值，可求得：

$$S_i^M = \frac{\partial \hat{F}}{\partial a_i^M} = \frac{\partial [(T - O)^T (T - O)]}{\partial a_i^M} = \frac{\partial [\sum_{j=1}^{N^M} (t_j - o_j)^2]}{\partial a_i^M} = -2(t_i - o_i) \frac{\partial o_i}{\partial a_i^M} \quad (6-23)$$

在(6-23)式中， $\frac{\partial o_i}{\partial a_i^M} = \frac{\partial o_i^M}{\partial a_i^M} = \frac{\partial f^M(a_i^M)}{\partial a_i^M} = \dot{f}^M(a_i^M)$ ，所以該式可再重新表示成(6-24)式，而(6-25)式為其矩陣表示式：

$$s_i^M = -2(t_i - o_i)\dot{f}^M(a_i^M) \quad (6-24)$$

$$S^M = -2\dot{F}^M(a^M) \cdot (T - O) \quad (6-25)$$

以(6-25)式求得最後一層之靈敏度參數值，並利用(6-22)式之疊代公式求得各網路層中的靈敏度參數 S^n ，再透過(6-16)與(6-17)式之網路連結參數與偏移量參數的修正公式，即可使多層架構之網路達到學習的目的。

6.3.2 學習調整率之設定

在(6-16)與(6-17)式中，學習調整率 a 值的設定將決定系統學習的速度，太大或太小的值($a > 0$)對系統均不利， a 愈大則對連結加權參數的修正量也愈大，如此可加快系統逼近性能指標的最小值；但太大的 a 值將會導致連結加權參數修正幅度過大，而造成系統性能指標參數值呈現振盪現象，難以收斂。在一般的應用中， a 值皆設定為常數值($0 < a \leq 1$)[23]，使得系統從開始到訓練結束皆採用相同的學習調整率，然而大多數系統在開始訓練時，往往需要較大的 a 值，經過幾回的訓練之後，為了使系統漸漸逼近性能指標的最小值，此時需要的便是較小的 a 值，所以在本文中，將 a 值設計成隨訓練次數的增加而衰減的型式，其數學式如下：

$$a_n = a \times (1 + A \cdot e^{-\frac{n}{T}}) \quad (6-26)$$

其中， a_n 為修正後第 n 次訓練時的學習調整率之值， n 為訓練次數， a 則為原本所設定之常數值學習調整率， A 為使用者希望系統開始訓練時之學習速率為一般學習速率的倍數， T 則為衰減指數，若希望系統在第 k 次就回復到一般的學習速度時，則 $T=k/5$ 。

系統採用(6-26)式之後，使系統在訓練初期的修正幅度較大，隨著訓練次數的增加，系統的學習調整率相反地會減小，如此一來，就可克服為加快系統學習速度，而採用較大之 a 值，所造成的振盪現象；也可減少因採用較小之 a 值，所需之訓練次數。

6.4 辨識系統訓練方法(Training Procedure)

在進行倒傳遞類神經網路訓練時，可分成三個步驟：

1.順向計算每一層網路的輸出值

$$O^n = f^n(a^n) = f^n(W^n O^{n-1} + B^n) \quad (6-27)$$

其中， $O^0 = P$ (即輸入之樣本)， $O^M = T$ (即輸出向量)，利用(6-27)式便可逐層求得各層網路神經元的輸出值。

2.逆向計算每一層的靈敏度參數值

$$S^n = \dot{F}^n(a^n) \cdot (W^{n+1})^T \cdot S^{n+1} \quad (6-28)$$

其中， $n=M-1, \dots, 2, 1$ ，而最後一層的靈敏度參數值為 $S^M = -2\dot{F}^M(a^M) \cdot (T - O)$ 根據(6-28)式從最後一層逐層往前推算，即可得各層之靈敏度參數值。

3.利用最大梯度下降法之修正式調整連結參數與偏移量參數值

$$W^n(k+1) = W^n(k) - \alpha s^n \cdot (O^{n-1})^T \quad (6-29)$$

$$b^n(k+1) = b^n(k) - \alpha s^n \quad (6-30)$$

利用步驟 1 和步驟 2 所求得 O^n 與 S^n 之值，代入(6-29)與(6-30)式之參數調整式中，即可逐層修正網路連結加權參數與偏移量參數之值。

如此重覆上述三個訓練步驟，便可使系統達到訓練的目的。為了避免系統永無止境的訓練，所以需要有停止訓練的條件，一般的做法有兩種，第一個是指定一個固定次數的訓練週期後即停止訓練，另一個則是設定性能指標值，系統反覆訓練直到系統的性能指標之值達到要求時才停止訓練。然而第二種方法，若參數設定不當，往往會陷入無窮迴圈的訓練中，所以在此採用採雙管齊下的方式，同時指定以上兩個終止訓練條件，當有一個條件滿足時，系統即可停止訓練。

要將倒傳遞類神經網路應用到語音辨識中時，系統的訓練過程根據上面所述三大步驟進行即可；然而實際的經驗顯示，不當的安排訓練樣本次序，將導致系統訓練時間增長，同時降低系統的辨識率；以非特定語者之中文數字語音辨識系統而言，若訓練時先全部訓練數字 0 的所有樣本，接著再訓練數字 1 的所有樣本，依此類推，直到全部的樣本都訓練過；如此之訓練安排，很容易使系統落入局部最小值的區域，甚至經過很長的訓練時間，系統亦無法達到所設定之辨識率的要求。

為了克服上述之問題，有人提出以隨機取樣的方式來決定訓練樣本之順序，這的確也是個好方法，然而為了簡化訓練流程，所以在文中並不採用此法，而是改以逐人逐字之方式來取代，所謂逐人逐字，就是先訓練第 1 個人的第 1 個數字 0 的樣本，接著訓練第 1 個數字 1 的樣本，依此類推直到數字 9；接著換第 2 個人的重覆同樣的訓練，直到所有人的第 1 個 0~9 數字樣本都訓練過了，再換訓練所有人的第 2 個 0~9 數字樣本，依此類推，直到所有的數字樣本都訓練過後，檢查系統的辨識率，當符合所需之標準時，即可停止訓練。

6.5 神經元移除法則

在倒傳遞類神經網路的實際應用時，將會發現神經元數量太多將會佔用很多的系統資源(記憶體)，而且會增加系統的計算量。對個人電腦而言，不管是運算速度或是記憶體大小，都足以應付以倒傳遞類神經網路所設計之語音辨識系統；然而對於 DSP 來說，估且不考慮運算速度，其本身所能支援的記憶體相當有限，使得採用太多神經元數目之模型，成為系統轉移到 DSP 平台上的一大瓶頸。

經由實驗發現，當辨識系統經過訓練後(辨識率高於 90%)，連結到某些神經元的加權值變動範圍變的很小。為降低系統記憶體的需求量，針對此一現象將隱藏層中，影響權重不是很重要的次要神經元予以移除，以減少神經元數量；首先定義神經元連結加權值之振幅：

$$A_i^l = \text{Max}_j[w_{i,j}^l] - \text{Min}_j[w_{i,j}^l] \quad (6-31)$$

其中 l 代表第 l 層， i 代表第 i 個神經元；接著設定移除的臨界值：

$$Thd_l = C \times \text{Max}_i[A_i^l] \times \text{Max}_i[A_i^{l+1}] \quad (6-32)$$

