

國立交通大學

電信工程學系

碩士論文

使用基頻資訊之國語分散式語音辨識系統

The Mandarin Distributed Speech Recognition  
System Using Pitch Information

研究生：魯柏暄

指導教授：王逸如 博士

中華民國九十四年六月

使用基頻資訊之國語分散式語音辨識系統  
**The Mandarin Distributed Speech Recognition  
System Using Pitch Information**

研究生：魯柏暄

Student : Bo-Xuan Lu

指導教授：王逸如 博士

Advisor : Dr. Yih-Ru Wang

國立交通大學



A Thesis

Department of Communication Engineering  
College of Electrical Engineering and computer Science  
National Chiao Tung University  
In Partial Fulfillment of Requirements  
for the Degree of  
Master of Science  
in Electrical Engineering

June 2005

Hsinchu, Taiwan, Republic of China

中華民國九十四年六月

# 使用基頻資訊之國語分散式語音辨識系統

研究生：魯柏暄

指導教授：王逸如 博士

國立交通大學電信工程學系碩士班



在本論文中，將在分散式語音辨認架構之標準—ETSI ES 202 212 XAFE 下，建立國語語音辨認之效能評估。論文中共作了國語數字串及國語大詞彙連續語音兩種語音辨認實驗。首先在實驗發現 ETSI 分散式語音辨認架構之基頻偵測器在語音信號的訊噪比低於 10dB 時，ETSI 架構之基頻偵測器的效能嚴重變壞；這使得國語語音辨識器在低訊噪比時，使用基頻資訊會比未使用基頻資訊的結果差；在論文中提出了一個小幅修改 ETSI 架構之基頻偵測方法後，可以增進在低訊噪比時的基頻偵測效能。論文中更藉由整合使用基頻資訊及未使用基頻資訊辨識器之辨認分數，可有效增進環境雜訊下的國語語音辨識率。最後在國語數字串可獲得 86.8% 辨認率，在國語大詞彙連續語音可獲得 65.3%、45.4% 的音節及字元辨認率。

關鍵詞：分散式語音辨識、基頻偵測器

# The Mandarin Distributed Speech Recognition System Using Pitch Information

Student : Bo-Xuan Lu

Advisor: Dr. Yih-Ru Wang

Department of Communication Engineering  
National Chiao Tung University

## Abstract

In the thesis, the performance of Mandarin digit-string and continuous large vocabulary Mandarin speech recognition were evaluated under ETSI ES-202-212 XAFE environment. First, the experimental results showed that the performance of the pitch detection algorithm degraded seriously when the SNR of speech signal was lower than 10dB. This makes the Mandarin speech recognizer using pitch information perform inferior to the recognizer without using pitch information in low SNR environments. A modification of the pitch detection algorithm is therefore proposed to improve the performance of ETSI's pitch detector in low SNR environments. The recognition performance of Mandarin speech can be improved for most SNR levels by integrating the recognizers with and without using pitch information. Finally, 86.8% recognition rate can be achieved for Mandarin digit-string. 65.3% syllable and 45.4% character recognition rates can be achieved for Mandarin continuous speech.

Keywords: DSR, Pitch detection

# 誌謝

在研究所的這兩年中，最需要感謝的人就是陳信宏老師與王逸如老師，尤其是王逸如老師每次都苦口婆心、不厭其煩的教導我，讓我學到了許多許多，在這邊要跟老師說一聲「老師，您辛苦了！」；再來要感謝由於有聯發科的支持，讓我們能順利進行研究。

再來要感謝實驗室的學長、同學及學弟；智合及性獸學長，在我碰到問題時由於有你們的幫忙，我才能順利解決，尤其要感謝性獸常常在幫我抓蟲；接著要感謝我的同學們：順哥、隆勳、金翰、希群以及我們實驗室之花—佩穎，幸虧有你們的陪伴，我才能順順利利地走完這兩年；還有可愛的學弟們，謝謝你們在我最後的一年裡，常常帶給我們許多的歡樂；最後還要感謝輝哥，因為你讓我知道原來世界是這樣大的阿！



最後我要感謝我的家人，以及我的女友，由於有了你們的支持，我才能夠努力到現在，謝謝你們！！

# 目錄

中文摘要.....	I
英文摘要.....	II
誌謝.....	III
目錄.....	IV
表目錄.....	VII
圖目錄.....	IX
第一章 導論.....	1
1.1 研究動機.....	1
1.2 研究方向與主要成果.....	2
1.3 章節概要.....	2
第二章 背景知識與基礎系統.....	3
2.1 分散式語音辨認系統介紹.....	3
2.2 分散式語音辨識系統環境下國語連續數字串之辨認.....	6
2.2.1 語料庫.....	6
2.2.2 環境雜訊.....	7
2.2.3 分散式語音辨識系統後級隱馬可夫模型之語音辨識器.....	13
2.2.4 實驗結果.....	16
第三章 使用基頻參數的分散式國語連續語音辨識系統.....	20
3.1 分散式語音辨識系統中的基頻抽取.....	20
3.2 基頻在不同環境雜訊及不同的訊噪比之下的分析.....	23
3.3 國語連續數字串之辨識---加入分散式語音辨識系統抽取的基頻 參數.....	28

3.3.1	實驗設定.....	28
3.3.2	實驗結果.....	30
第四章	改良基頻參數抽取的方法.....	35
4.1	改良式分散式語音辨識系統之基頻參數抽取.....	35
4.2	改良式分散式語音辨識系統前級之基頻參數抽取器之效能分析.....	36
4.3	國語連續數字串之辨識---加入改良式分散式語音辨識系統抽取 的基頻參數.....	37
4.3.1	實驗設定與訓練模型建立.....	38
4.3.2	實驗結果.....	39
4.4	國語連續數字串之辨識---整合沒有加入基頻參數的辨識器與加入改良 式分散式語音辨識系統抽取之基頻參數的辨識器.....	44
4.4.1	實驗設定.....	45
4.4.2	實驗結果.....	46
4.5	國語連續數字串之辨識---使用乾淨語音的基頻參數之辨識器.....	48
4.5.1	實驗設定.....	48
4.5.2	實驗結果.....	48
第五章	大字彙國語連續語音辨認.....	50
5.1	語料庫介紹---TCC300.....	50
5.2	大字彙國語連續語音之辨識---沒有加入基頻參數.....	51
5.2.1	實驗設定.....	51
5.2.2	實驗結果.....	54
5.3	大字彙國語連續語音之辨識---加入改良式的分散式語音辨識系統抽取 之基頻參數.....	56
5.3.1	實驗設定.....	56
5.3.2	實驗結果.....	57

5.4 大字彙國語連續語音之辨識---整合沒有加入基頻參數的辨識器與加入改良式分散式語音辨識系統抽取之基頻參數的辨識器.....	64
5.4.1 實驗設定.....	64
5.4.2 實驗結果.....	65
5.5 加入語言模型至使用改良式分散式語音辨識系統抽取之基頻參數的大字彙國語連續語音辨識.....	66
5.5.1 建立語言模型.....	67
5.5.1.1 訓練語料及詞典.....	67
5.5.1.2 訓練語言模型的方法.....	69
5.5.2 基本辨識器加入語言模型之辨識分析.....	70
5.5.3 實驗結果.....	71
第六章 結論與展望.....	72
6.1 結論.....	72
6.2 展望.....	73
參考文獻.....	74