其中 C 為神經元移除係數(20%~35%)，係數愈大移除的神經元數愈多，係數愈小則較不會影響到原系統之辨識率，但可移除的神經元相對的也會變少。

而神經元移除之原則為：當第 l 層中的神經元連結加權值之值域小於 Thd_l 時，便將該神經元移除，保留剩下之神經元為系統所用，從實驗中發現移除後之系統其辨識率會減低，但若對系統進行再訓練，則其辨識率亦會再回增。如此一來，將系統在個人電腦上開發完成後，將次要之神經元予以移除，再將剩下之神經元轉移到 DSP 平台上，即可應用於內建式語音指令辨識器。

第七章 實驗方法

7.1 前言

本論文以倒傳遞類神經網路(Backpropagation Neural Network : BPNN)為架構設計一非特定語者之中文數字語音辨識系統。首先針對動態時間校準法(DTW)與倒傳遞類神經網路演算法，分析與比較兩者在性能上的優劣，進而證明倒傳遞類神經網路模型，更適於轉移到數位處理器(DSP)平台上發展。其次針對類神經網路模型，加入神經元移除法則之處理程序，再分析移除次要神經元後對辨識系統的辨識率之影響。關於倒傳遞類神經網路的設計，為使系統在現有之神經元數的條件下，能辨識更多的語音樣本數量，故針對此一問題，將訓練過程中，原本一對一的輸出神經元，改以二進位編碼對應，比較此兩不同輸出架構之系統的效能差異。關於其他更詳細的實驗設計將在 7.3 節中做說明。

7.2 語音樣本資料庫

關於本論文中，所有語音辨識系統的訓練與測試，皆採用多媒體控制實驗室之中文數字語音樣本資料庫，此語音樣本資料庫為 21 人(12 男 9 女)，0~9 每個字錄音四次，錄音環境為一般實驗室，取樣頻率為 8 kHz，單聲道 16-bit 表示之取樣點；每個錄音訊號經自動端點偵測後，擷取出有聲段訊號部份做前置強波處理，再將該有聲段字音分割成 25~35 個分析框(含分析框重疊處理)來擷取特徵參數，包括：線性頻譜參數(LFS)、線性倒頻譜參數(LFCC)、梅爾頻譜參數(MFS)以及梅爾倒頻譜參數(MFCC)等四類語音特徵參數。如表 7-1 表 7-3 即為中文數字語音樣本資料庫之詳細規格表。

表 7-1 多媒體控制實驗室中文數字語音樣本資料庫

多媒體控制實驗室中文數字語音樣本資料庫					
性別	數字(0~9)	錄音次數	人數	樣本數量	樣本總數
男性	10	各 4	15	$10 \times 4 \times 15 = 600$	840
女性	10	各 4	6	$10 \times 4 \times 6 = 240$	
附註	1.錄音環境：一般實驗室 2.錄音規格：單聲道、16-bit 格式、取樣頻率 8 kHz 3.檔案格式：Wave-file format.				

表 7-2 中文數字語音樣本頻譜特徵參數資料庫

參數	線性頻譜特徵參數(LFS)			梅爾頻譜特徵參數(MFS)		
濾波器	25-frame	30-frame	35-frame	25-frame	30-frame	35-frame
12	F25LFS12	F30LFS12	F35LFS12	F25MFS12	F30MFS12	F35MFS12
14	F25LFS14	F30LFS14	F35LFS14	F25MFS14	F30MFS14	F35MFS14
16	F25LFS16	F30LFS16	F35LFS16	F25MFS16	F30MFS16	F35MFS16
18	F25LFS18	F30LFS18	F35LFS18	F25MFS18	F30MFS18	F35MFS18
20	F25LFS20	F30LFS20	F35LFS20	F25MFS20	F30MFS20	F35MFS20
22	F25LFS22	F30LFS22	F35LFS22	F25MFS22	F30MFS22	F35MFS22
24	F25LFS24	F30LFS24	F35LFS24	F25MFS24	F30MFS24	F35MFS24

表 7-3 中文數字語音樣本倒頻譜特徵參數資料庫

參數	線性倒頻譜特徵參數(LFCC)			梅爾倒頻譜特徵參數(MFCC)		
階數	25-frame	30-frame	35-frame	25-frame	30-frame	35-frame
6	F25LFCC6	F30LFCC6	F35LFCC6	F25MFCC6	F30MFCC6	F35MFCC6
8	F25LFCC8	F30LFCC8	F35LFCC8	F25MFCC8	F30MFCC8	F35MFCC8
10	F25LFCC10	F30LFCC10	F35LFCC10	F25MFCC10	F30MFCC10	F35MFCC10
12	F25LFCC12	F30LFCC12	F35LFCC12	F25MFCC12	F30MFCC12	F35MFCC12
14	F25LFCC14	F30LFCC14	F35LFCC14	F25MFCC14	F30MFCC14	F35MFCC14
16	F25LFCC16	F30LFCC16	F35LFCC16	F25MFCC16	F30MFCC16	F35MFCC16
18	F25LFCC18	F30LFCC18	F35LFCC18	F25MFCC18	F30MFCC18	F35MFCC18

7.3 實驗設計

本論文之主要目的為分析與評估動態時間校準法(DTW)與倒傳遞類神經網路模型(BPNN)，兩不同演算法其在性能上的差異，期能從中找出適於 DSP 平台上發展之辨識架構。並針對倒傳遞類神經網路模型，做一些修正或改良，期能使以此架構為核心的辨識系統，能達到辨識速度快、辨識率高以及佔用較少的系統資源的優勢。為評估所設計之語音辨識系統的各項性能優劣，在此設計八項實驗，用以呈現各子系統中所做之修正的影響，其詳細說明如下。此八項實驗包括：

- (1).端點偵測演算法之比較；
- (2).特徵參數的選定；
- (3).學習調整率 a 對系統學習效能的影響；
- (4).非特定語者之中文數字辨識系統；
- (5).倒傳遞類神經網路輸出架構之設計；
- (6).神經元移除法則的應用；
- (7).辨識演算法 DTW 與 BPNN 之比較；
- (8).非特定語者之系統應用於特定語者之影響；

7.3.1 端點偵測演算法之比較

在獨立字音辨識系統中，精準的端點偵測是很重要的，錯誤的判斷會徒增系統的計算量並降低其辨識率，如第三章中所介紹的兩種端點偵測法：能量曲線判別法以及 R-S 端點偵測法。由於前者無法找出氣音段之訊號，而後者雖可克服此一問題，但卻易受雜訊干擾而失效；而能對抗雜訊干擾有效找出端點的演算法，其過程相當煩雜計算量大，所以本文中，結合能量曲線與 R-S 偵測法，透過雜訊干擾參數的指標，自動選擇端點偵測判別模式。