## 表目錄

表 2-1	國語連續數字串語料庫.....	7
表 2-2	八種環境雜訊的音檔長度.....	8
表 2-3	加上環境雜訊的國語連續數字串內容介紹.....	14
表 2-4	語音特徵參數抽取之參數設定.....	15
表 2-5(a)	國語連續數字串---乾淨語音訓練模式之辨識結果.....	16
表 2-5(b)	國語連續數字串---複合情境訓練模式之辨識結果.....	17
表 3-1(a)	基頻在地下鐵環境下不同訊噪比之分析.....	24
表 3-1(b)	基頻在嘈雜的人聲環境下不同訊噪比之分析.....	24
表 3-1(c)	基頻在汽車環境下不同訊噪比之分析.....	24
表 3-1(d)	基頻在展覽會場環境下不同訊噪比之分析.....	25
表 3-1(e)	基頻在餐廳環境下不同訊噪比之分析.....	25
表 3-1(f)	基頻在街道環境下不同訊噪比之分析.....	25
表 3-1(g)	基頻在機場環境下不同訊噪比之分析.....	26
表 3-1(h)	基頻在火車環境下不同訊噪比之分析.....	26
表 3-1(i)	基頻在不同訊噪比之分析.....	27
表 3-2(a)	加入基頻參數後的國語連續數字串之乾淨語音訓練模式辨識結果.....	30
表 3-2(b)	加入基頻參數後的國語連續數字串之複合情境訓練模式辨識結果.....	31
表 3-3	八種環境雜訊在兩個實驗中的進步情形.....	33
表 4-1	比較原本基頻參數抽取的作法與改進後基頻參數抽取的作法.....	37
表 4-2(a)	加入改良式分散式語音辨識系統抽取之基頻參數的國語連續數字串 辨認實驗中乾淨語音訓練模式之辨識結果.....	39
表 4-2(b)	加入改良式分散式語音辨識系統抽取之基頻參數的國語連續數字串 辨認實驗中複合情境訓練模式之辨識結果.....	40

表 4-3(a) 整合有由改良式的分散式語音辨識系統抽取之基頻參數以及不含基 頻參數的國語連續數字串辨識實驗中乾淨語音訓練模式之辨識結果.....	46
表 4-3(b) 整合有由改良式的分散式語音辨識系統抽取之基頻參數以及不含基 頻參數的國語連續數字串辨識實驗中複合情境訓練模式之辨識結果.....	47
表 4-4 使用乾淨語音之基頻參數的國語連續數字串辨識中乾淨語音訓練模 式之辨識結果.....	49
表 5-1 大字彙連續國語語音的語料庫.....	50
表 5-2 加上環境雜訊的大字彙連續國語語音內容介紹.....	52
表 5-3 在測試語料中 8 種環境雜訊下之音節數.....	52
表 5-4(a) 沒有加入基頻參數的大字彙國語連續語音辨識實驗中的乾淨語音訓 練模式之辨識結果.....	54
表 5-4(b) 沒有加入基頻參數的大字彙國語連續語音辨識實驗中的複合情境訓 練模式之辨識結果.....	55
表 5-5 每個聲調的出現次數.....	57
表 5-6(a) 加入改良式的分散式語音辨識系統抽取之基頻參數的大字彙國語連 續語音辨識實驗中的乾淨語音訓練模式 1,515 個音節之辨識結果.....	58
表 5-6(b) 加入改良式的分散式語音辨識系統抽取之基頻參數的大字彙國語連 續語音辨識實驗中的複合情境訓練模式 1,515 個音節之辨識結果.....	59
表 5-7(a) 加入改良式的分散式語音辨識系統抽取之基頻參數的大字彙國語連 續語音辨識實驗中的乾淨語音訓練模式之聲調辨識結果.....	60
表 5-7(b) 加入改良式的分散式語音辨識系統抽取之基頻參數的大字彙國語連 續語音辨識實驗中的複合情境訓練模式之聲調辨識結果.....	61
表 5-8(a) 加入改良式的分散式語音辨識系統抽取之基頻參數的大字彙國語連 續語音辨識實驗中的乾淨語音訓練模式 411 個音節之辨識結果.....	62
表 5-8(b) 加入改良式的分散式語音辨識系統抽取之基頻參數的大字彙國語連 續語音辨識實驗中的複合情境訓練模式 411 個音節之辨識結果.....	63

表 5-9	整合有由改良式的分散式語音辨識系統抽取之基頻參數以及不含基 頻參數的大字彙國語連續語音辨識實驗中乾淨語音訓練模式之辨識結果..	65
表 5-10	詞典中之詞長分佈.....	68
表 5-11	通用語料庫之詞數表.....	69
表 5-12	在複合情境訓練、測試一人聲環境雜訊.....	71



## 圖目錄

圖 2-1	分散式語音系統架構圖.....	4
圖 2-2	降低雜訊處理系統流程圖.....	6
圖 2-3	在乾淨語料中加入環境雜訊示意圖.....	8
圖 2-4	八種環境雜訊的長時間頻譜.....	11
圖 2-5	八種環境雜訊的頻譜—時間圖（橫軸：時間；縱軸：頻率）.....	12
圖 2-6	八種環境雜訊在兩個訓練模式的比較.....	18
圖 2-7	不同的訊噪比在兩個訓練模式的比較.....	19
圖 3-1	分散式語音辨識前級估計基頻與語音狀態資訊系統架構.....	21
圖 3-2	經過指數函數補償的基頻.....	22
圖 3-3	比較在乾淨語音與汽車環境雜訊中訊噪比為 0dB 的基頻值.....	28
圖 3-4	八種環境雜訊的辨識結果比較.....	33
圖 3-5	不同的訊噪比的辨識結果比較.....	34
圖 4-1	改良式分散式語音辨識系統前級之基頻參數抽取的架構.....	35
圖 4-2(a)	在乾淨語音訓練模式中比較沒有加入基頻參數、使用 DSR XAFE 與 Modified XAFE 抽取之基頻參數的辨識結果.....	41
圖 4-2(b)	在複合情境訓練模式中比較使用沒有加入基頻參數、DSR XAFE 與 Modified XAFE 抽取之基頻參數的辨識結果.....	41
圖 4-3	整合含有基頻參數以及不含基頻參數的辨識器之系統方塊圖.....	45
圖 5-1	LM 訓練流程圖.....	67
圖 5-2	LM 轉 Word-Net 之流程圖.....	70

# 第一章 導論

## 1.1 研究動機

由於科技產業的蓬勃發展，電子產品的功能日益強大，而體積卻是越做越小，這不但是代表著人類的卓越的發明、創新的能力，同時這也為我們的生活帶來了許許多多的便利；加上無線網路的進步，滿足了人們隨心所欲、隨時隨地交流資訊的渴望。例如行動電話和筆記型電腦的普及，讓我們無論是在路上或是車上，時時刻刻都能夠利用電子產品與網際網路接軌，可以接收或是發送資訊，尤其是以行動電話的普及率更是高達到幾乎人手一機了；但是為了便利性、可攜帶性，我們不斷的追求這些產品的輕薄短小，使得傳統式的鍵盤或按鍵輸入已經漸漸不是最方便的輸入方式了，我們需要的是能夠更快速且能處理繁雜指令的輸入介面。而使用語音當成輸入介面就是一個很好的方法。