選擇錄音訊號的前 $N(3\sim 5)$ 個分析框做為背景雜訊參數值估測，求得雜訊干擾指標：

$$\text{雜訊能量} : E_n = \frac{1}{N} \sum_{i=1}^N E(i) \quad (7-1)$$

$$\text{雜訊越零率} : ZCR_n = \frac{1}{N} \sum_{i=1}^N ZCR(i) \quad (7-2)$$

再視此兩參數值選用不同之判別法，選定方式如表 7-4 所列，其中 $Max(E)$ 與 $Max(ZCR)$ 分別為能量曲線與越零率曲線的最大值。在此實驗中，分別以能量判別法、R-S 判別法、能量判別法與 R-S 判別法之混合模式以及人工端點定位等四種不同方法，比較在有雜訊與無雜訊的環境下的端點偵測性能。

表 7-4 端點偵測法選用表

雜訊指標	$ZCR_n \leq 0.1 \times Max(ZCR)$	$ZCR_n > 0.1 \times Max(ZCR)$
$E_n \leq 0.05 \times Max(E)$	R-S 端點偵測法	能量曲線判別法
$E_n > 0.05 \times Max(E)$	能量曲線判別法	能量曲線判別法

關於此實驗中，雜訊的部份採用訊號雜訊比為 20dB 的白雜訊(white noise)，加入原錄音訊號之中，再進行有聲段端點偵測之測試。

7.3.2 特徵參數的選定

對語音辨識系統而言，語音訊號特徵參數的選擇是非常重要的，可用來表示語音訊號的特徵參數有很多種，如第四章中所介紹，以數位濾波器組(多個帶通濾波器)來處理語音訊號，再將每個濾波器的頻譜能量值做參數轉換，所得到之語音訊號特徵參數包括：線性頻譜參數(LFS)、線性倒頻譜係數(LFCC)、梅爾頻譜參數(MFS)梅爾倒頻譜係數(MFCC)等。

在特徵參數選定的實驗中，以倒傳遞類神經網路(BPNN : 600-30-10)為系統之辨識核心來進行頻譜特徵參數(LFS 與 MFS)的比較，即每個有聲段分割成 30 個分析框，且每個分析框以 20 個帶通濾波器，來擷取該分析框內之語音訊號特徵參數，而系統的輸出架構採一對一的模式，即一個數字對應一個輸出神經元，所以系統將會有 10 個輸出單元。

對於倒頻譜參數(LFCC 與 MFCC)，也以倒傳遞類神經網路(BPNN : 360-30-10)為系統之辨識核心來進行實驗，此時每個有聲段亦分割成 30 個分析框，而每個分析框擷取 12 階之倒頻譜參數，系統之輸出架構同樣是採一對一的模式。

系統在訓練時，以 $\frac{2}{3}$ 之語音資料庫樣本(8 男 6 女)對系統進行訓練，再以 $\frac{1}{3}$ 的樣本(4 男 3 女)來做系統測試，對系統進行辨識率之評估。最後將四個採用不同語音特徵參數的系統之辨識率進行比較，其結果將會是往後從事語音辨識系統設計時之重要參考依據。

7.3.3 學習調整率 a 對系統學習效能的影響

在類神經網路領域中，關於訓練過程中所用到的學習調整率 a 值的選定，有很多相關的研究，一般最常用的方法便是針對特定的應用場合，選定一適當的常數 a 值。在第六章中我們提出隨訓練次數增加而衰減的變動型 a 值之設定方法。所以在此實驗中，主要測試常數型的學習調整率與變動型學習調整率，兩者對系統學習效能的影響。

在常數型的學習調整率 a 的實驗中，根據經驗， a 介於 0.001 至 0.01 為佳，若太大則系統容易發散或發生振盪的現象，若選的太小則增加系統訓練的時間，所以在此實驗中，設計兩個辨識系統，分別採用(7-3)與(7-4)之值來進行系統測試。

$$a = 0.001 \quad (7-3)$$

$$a_n = a \times (1 + e^{-n}) = 0.001 \times (1 + e^{-n}) \quad (7-4)$$

7.3.4 非特定語者之中文數字辨識系統

非特定語者之中文數字辨識系統為本論文主要研究之主題，用以探討語音辨識過程中，每次子系統對辨識性能的影響，同時針對各各環節提出改善之對策，如：變動重疊寬度之分析框處理、自動選擇之端點偵測法、辨識效果較佳之特徵參數、可加快學習速度之衰減型學習調整率的設定等等；其主要目的乃希望可設計出一簡單、辨識速度快、辨識率高、系統資源需求少之語音辨識系統，同時為下一階段目標 DSP 平台之語音辨識系統做準備。

在這個實驗中，所採用之架構為倒傳遞類神經網路(BPNN : 360-30-10) 模型，如圖 7-1 所示，實驗的過程分為兩個步驟進行：訓練與測試。在訓練時，以 $\frac{2}{3}$ 之語音資料庫樣本(8 男 6 女)對系統進行訓練，每訓練一個循環，便以 $\frac{1}{3}$ 的樣本(4 男 3 女)來對系統進行測試，評估系統之辨識率。其中關於語音樣本的特徵參數，採用 12 階之梅爾倒頻譜參數，且每個字音段分割成 30 個分析框。

為使實驗結果更客觀，所以隨機取樣 $\frac{2}{3}$ 之語音資料庫做為訓練樣本，而剩下之 $\frac{1}{3}$ 則為測試樣本，如此重覆進行三次，再將所得之辨識率結果做平均化之處理。

7.3.5 倒傳遞類神經網路輸出架構之設計

根據倒傳遞類神經網路的架構設計，關於輸出的神經元數量多寡皆取決於輸入訊號的分類數量，對於分類數量較少的系統，此法尚可行；然而對於分類數量大的系統，則此法就顯得沒有效益，故在此實驗中用以比較兩種不同的輸出架構之設計方式，其在性能上的差異。

a.一對一輸出型

在類神經網路的應用中，關於網路輸出架構的設計，大部份都採一對一的方式，也就是說當輸入訊號經系統處理後，若其輸出有 n 個不同的結果，則其輸出架構便需有 n 個神經元與之對應。則其所採用之一對一輸出型架構如圖 7-1 所示。

b.二進位編碼輸出型

除了一對一輸出架構外，可將輸出之神經元在訓練的過程，即以二進位編碼來分類其輸入之訊號，以輸出架構為 4 個神經元的系統而言，則此系統最多可區分為 16 種不同的輸出對映值。同樣以中文數字辨識系統來進行實驗，則所採用之二進位編碼輸出型架構如圖 7-2 所示，編碼後所對應的關係如表 7-5 所列。

7.3.6 神經元移除法則的應用

在第六章最後提出了神經元移除法則，對於採用該法則移除神經元後之系統進行性能測試；首先設計一辨識系統，並加以訓練，直到系統辨識率達 95% 時，再利用神經元移除法則，將次要的神經元移除，保留剩下之神經元做為系統新的架構，再對系統進行測試與訓練，最後比較兩者在辨識功能上之差異，以做為往後設計時之參考。