根據前面所述，若是使用語音作為新的輸入介面，勢必碰上許多問題：手持設備（Handheld device）的體積太小，其計算能力以及儲存用的記憶體將嚴重受限，使得我們要在手持設備上處理整個語音辨識的程序是有困難的。因此分散式語音辨識（Distributed Speech Recognition; DSR）架構就此產生。分散式語音辨識的想法是將整個語音辨識工作分成兩個部分：在手持設備（Client）上，因為有許多限制，因此只做簡單的語音參數的抽取與壓縮，再將這些資料透過無線通道傳送到遠端的伺服器（Server）執行語音辨識。也因為要透過無線通道來傳遞資訊，無可避免的會有因多通道衰減造成（Multi-path fading）的群集錯誤（Burst error）等；而且當使用者在使用手持設備的同時，會受到週遭環境的影響，是造成使辨識率下降的最重要原因之一。

## 1.2 研究方向

本論文主要的研究方向在分散式語音辨識系統的架構下，建立國語連續語音之辨識器；並且希望能夠加入基頻參數於國語連續語音辨識器中，以對抗環境雜訊的干擾。本論文同時也提出了改良式的分散式語音辨識系統基頻參數抽取的方法，此作法能夠有效的改進高訊噪比下之基頻偵測效能並提升辨識率。

## 1.3 章節概要

第一章 導論：說明本篇論文的研究動機、研究方向及章節概要。

第二章 背景知識與基礎系統：介紹分散式語音辨認系統，並且做了一個國語連續數字串辨認的實驗。

第三章 結合聲調辨識器與分散式語音辨識系統：說明如何建立一個加入由分散式語音辨認系統抽取之基頻參數的辨認器，並且對分散式語音辨認系統抽取之基頻參數所做的效能分析，以及實驗的結果分析。

第四章 改良基頻參數抽取方法：介紹的是如何改良分散式語音辨識系統之抽取基頻參數的方法，使其應用在辨識器時，可以提升辨識率；並且分析比較改良式分散式語音辨識系統抽取之基頻參數與原本的分散式語音辨識系統抽取之基頻參數。

第五章 大字彙國語連續語音辨認：將對大字彙國語連續語音辨認比較其有使用基頻參數與沒有使用基頻參數的差異。

第六章 結論與展望：對本論文的方法結果作結論，並說明未來改進的方向。

## 第二章 背景知識與基礎系統

本章將會介紹分散式語音辨認系統，並且做了建立在分散式語音辨識系統，多種環境雜訊下的國語連續數字串辨識實驗。

### 2.1 分散式語音辨認系統介紹

分散式語音辨識系統主要的構想是來自：想要應用在手持設備可以使用語音輸入更多更複雜的指令，但是手持設備又受限於其計算能力以及記憶體之不足。因此分散式語音辨識系統的架構是將語音辨識分成兩個部分：在手持設備也就是分散式語音辨識系統的前級（DSR front-end）接收語音輸入，繼而抽取語音的特徵參數，經過壓縮、編碼，透過無線通道傳送到伺服器也就是分散式語音辨識的後級（DSR back-end）端進行解碼以及辨識。本論文中之語言辨識前級是使用歐洲電信標準協會編號202 212 V1.1.1 (ETSI ES 202 212 V1.1.1) [1]之分散式語音辨識系統前級的標準（Extended Advance Feature Extraction; XAFE），圖 2-1則是歐洲電信標準協會編號202 212 V1.1.1之分散式語音辨識系統的架構圖。

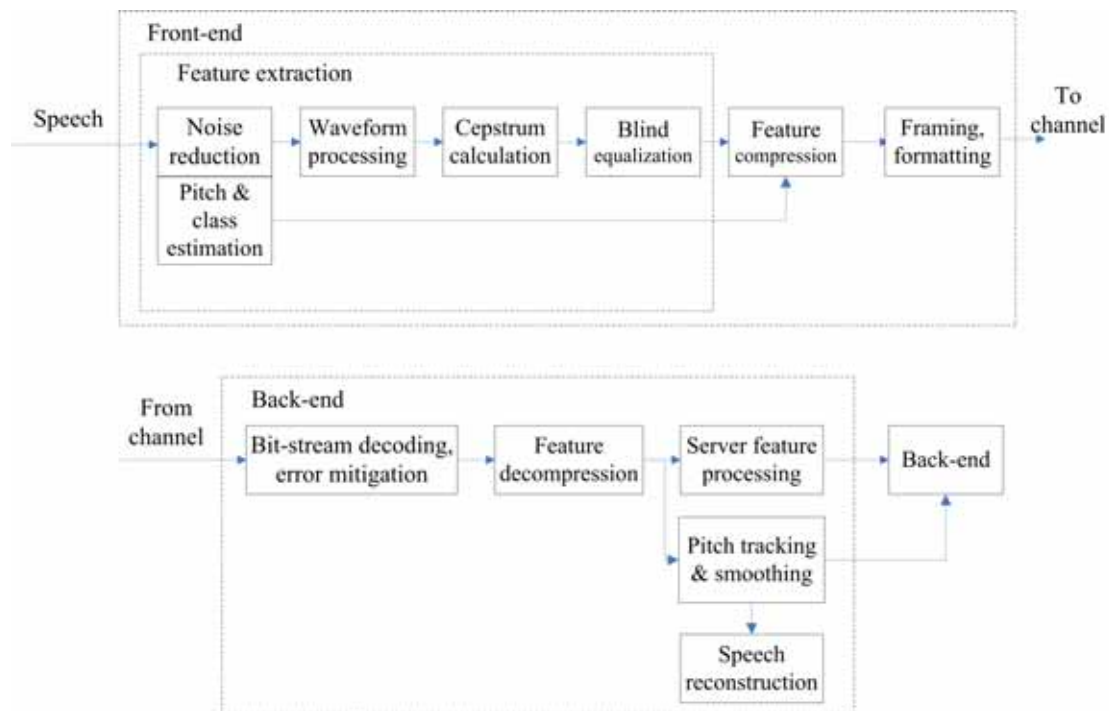


圖 2-1：分散式語音系統架構圖



當使用者在使用手持設備時，週遭普遍都有一些環境雜訊的干擾，為了因應此種情況，在分散式語音辨識系統前級在參數抽取之前特別加入了降低雜訊 (Noise reduction) 的處理。

在歐洲電信標準協會編號 202 212 V1.1.1 分散式語音辨識系統前級中，降低語音雜訊的處理方法是利用一個二階式維納濾波器 (Wiener filter)，如圖 2-2 所示，這是一種能夠有效降低雜訊的方法，圖 2-2 顯示它的方塊圖，它是由兩個串聯的維納濾波器組成，它們的輸出及輸入都是聲音的波形 (Waveform) 訊號；第一個維納濾波器的輸入是未經處理且帶有雜訊的語音波形訊號，輸出的是經過初步處理的語音波形訊號，它同時也是第二個維納濾波器的輸入波形訊號，第二個維納濾波器輸出的是已除去大部分雜訊的波形訊號。在第一個維納濾波器中包含了語音偵測的技術 (Voice Activity Detection, VAD)，用以進行雜訊頻譜的估測 (Noise spectrum estimation)，第二個維納濾波器則假設經過第



一個維納濾波器的處理，剩餘的加成性雜訊可以用白雜訊 (White noise) 近似，不再含語音偵測技術。兩個維納濾波器都是隨著各個音框 (Frame) 內不同的雜訊特性及訊噪比而設計的；首先依照不同頻率的訊噪比，得到線性頻率上維納濾波器的係數 (Linear-frequency Wiener filter coefficients)，再將其通過梅爾濾波器組 (Mel filter-bank) 以得到較平滑且和聽覺系統相關的梅爾維納濾波器係數 (Mel-warped Wiener filter coefficient)，接著將此梅爾維納濾波器係數作梅爾反離散餘弦轉換 (Mel-warped Inverse Discrete Cosine Transformation, Mel-warped IDCT)，以得到在時域上的脈衝響應 (Impulse response)，最後再把目前音框中的波形訊號通過此脈衝響應以得到輸出的波形訊號。在第二階維納濾波器輸出之前，有一個偏移補償 (Offset compensation) 的區塊，用以移除輸出波形中的直流偏移量 (DC offset)。

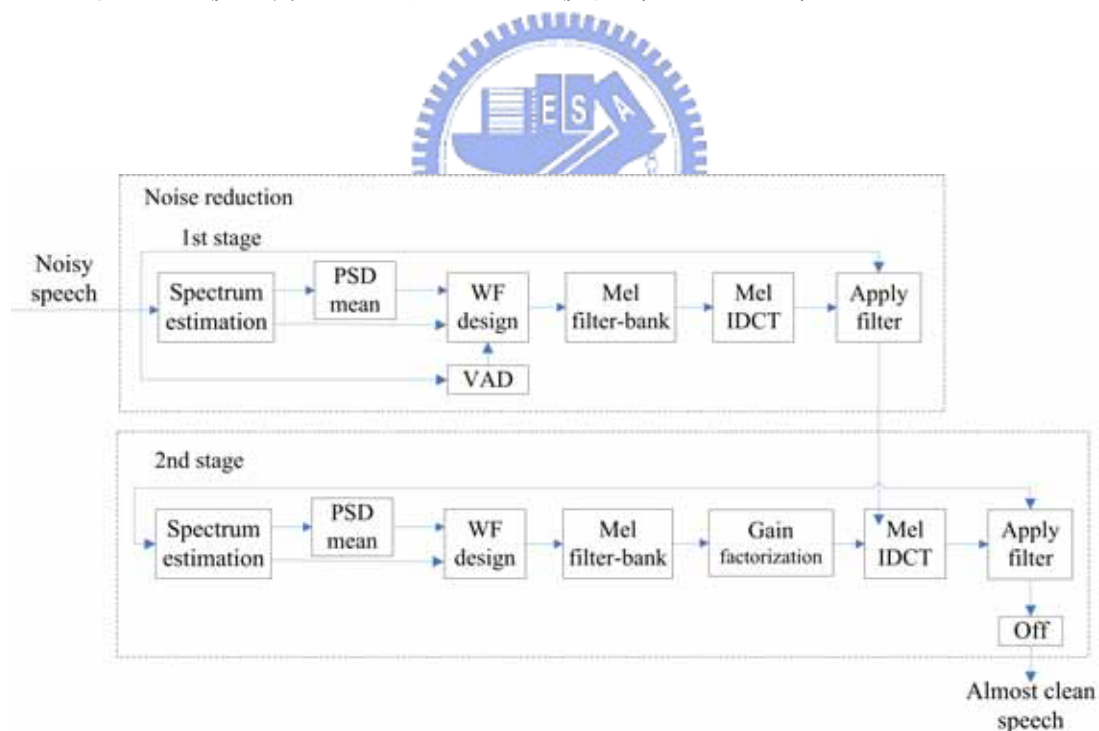


圖2-2：降低雜訊處理系統流程圖

在歐洲電信標準協會編號 202 212 V1.1.1 的系統架構中，也訂定了偵測基頻的方法，其用途可以用來做聲調語言之語音的辨識 (Tonal language

recognition) 以及重建語音訊號 (Speech reconstruction) ，所以在本論文中將會利用歐洲電信標準協會編號202 212 V1.1.1所偵測的基頻資訊來做國語連續語音的辨認。

## 2.2 分散式語音辨識系統環境下國語連續數字串之辨認

在一套新的語音辨識系統架構中，一開始大多選定連續數字串的辨識工作，因為連續數字串是有最多應用的一個語音辨識系統。在分散式語音辨識系統的環境下，加上在分散式語音辨識系統的環境下，都會考慮環境雜訊對語音辨識效能的影響。所以 AURORA-2 為例，它是由 TI-Digits 這套不含雜訊的英文連續數字串的語料，並以人工處理的方式加上了八種環境雜訊而成的語料。在乾淨語音的時候，辨識率已經高達 99.02%，但是加了環境雜訊之後，辨識率隨著訊噪比越低，下降越快，平均從訊噪比 20dB 到 0dB 辨認率大幅降低了 30%[2]。在本論文中將先建立一個在分散式語音辨識系統環境下國語連續數字串的基本辨識系統。

### 2.2.1 語料庫

在實驗中所使用的國語連續數字串語料庫，是一套由交通大學語音實驗室所錄製的麥克風語料。表 2-1 列出此套語料的錄製方式，取樣頻率、句數，以及語料統計特性。

表 2-1：國語連續數字串語料庫

錄製方式	麥克風
取樣頻率	16 kHz
編碼格式	16 位元 PCM
語料內容	男性語者和女性語者各 50 人，每人 10 句，共 1000 句，6,438 個數字
統計特性	每句有 1~11 個數字不等，平均每句含有 6~7 個數字

一般大眾使用的 GSM 手機，其內部對於聲音的取樣頻率，是依據傳統公眾交換電話網路（PSTN）取樣頻率為 8k Hz 的標準。為了相容於此標準，使我們的實驗更符合實際情況，所以將所取得的麥克風語料降頻（down-sample）為 8kHz。



## 2.2.2 環境雜訊

實際上當使用者在使用分散式的語音辨認系統時，系統的辨識率會受到使用者週遭的環境雜訊影響，為了使我們的實驗與實際狀況更符合，所以要在語料中加上環境雜訊。

在本論文中，環境雜訊是採用 AURORA 2 中提供的環境雜訊[3]，總共有八種環境雜訊（地下鐵、人聲、汽車、展覽會館、餐廳、街道、機場、火車站），取樣頻率是 8kHz，16 bit 的 PCM 檔案。表 2-2 表示每個環境雜訊的音檔長度。

表 2-2：八種環境雜訊的音檔長度

地下鐵	20:24
人聲	3:55:06
汽車	22:12
展覽會館	19:06
餐廳	4:46:12
街道	57:11
機場	2:59:29
火車站	2:59:29

在加入環境雜訊時，是以乾淨語料的長度為基準，隨機選擇一段環境雜訊與乾淨語料相同長度作相加的動作，但是八種環境雜訊的音長不盡相同，也不一定比乾淨語料還要長，所以又可以分成兩種情形：1. 乾淨語料的音長比環境雜訊的音長短；2. 乾淨語料的音長比環境雜訊的音長還長。