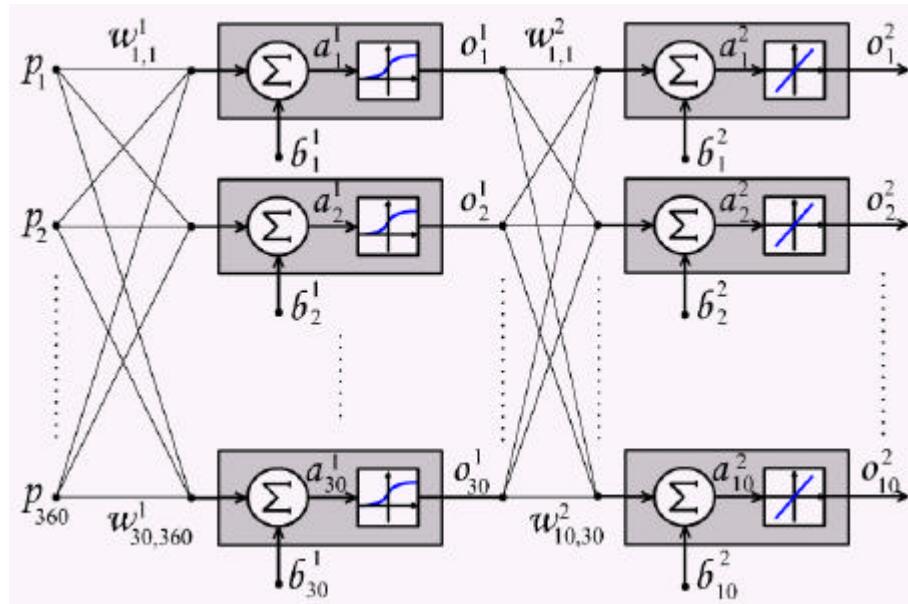


圖 7-1 一對一輸出型之類神經網路架構

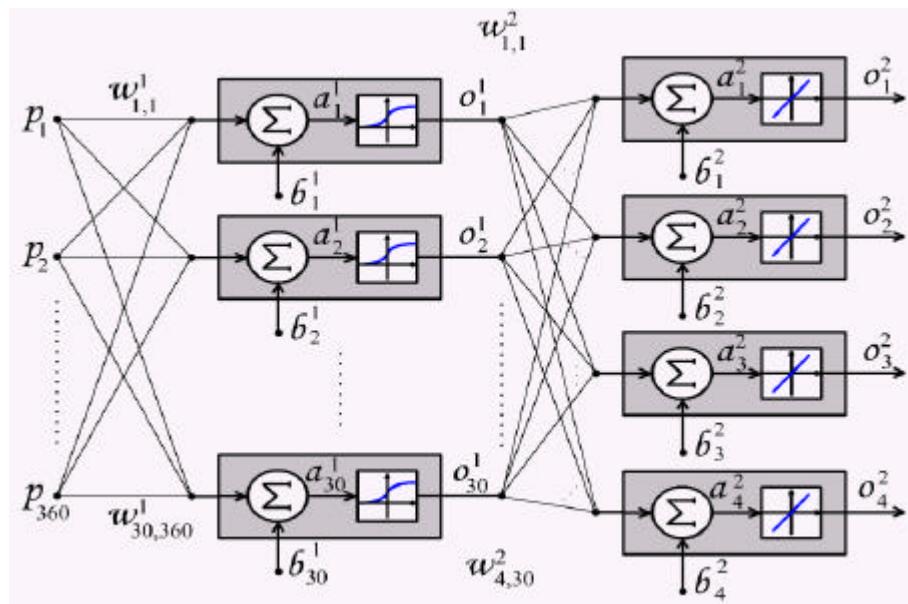


圖 7-2 二進位編碼輸出型之類神經網路架構

表 7-5 二進位編碼輸出型之中文數字對應表

$o_1^2, o_2^2, o_3^2, o_4^2$	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001
中文數字	0	1	2	3	4	5	6	7	8	9

7.3.7 辨識演算法 DTW 與 BPNN 之比較

一般常用的辨識演算法為第五章所介紹的動態時間校準法(DTW)，與第六章所介紹的倒傳遞類神經網路模型(BPNN)等兩大辨識架構，為了比較此兩辨識演算法何者較適合於 DSP 平台上發展，特別設計此一實驗，實驗之系統如圖 7-3 所示。

實驗中，動態時間校準法的語音資料庫採每個字音錄音兩次，並擷取兩字音之特徵參數，進行第五章所介紹之特徵參數平均化的處理，再將處理後的特徵參數做為語音資料庫中的樣板圖。而測試樣本則對同一字音各錄音三次，用以供系統進行測試。

而倒傳遞類神經網路的語音樣本同樣採每個字音錄音兩次，並擷取每個字音之特徵參數，對倒傳遞類神經網路進行訓練的程序，而測試樣本也同樣是對每一個字音錄音三次，再對系統進行測試。

在語音辨識系統的評估中，主要的評估參數包括：系統計算量之多寡、訓練時程的長短、辨識速度的快慢、辨識率之高低以及系統資源的需求等。故在此實驗中，即針對此評估參數來比較 DTW 與 BPNN 兩者之差異。

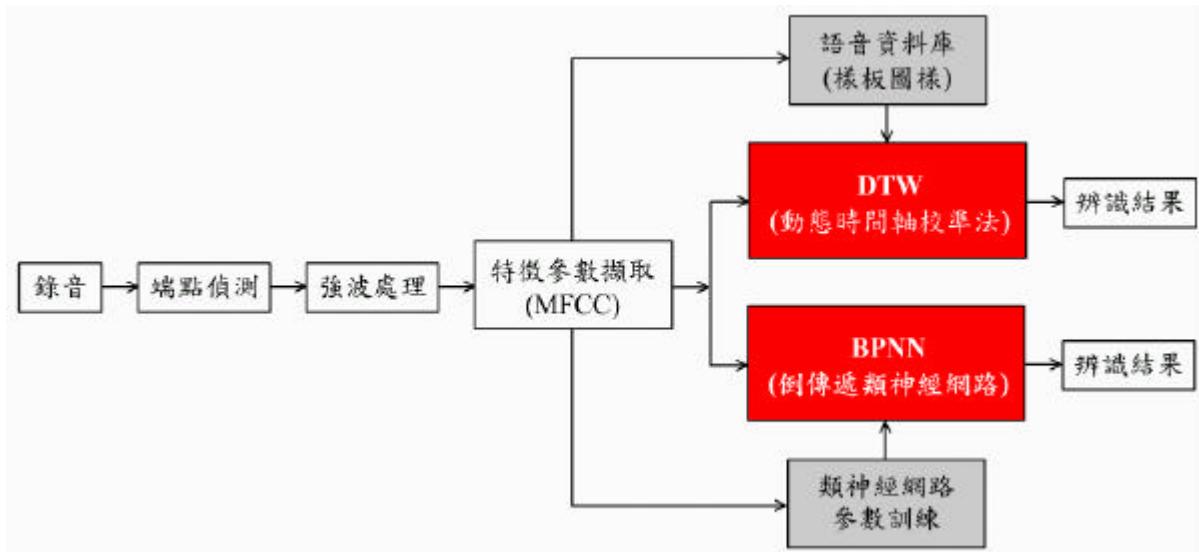


圖 7-3 實驗測試方塊圖

7.3.8 非特定語者之系統應用於特定語者之影響

如同在實驗 D 中所進行的非特定語者之中文數字辨識系統，同樣採用倒傳遞類神經網路(BPNN : 360-30-10)模型，如圖 7-1 所示，在此實驗的過程分為兩個階段進行：非特定語者系統與特定語者系統。首先以 $\frac{2}{3}$ 之語音資料庫樣本(8 男 6 女)對非特定語者之辨識系統進行訓練，每訓練一個循環，便以 $\frac{1}{3}$ 的樣本(4 男 3 女)來對系統進行測試，評估系統之辨識率，當系統之辨識率高於 90% 時，將此一系統應用於特定語者中進行測試且每測試一回，隨即對系統的參數再做適應調整的處理，以提高其辨識率。