當乾淨語料的音長比環境雜訊的音長短的時候，便是直接以乾淨語料的長度為基準，隨機選擇一段與乾淨語料相同長度的環境雜訊，來與乾淨語料做相加的動作；若是乾淨語料的音長比環境雜訊的音長還長的時候，先重覆環境雜訊，直到環境雜訊的音長超過乾淨語料的音長，再以乾淨語料的長度為基準，隨機選擇一段與乾淨語料相同長度的環境雜訊，來與乾淨語料做相加的動作。圖 2-3 以圖示說明。在圖 2-3 中，S 為乾淨語料的音長，N 為環境雜訊的音長，L 是環境雜訊上與乾淨語料相加區段的起始點。

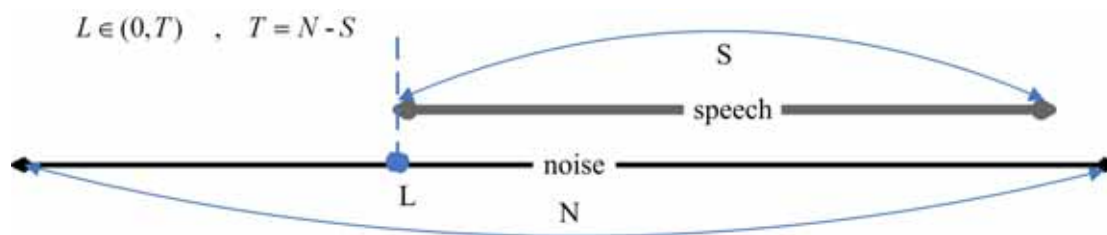


圖 2-3：在乾淨語料中加入環境雜訊示意圖

接著介紹當我們如何在乾淨語料中加上環境雜訊，並且控制訊噪比

(Signal-to-Noise Ratio, SNR) 在某一定值的方法。首先要先計算乾淨語料以及環境雜訊的平均能量 (Average Power)，其中乾淨語料只計算有語音部份的平均能量，環境雜訊只計算與乾淨語料相加部份的平均能量。平均能量可以下式表示：

$$P = \frac{1}{M} \sum_{i=1}^M x^2(i) \quad (2-1)$$

$P$  為平均能量， $M$  為取樣的個數， $x(i)$  代表是第  $i$  個取樣點的振幅大小。

在乾淨語料與環境雜訊相加時，想要控制訊噪比在某一定值，又因為訊噪比為語音訊號與環境雜訊能量大小的比值，即為聲音振幅大小的比值[4]，所以固定乾淨語料振幅的大小，只調整環境雜訊振幅的大小；將環境雜訊的振幅大小乘以  $G = \left( \frac{P_S}{P_N} 10^{\frac{-SNR}{10}} \right)^{\frac{1}{2}}$  倍，再與乾淨語料的振幅相加，即可控制乾淨語音訊號與環境雜訊相加後的訊噪比。

$$SNR = 10 \log(P_S) - 10 \log(P_N') \Rightarrow G = \left( \frac{P_S}{P_N} 10^{\frac{-SNR}{10}} \right)^{\frac{1}{2}} \quad (2-2)$$

其中  $P_S = \frac{1}{M} \sum_{i=1}^M x_S^2(i)$ ， $P_N = \frac{1}{K} \sum_{i=1}^K x_N^2(i)$ ， $P_N' = G^2 * P_N$

$SNR$  為乾淨語料與環境雜訊相加後的訊噪比， $P_S$  代表乾淨語料的平均能量， $P_N$  代表環境雜訊的平均能量， $P_N'$  代表調整過後的環境雜訊的平均能量。

圖 2-4是各種環境雜訊的長時間頻譜(long-term spectrum)圖，由此圖可看出：汽車雜訊、機場雜訊及火車站雜訊長時間平均頻譜在低頻處能量最高，隨著頻率增加，能量逐漸減少，至4000Hz(二分之一的取樣頻率)時的能量大小和能量最高處相差約有40dB；人聲雜訊、餐廳雜訊及街道雜訊的長時間頻譜特性大致和前述三種類似，但高頻及低頻能量的差距不像前述三種雜訊明顯，且能量峰值的位置亦較前述三種雜訊來的高；剩下兩種雜訊的特性則較為不同，地下鐵雜訊在500Hz及2500Hz這兩處能量都有明顯峰值，展覽會館雜訊和其他雜訊相比之下，其長時間頻譜則是較接近平坦的白雜訊特性。由圖 2-4只能觀察出各種雜訊長時間平均後的特性，卻無法得知其特性是否穩定(Stationary)。圖 2-5 則是它們的頻譜-時間圖(Spectrogram)橫軸及縱軸分別代表時間及頻率，較亮的顏色代表較強的能量，由此圖較易了解雜訊的穩定性如何；由此圖我們看到較穩定的雜訊(如：汽車雜訊及展覽會館雜訊)在任一時間點的頻譜都很接近其長時間頻譜；而不穩定的雜訊(如：街道雜訊、機場雜訊及火車站雜訊)，則隨著不同的時間點，可能有著變動很大的頻譜特性，所以其長時間頻譜和實際上的雜訊特性是有較多出入的。

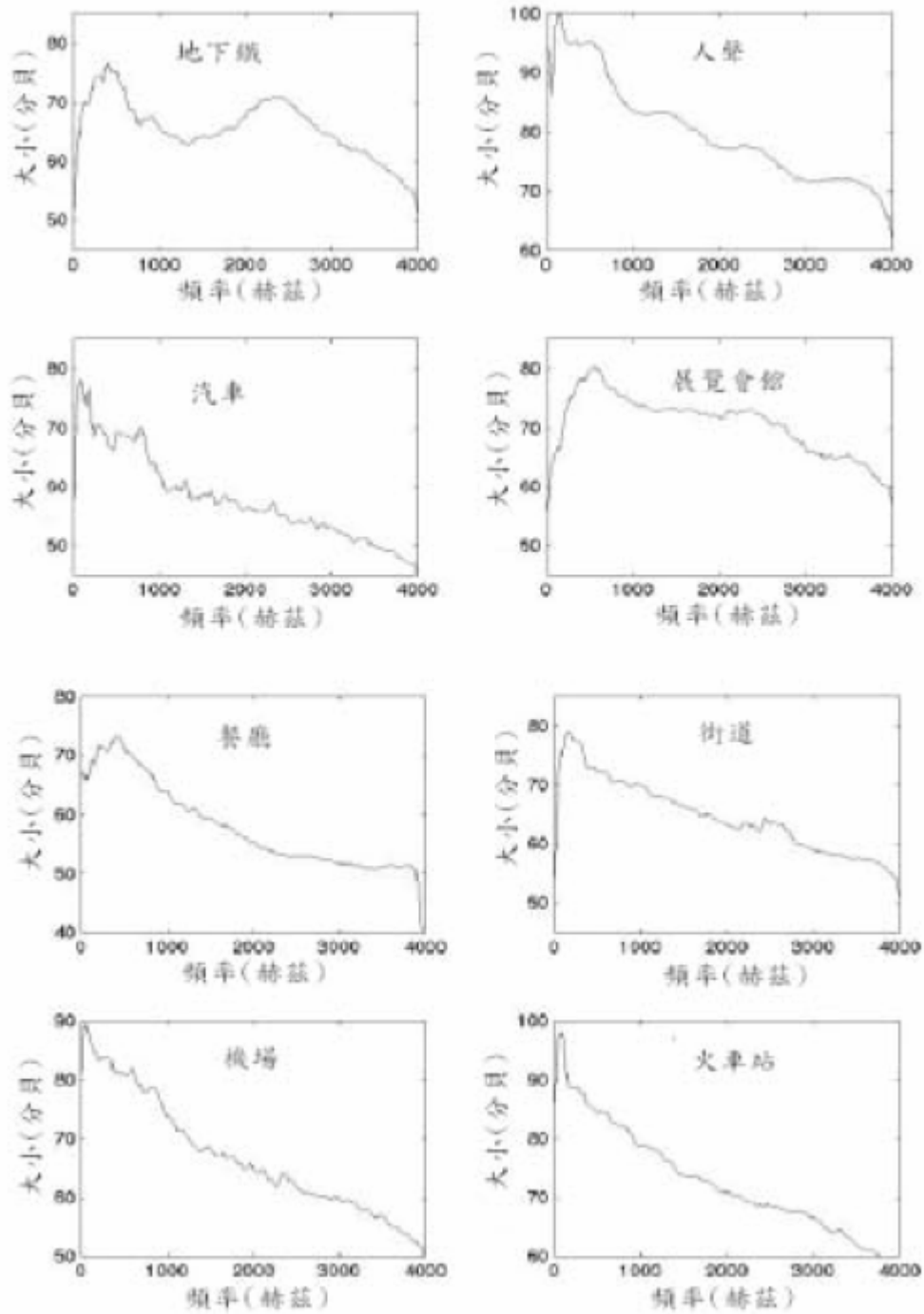


圖 2-4：八種環境雜訊的長時間頻譜

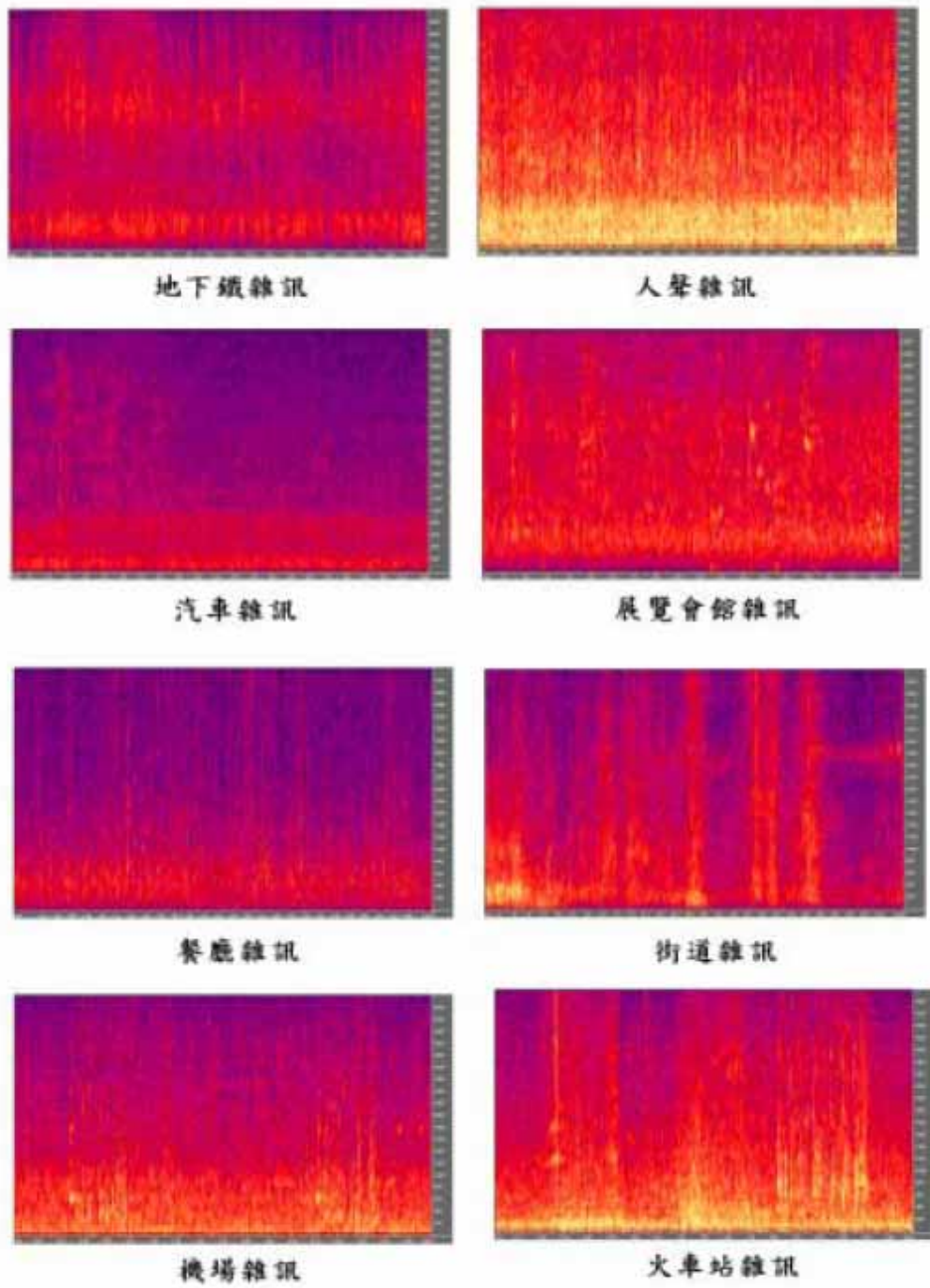



圖 2-5：八種環境雜訊的頻譜—時間圖（橫軸：時間；縱軸：頻率）



### 2.2.3 分散式語音辨識系統後級隱藏式馬可夫模型之語音辨識器

在實驗中，分散式語音辨識系統後級採用隱藏式馬可夫模型（Hidden Markov Model, HMM）語音辨識器。隱藏式馬可夫模型的產生也可以分成只用乾淨語料訓練，或是用加入不同的環境雜訊、以及不同訊噪比的語料做訓練，分別對應到「乾淨語料訓練」和「複合情境訓練」這兩種訓練模式；而且依照各種訊噪比加上八種不同的環境雜訊，按照所加環境雜訊的種類，分成 A、B 兩種測試組合（Testing set），其中 A 組所加入的環境雜訊是與訓練語料所加入之環境雜訊匹配（Match），B 組所加入的環境雜訊與是訓練語料所加入之環境雜訊不匹配（Mismatch）。詳細內容如表 2-3 所示。