為使實驗結果更客觀，所以在非特定語者訓練階段，隨機取樣 $\frac{2}{3}$ 之語音資料庫做為訓練樣本，而剩下之 $\frac{1}{3}$ 則為測試樣本；而要進行特定語者測試時，則從 $\frac{1}{3}$ 的測試樣本中，抽出一個語者之樣本來進行測試；如此重覆進行三次，再將所得之辨識率結果做平均化之處理。

在實驗 E 中曾提及二進位編碼之輸出架構，在本實驗中亦針對此編碼架構進行同樣的實驗，比較非特定語者之系統應用於特定語者之情況下，二進位編碼之輸出架構的性能是否會比一對一輸出架構之性能來得佳。

第八章 實驗結果

關於第七章中所進行的八項實驗所獲得的結果，將在本章中做更進一步的闡述與說明。

8.1 端點偵測演算法之比較結果

圖 8-1 與圖 8-2 分別針對未受雜訊干擾與含雜訊干擾之錄音訊號，使用不同端點偵測法所得到之結果。可發現，能量曲線判別法不易受雜訊干擾，但卻無法判別出氣音段之所在；而 R-S 判別法則能有效找出有聲段(含氣音段)之正確端點位置，但對於受雜訊干擾的錄音訊號，則容易造成嚴重誤判的結；而利用雜訊指標來選用不同判別法的方式，其結果與人工設定的端點較為接近，從實驗的結果也證實此法可適時避開雜訊干擾所造成的誤判。

因此在本文中關於端點偵測的部份(包括建立語音資料庫)，皆是採用自動選定判別法。其原因有二，首先即是在此實驗中所獲得的結果，其能適時的避開雜訊干擾所造成的誤判率；其次便是在實驗的過程中發現，當建立語音樣本時，若所使用的端點偵測演算法與測試所用的判別法不同時，則系統的辨識率將會受到影響，而造成辨識率降低。

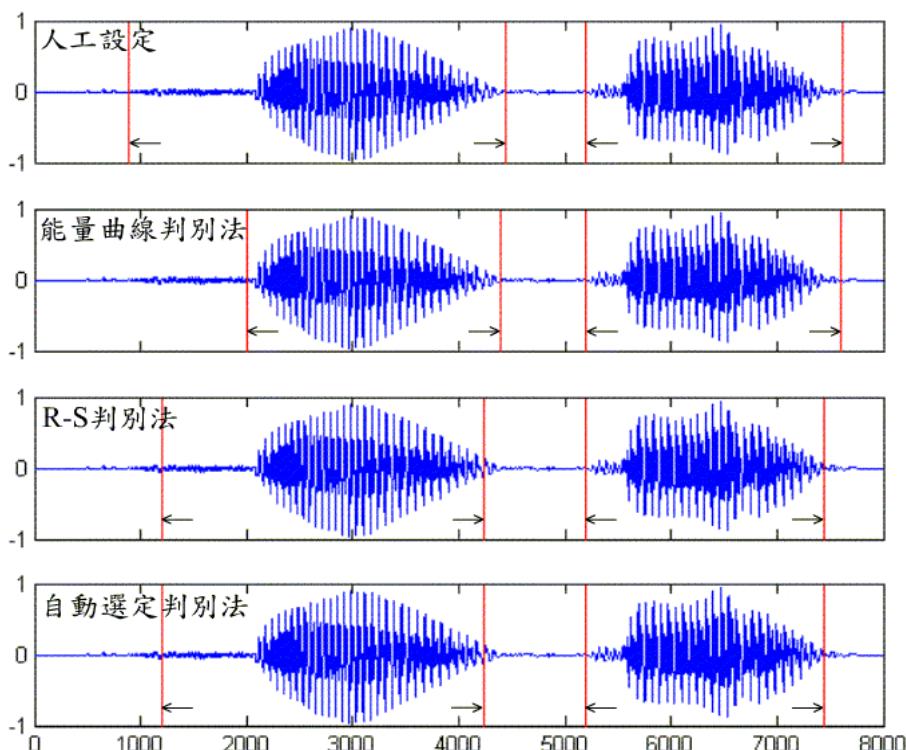


圖 8-1 訊號未受雜訊干擾之端點偵測結果

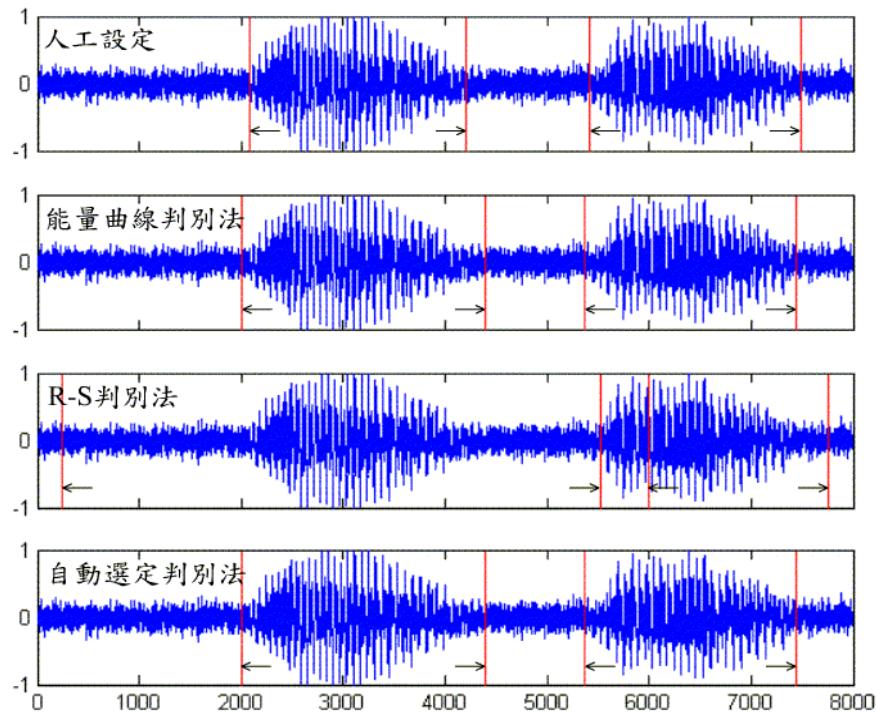


圖 8-2 訊號受雜訊干擾之端點偵測結果(SNR=20dB)

8.2 特徵參數的選定

如圖 8-3 所示，分別採用此四個參數 LFS、MFS、LFCC 與 MFCC 所做的測試結果。從實驗中可以發現，採用倒頻譜參數 LFCC 或 MFCC 做為語音特徵參數的系統，會比採用頻譜參數 LFS 或 MFS 的系統，在收斂速度上較快，亦即要達到相同水準之辨識率，採用倒頻參數所需的訓練次數較少即可辦到。再加上從參數的定義中可知，MFCC 特徵參數較接近人類聽覺頻率反應，所以在本文中往後的設計將以梅爾倒頻譜係數(MFCC)做為語音訊號特徵參數。

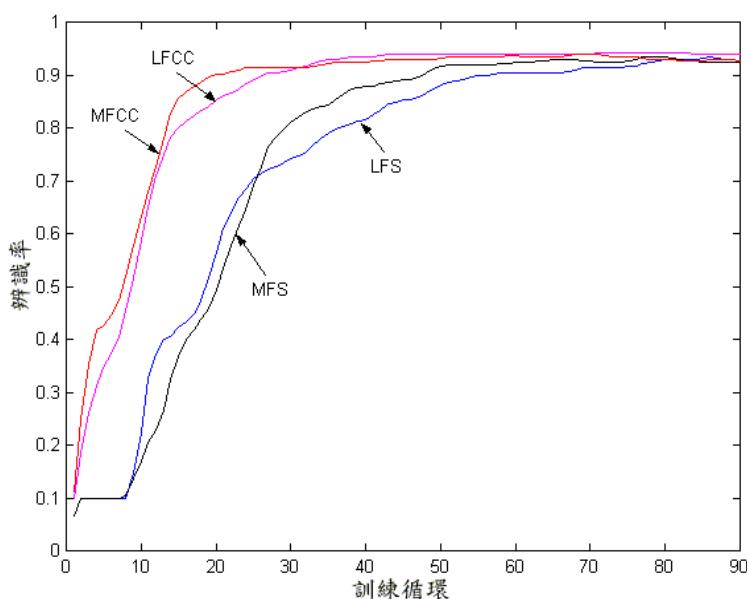


圖 8-3 不同特徵參數對系統辨識效能的影響

8.3 學習調整率 α 對系統學習效能的影響

從實驗結果可發現，採用本文中所介紹之衰減型學習調整率設定法，比傳統使用固定常數之方法所獲得的學習速度來的快；如圖 8-4 中所示之曲線。從實驗的過程中發現，系統學習的速度取決於學習調整率 α 值的設定，若 α 值設定太大則系統容易發散或發生振盪的現象，但若選的太小則會增加系統訓練的時間。

此後系統在學習調整率的設定，將採用(6-26)式之公式，且取 $A = 5, T = 1$ 之值，如此之設定可使系統在開始之學習調整率為平常的 5 倍，使系統可大幅修正連結加權參數與偏移量參數，加快系統學習速度；且在第 5 個訓練週期之後，系統的學習調整率又可恢復平常之值，避免發生振盪的現象。

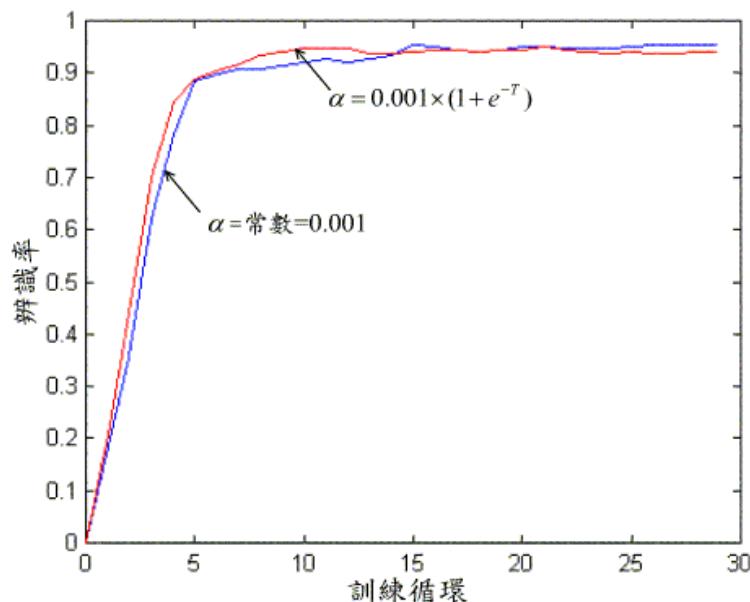


圖 8-4 採用不同學習調整率所得之學習過程

8.4 非特定語者之中文數字辨識系統

從非特定語者之中文數字辨識系統的訓練與測試中發現，系統在經過約 10 個訓練週期後，系統之辨識率即可達 90% 以上，當訓練週期超過 20 個之後，系統的辨識率維持在 94% ~ 96% 之間變動，如圖 8-5 所示為系統的學習曲線。

本論文中所設計之非特定語者中文數字辨識系統，應用於未經訓練之使用者，可得 95% 之辨識率，而對於語音樣本包含在訓練樣本中的使用者，其辨識率將可高於 99%。所以對於新的使用者，只要對參數進行適應調整處理，則此系統將可應用於所有的人使用中，此情況的實際試驗將在實驗 H 中進行。

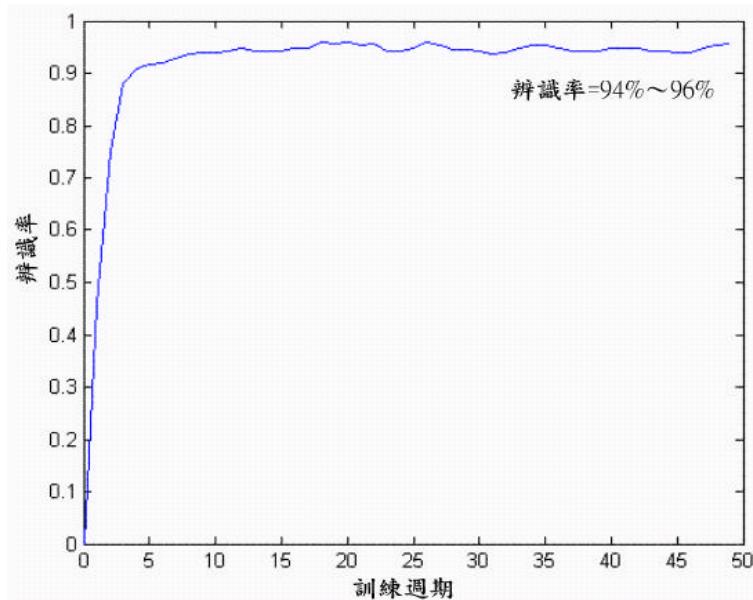


圖 8-5 非特定語者之中文數字辨識系統響應圖

8.5 倒傳遞類神經網路輸出架構之設計

從兩種不同的輸出架構測試中可發現，採用一對一輸出的架構之辨識率高於以二進位編碼輸出的架構，兩系統之辨識率相差約 10%，造成此一差異最大的主因，乃類神經網路本身即可視為一種維度空間轉換的關係，以此測試系統而言，一對一輸出的架構為 360 維空間轉換到 10 維空間，而二進位編碼之輸出架構為 360 維空間轉換到 4 維空間，使得前者之系統輸出間的差異度遠大於後者所致。

在此實驗乃為非特定語者之情況，若所要應用的場合為特定的使用者，則採用二進位編碼之輸出架構將會比一對一輸出架構佔優勢，尤其是在輸出的分類很多的時候，此差異會更明顯。關於此情況的實際試驗將在實驗 H 中進行。