在乾淨語音訓練模式中，將語料庫的十分之九當作訓練語料，其中男性語者和女性語者各 45 人，每人 10 句，共 900 句，5,796 個數字；在複合情境訓練模式中，因為語料數不夠的因素，所以將在乾淨語音訓練模式的 900 句的訓練語料，重複使用兩次，總共 1,800 句訓練語料，再平均分為 20 組，每組中沒有重複出現的句子，且每組分別是加入不同環境雜訊、不同訊噪比的情境。在兩種訓練模式中，都是將語料庫的另外十分之一當作測試語料，男性語者和女性語者各 5 人，每人 10 句，共 100 句，642 個數字。同樣也是有語料數不足夠的問題，所以將 100 句測試語料重複使用於各個不同的環境雜訊與不同的訊噪比的組合中，總共有 49 組測試組，分別是八種環境雜訊與六種訊噪比合併組合的 48 組，以及一組乾淨語料測試。

表 2-3：加上環境雜訊的國語連續數字串內容介紹

國語連續數字串語料庫		
取樣頻率	8 kHz	
訓練模式	乾淨語音訓練	複合情境訓練
	音段數：900 環境雜訊： 無	音段數：1,800 環境雜訊： ● 種類：地下鐵、人聲、汽車、展覽會館 ● 訊噪比：20dB、15dB、10dB、5dB 和完全乾淨 ● 4 種雜訊乘以 5 種 SNR，共 20 種情境
測試組合	A 組	B 組
	音段數：2,800 環境雜訊： 地下鐵 人聲 汽車 展覽會館	音段數：2,800 環境雜訊： 餐廳 街道 機場 火車站
	對於上述的每種環境雜訊，訊噪比都控制在 20dB、15dB、10dB、5dB、0dB、-5dB 以及完全乾淨七種程度，並且對於每種雜訊的每個訊噪比程度都計算一組辨識結果	

本實驗使用的語音辨識參數是 12 維梅爾倒頻譜係數(Mel Frequency Cepstral Coefficients, MFCC)，加上一維與二維的變化量，以及能量的一維與二維的變化量，共 38 維特徵向量。表 2-4 列出特徵參數抽取過程中各項參數設定。其中前五項是分散式語音辨識系統前級的標準設定，而語音特徵向量之選取則是後級隱藏式馬可夫模型辨識器之設定。

表 2-4：語音特徵參數抽取之參數設定

取樣頻率(Sampling rate)	8 kHz
音框長度(Frame window size)	25 ms
音框平移量(Frame window shift)	10 ms
預強調的轉換函數(Pre-emphasis)	$1-0.9z^{-1}$
梅爾濾波器組(Mel-frequency filter bank)	23 個濾波器
語音特徵向量(Speech feature vector)	38 維 (靜態 [12-MFCCs, log E ]、一次及二次動態係數)

隱藏式馬可夫語音辨識模型的建立則詳述如下：首先建立國語數字從 0 到 9 的聲學模型，每個聲學模型設定為 8 個狀態 (State)，每個狀態含有 8 個混合高斯數 (Mixtures)；除了國語數字的聲學模型外，還有兩個模型——靜音模型 (Silence model) 與停頓模型 (Short pause model) 的聲學模型，是用來描述語音信號中靜音部分，其中靜音聲學模型是描述句首和句尾之靜音，設定為 3 個狀態，停頓聲學模型則用來描述字與字之間的靜音，設定為 1 個狀態，此狀態允許跳躍 (Skip)，並且與靜音模型的中間狀態合併 (Tying)，兩個聲學模型中每個狀態則含有 16 個混合高斯數。

## 2.2.4 實驗結果

表 2-5(a)與表 2-5(b)分別列出乾淨語音訓練模式與複合情境模式下的辨識結果。其中各種環境雜訊下之平均辨識率是依照 AURORA-2 平均辨識率的計算方式，只對訊噪比 20dB 到 0dB 環境下的辨識率做平均。

表 2-5(a)：國語連續數字串---乾淨語音訓練模式之辨識結果

乾淨語音訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	98.1%				
20	94.9%	93.3%	97.0%	94.7%	95.0%
15	90.3%	91.7%	95.6%	91.4%	92.3%
10	84.4%	87.5%	93.8%	84.7%	87.6%
5	66.7%	77.4%	86.0%	70.6%	75.2%
0	41.1%	52.0%	60.4%	42.1%	48.9%
-5	16.4%	20.1%	19.9%	15.4%	18.0%
平均值(20dB~0dB)	75.5%	80.4%	86.6%	76.7%	79.8%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	98.1%				
20	90.0%	95.5%	90.5%	95.3%	92.8%
15	86.1%	94.4%	89.4%	94.9%	91.2%
10	80.7%	87.4%	86.6%	90.3%	86.3%
5	67.0%	80.2%	81.5%	86.3%	78.8%
0	48.6%	48.3%	57.2%	65.4%	54.9%
-5	21.7%	24.0%	31.3%	38.6%	28.9%
平均值(20dB~0dB)	74.5%	81.2%	81.0%	86.4%	80.8%
八種環境雜訊及五種訊噪比的平均值					80.3%

表 2-5(b)：國語連續數字串---複合情境訓練模式之辨識結果

複合情境訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	96.6%				
20	95.8%	96.9%	97.2%	97.2%	96.8%
15	95.0%	96.0%	96.9%	95.5%	95.9%
10	89.1%	93.0%	96.1%	91.1%	92.3%
5	77.1%	84.1%	91.4%	80.7%	83.3%
0	48.0%	62.0%	68.1%	52.7%	57.7%
-5	17.3%	25.9%	26.2%	15.7%	21.3%
平均值(20dB~0dB)	81.0%	86.4%	89.9%	83.4%	85.2%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	96.6%				
20	94.2%	95.5%	92.7%	95.5%	94.5%
15	94.1%	95.0%	94.7%	96.0%	95.0%
10	88.6%	89.6%	92.7%	93.8%	91.2%
5	78.5%	84.6%	86.3%	90.5%	85.0%
0	56.9%	53.4%	69.6%	76.5%	64.1%
-5	25.7%	27.4%	40.7%	46.7%	35.1%
平均值(20dB~0dB)	82.5%	83.6%	87.2%	90.5%	86.0%
八種環境雜訊及五種訊噪比的平均值					85.6%

從實驗結果中，我們可以獲得以下觀察：

- (1) 乾淨語音訓練模式的辨識率幾乎都是比複合情境訓練模式的辨識率還差，這和我們的預期是一致的，因為複合式情境訓練模式所訓練出來的聲學模型跟測試語料較匹配的原因；只有在沒有任何環境雜訊的測試情形下，乾淨語音訓練模式的辨識率比複合情境訓練模式的辨識率好，這是因為此時複合情境訓練模式所產生的聲學模型反而和測試語料存在較大的不匹配了。
- (2) 火車站與汽車環境雜訊有比較多低頻聲音，所以對語音辨認的影響較小。

(3) 比較在不同的環境雜訊之下，兩種訓練模式的差異。從乾淨語料訓練模式到複合情境訓練模式，辨識率提高最多的是加了餐廳環境雜訊的情況，次之的是加了展覽會館環境雜訊的情況；進步最少的是加了街道環境雜訊的情況。如圖 2-6 所示。

(4) 兩種訓練模式都是隨著訊噪比越低，辨識率也會越低，當訊噪比低於 5dB 時，辨認率會急速下降；而且在訊噪比在 10dB 以上時，測試組合的 A 組的辨識率，都比測試組合的 B 組還要高；但是訊噪比在 5dB 以下，情形便顛倒過來，測試組合的 A 組的辨識率，都比測試組合的 B 組還要低。如圖 2-7 所示。