8.6 神經元移除法則的應用

當系統辨識率達 95% 時，利用神經元移除法則，將次要的神經元移除，保留剩下之神經元做為系統新的架構，重新對系統進行測試與訓練，可發現（參考圖 8-7）：當次要神經元被移除時，系統的辨識率衰減了將近 10%。接著再對此新架構之系統進行同樣的訓練，可使系統提高約 5% 的辨識率。移除次要神經元之系統，其辨識率約為 82% ~ 90% 之間。

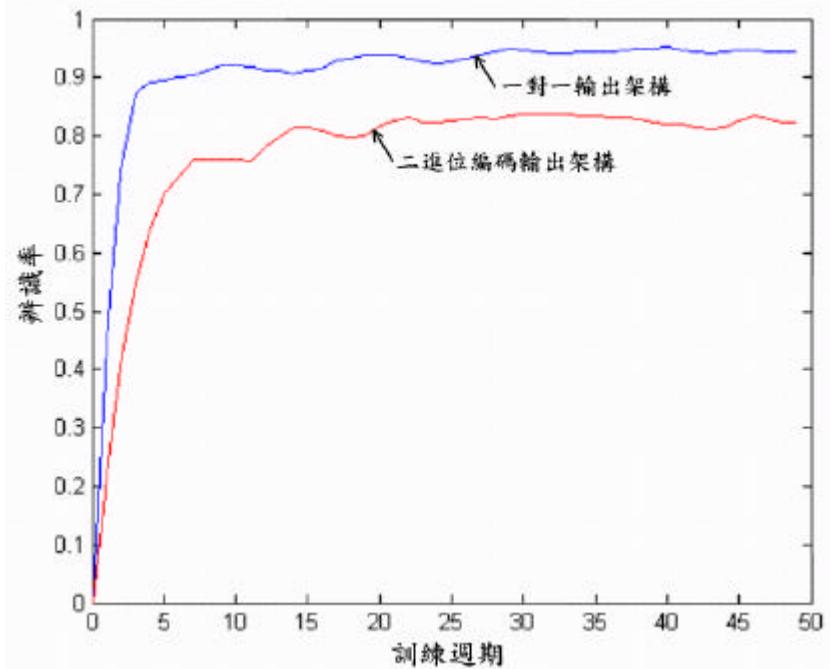


圖 8-6 一對一與二進位編碼之輸出架構性能比較圖

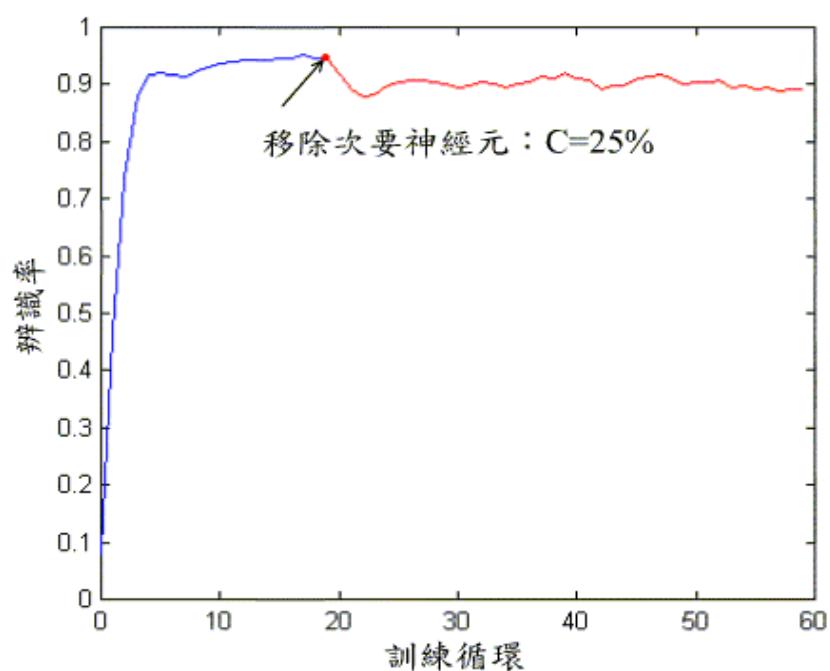


圖 8-7 移除次要神經元之學習曲線

8.7 辨識演算法 DTW 與 BPNN 之比較

表 8-1 為分別以 DTW 與 BPNN 為辨識系統架構，所進行的實驗結果；可以發現，就辨識率而言，DTW 的精確度略高於 BPNN 之系統；且 DTW 不需對系統進行額外的訓練，只需取得每個字音的樣本即可；然而在辨識速度與記憶體需求上，BPNN 却遠超過 DTW 後很多，以採用神經元移除法之 BPNN 來說，其辨識速度約為 DTW 的 120 倍；記憶體的需求量也減少了 76%。

且當系統的語音資料庫增多時，DTW 的比對時間會相對的跟著增加，系統的記憶體需求量也會變大；而對 BPNN 而言，當系統架構確定，只要是用於小量字庫範圍內(100 個字詞)之場合，則即使字庫量增加，也不會影響系統的辨識速度與記憶體需求量，所以 BPN 的架構的確優於 DTW 的系統。

從實驗中還發現一個問題，以 DTW 所設計的系統在每次的使用中(如訓練後，一個星期再進行測試)，若沒有加入適當的樣本更新方法(或特徵參數平均化)，則其辨識率會嚴重衰減，甚至不到 50%的辨識率。而 BPNN 在這方面，由於類神經網路本身的容錯能力較強，故其影響較小。

表 8-1 DTW 與 BPNN 特性測試結果

性能比較	動態時間校準法	倒傳遞類神經網路	
		固定神經元數(30)	採用神經元移除法
訓練時間	不需進行訓練	2.45	1.0
辨識速度	120.0	1.75	1.0
記憶體需求	30K Byte	12K Byte	7K Byte
辨識率	97%	96%	90%

8.8 非特定語者之系統應用於特定語者之影響

從實驗結果中可發現(參考圖 8-8)，當系統應用於一新的使用者時，其辨識率會立刻降低，然而只要針對該使用者，擷取其字音樣本特徵參數，對系統再進行訓練，將會發現經過約 5 個訓練週期，系統的辨識率可大幅地提高，使得系統的辨識率高於 99%。

在圖 8-8 中，還有另一辨識系統，即採用二進位編碼輸出架構之系統，從實驗中可發現，此系統應用於非特定語者時，其辨識率約為 80%~85%之間(參考圖 8-6)，遠不如一對一輸出架構之系統，然而當此系統應用於特定語者時，系統參數經適應調整後，其最後的辨識率將可與一對一輸出架構之系統一樣。故若使用的場合為特定語者，則採用二進位編碼輸出架構之系統為佳。

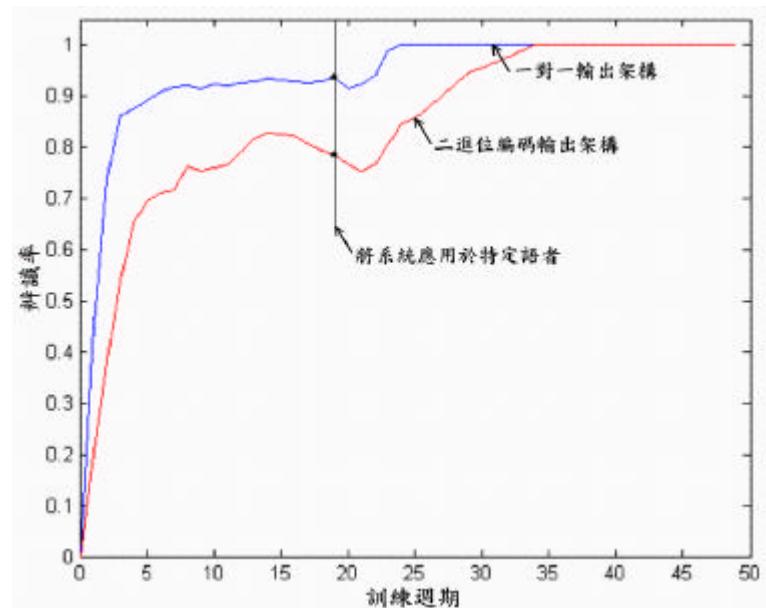


圖 8-8 非特定語者系統應用於特定語者之系統響應圖

第九章 結論與展望

9.1 結論

本論文以倒傳遞類神經網路(BPNN)為架構設計一非特定語者之中文數字語音辨識系統，透過變動型之學習調整率的設定，可加快系統的學習速度，並對訓練樣本之訓練順序做有效之安排，使系統之辨識率可達 95%。

當此所設計之非特定語者辨識系統應用於特定語者時，經即時的適應修正後，更可使系統之辨識率高於 99%，大幅提高系統之辨識率，使得以倒傳遞類神經網路為架構之辨識系統，更適合於內建式語音控制電子產品之發展。

為能使辨識系統轉移到數位處理器(DSP)平台，針對類神經網路模型，提出了神經元移除法則，將訓練後之連結加權值振幅較小者予以移除，利用此法之系統，約可減去 $\frac{1}{3}$ 數量之神經元，降低系統 20%~40% 的記憶體需求，且系統之辨識率仍可達 85%。除此之外，在 BPNN 網路模型的輸出架構中，提出以二進位編碼之方式取代傳統一對一之架構，除可增加系統之辨識字彙之數量外，亦可減少輸出神經元之數量，降低系統的記憶體需求與減少計算量。

關於語音訊號端點偵測中，採用能量曲線與 R-S 判別法之混合模式，根據背景雜訊的能量參數與越零率參數，定出端點偵測法對照表，如此一來使得系統在端點偵測過程中，可適時的避開雜訊的干擾，且不需複雜之運算，即可有效定位出有聲段所在之處。

從 DTW 與 BPNN 的比較結果也證實，本論文中所提之 BPNN 架構比 DTW 更適合在 DSP 平台上發展，不論是在計算量、記憶體需求、辨識速度與辨識率等各方面的考量，BPNN 都遠勝於 DTW。而近年來更人性化的高科技產品發展蓬勃，對於語音操作界面之電子產品，此架構可做為實際應用設計時之參考依據。

9.2 展望

本論文已完成階段性之任務，明確找出適用於 DSP 上發展之辨識演算架構，即倒傳遞類神經網路(BPNN)，同時也做了適當的修正與改良，使得修改後之 BPNN 更加適合於 DSP 上應用與發展。而實驗室下一階段之研究目標即可根據此倒傳遞類神經網路之架構，著手於 DSP 平台上開發語音辨識電子產品之應用，繼續往前邁進。

語音訊號處理部份，除循著實驗室規劃之五大階段發展目標外，在本論文的執行過程中卻也發現幾個相當值得深入研究探討的主題，在此提出供有興趣之研究人員參考：

- A.雜訊干擾的抑制或消除。
- B.如何消除使用不同錄音系統所造成的不良影響。
- C.如何修改系統，使其能做到即時(real-time)的辨識。

當然除了這三個問題之外，在語音辨識領域中仍有很多瓶頸，正等待更多充滿熱忱的同學，一起投入心力，克服種種障礙，為語音辨識領域開拓出另一能為人類帶來更多方便性的技術。

Reference

- [1] X.D. Huang and K.F. Lee, “*On Speaker-Independent, Speaker-Dependent, and Speaker-Adaptive Speech Recognition,*” IEEE Trans on ASSP, 1991
- [2] P. Woodland, “*Speech Recognition,*” IEE, 1998.
- [3] H. Sakoe and S. Chiba, “*Dynamic Programming Optimization for Spoken Word Recognition,*” IEEE Trans on ASSP, Vol.26, pp 43-49, Feb. 1978.
- [4] C. Myers and L.R. Rabiner, “*Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition,*” IEEE Trans on ASSP, Vol.28, No.6, pp 623-635, Dec. 1980.
- [5] D.P. Morgan and C.L. Scofield, *Neural Networks and Speech Processing*, Kluwer Academic, 1991.
- [6] L.R. Rabiner, “*A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,*” IEEE Trans on ASSP, Vol.77, No.2, pp 257-286, Feb. 1989.
- [7] L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, pp 200-232, 1993.
- [8] F. Runstein and F. Violaro, “*An Isolated-Word Speech Recognition System Using Neural Networks,*” IEEE Trans on ASSP, pp 550-553, 1996.
- [9] G.D. Wu and C.T. Lin, “*A Recurrent Neural Fuzzy Network for Word Boundary Detection in Variable Noise-Level Environments,*” IEEE Trans on System, Man, and Cybernetics, Vol. 31, No. 1, pp 84-97, Feb. 2001.
- [10] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978.
- [11] A. Hussain, S.A. Samad and L.B. Fah, “*Endpoint Detection of Speech Signal using Neural Network,*” IEEE Trans on ASSP, pp 271-274, 2000.
- [12] L.F. Lamel and L.R. Rabiner, “*An Improved Endpoint Detector for Isolated Word Recognition,*” IEEE Trans on ASSP, Vol.29, No.4, pp 777-785, Aug. 1981.
- [13] E. Keller, *Fundamentals of Speech Synthesis and Speech Recognition Basic Concepts, State of the Art and Future Challenges*, John Wiley and Sons, 1994.

- [14] S.B. Davis and P. Mermelstein, “*Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences*,” IEEE Trans on ASSP, Vol.28, No.4, pp357-366, Aug. 1980.
- [15] J.R. Deller, J.G. Proakis and J.H.L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan, 1993.
- [16] J.N. Holmes, *Speech Synthesis and Recognition*, Van Nostrand Reinhold, 1988.
- [17] 葉怡成, *類神經網路模式應用與實作*, 儒林出版社, 1993.
- [18] 中國科學技術大學生物醫學工程跨系委員會, *神經網路及其應用*, 儒林出版社, 1993.
- [19] H.A. Bourlard and N. Morgan, *Connectionist Speech Recognition A Hybrid Approach*, Kluwer Academic, 1994.
- [20] M.T. Hagan, H.B. Demuth and M. Beale, *Neural Network Design*, PWS, 1996.
- [21] J.S. Jang, C.T. Sun and E. Mizutani, *Neuro-Fuzzy and Soft Computing*, Prentice Hall, 1997.
- [22] K.J. Astrom, B. Wittenmark, *Adaptive Control*, 2nd, Addison Wesley, 1995.
- [23] 蘇木春, 張孝德, *機器學習：類神經網路、模糊系統以及基因演算法則*, 全華科技圖書, 1997.