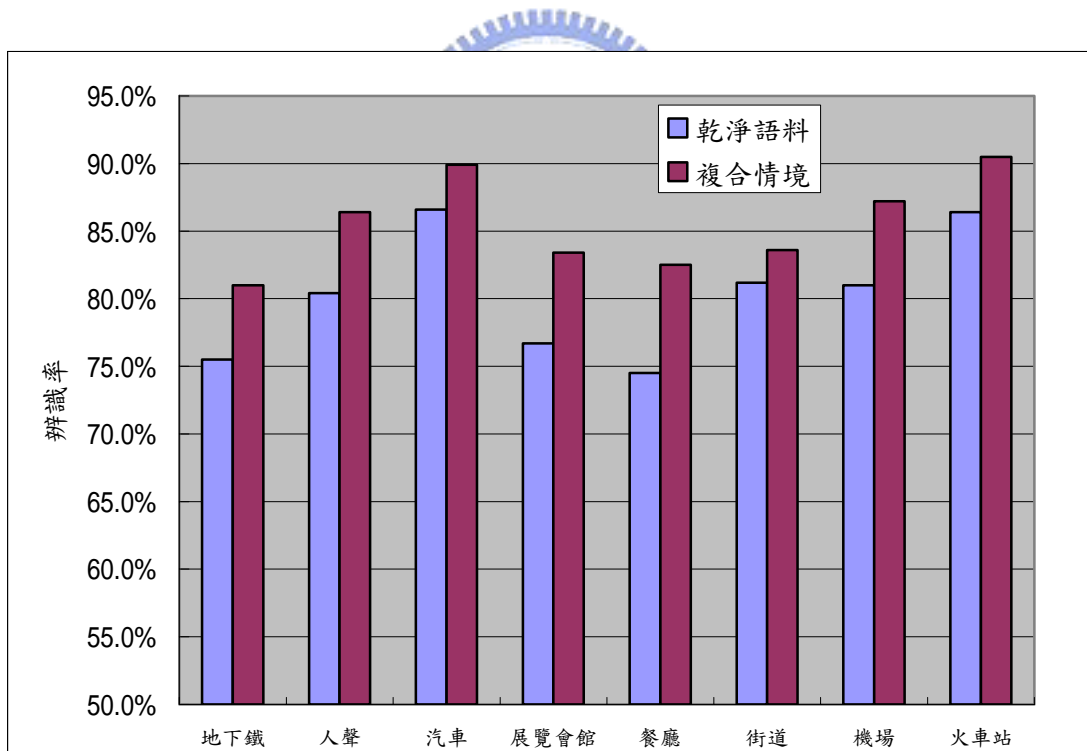


圖 2-6：八種環境雜訊在兩個訓練模式的比較

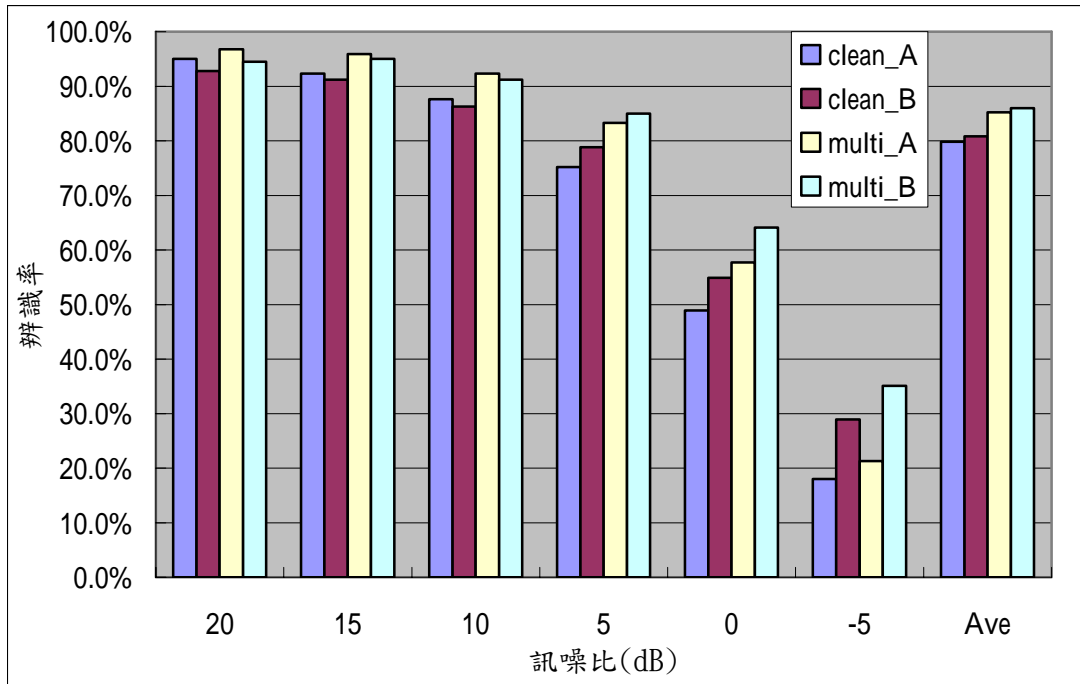


圖 2-7：不同的訊噪比在兩個訓練模式的比較



## 第三章 使用基頻參數的分散式國語連續語音辨識系統

從上一章中，我們知道在有環境雜訊的情況下，語音的辨識率會隨著訊噪比越低而降低，而且在訊噪比為 5dB 以下時，辨識率下降的情況更為嚴重。為了減緩這種情況發生，本章中我們將加上在歐洲電信標準協會編號 202 212 V1.1.1 的分散式語音系統架構中一項新的參數---「基頻」，相信如果將這項基頻參數使用在辨識器中，對國語語音應當可以獲得辨識率的增益[5]。本章說明如何使用分散式語音辨認系統中求得的基頻參數，建立一個帶聲調的國語連續數字串辨認器，以及所做的實驗與分析。

### 3.1 分散式語音辨識系統之基頻抽取



在這一節將先介紹歐洲電信標準協會編號 202 212 V1.1.1 的分散式語音系統架構中基頻參數，是如何求得的。接著再介紹在聲調辨識器中所使用的基頻參數。

在分散式語音辨識系統的前級中的「Pitch & class estimation」，參考圖 2-1，是用來估計基頻以及語音的狀態資訊。當語音信號數入，中間經過波形處理(Waveform processing)、計算頻譜及能量(Spectrum and energy computation; SEC)，基頻以及是否為語音的預先估計 (Pre-processing for pitch and class estimation; PP)；梅爾濾波器組---用以得到較平滑且和聽覺系統相關的梅爾維納濾波器係數，之後再經過語音偵測處理，得到哪一段是語音、哪一段不是語音的資訊；低頻的雜訊偵測 (Low-band noise detection; LBND) ---偵測在低頻中哪一個音框有背景雜訊，用以預先加強由 PP 求得的功率頻譜；然後在經由最後的基頻估計 (Pitch estimation; PITCH) 得到最後的基頻值；最後再由「CLS」



得到最後的語音狀態資訊。其系統方塊圖於圖 3-1。

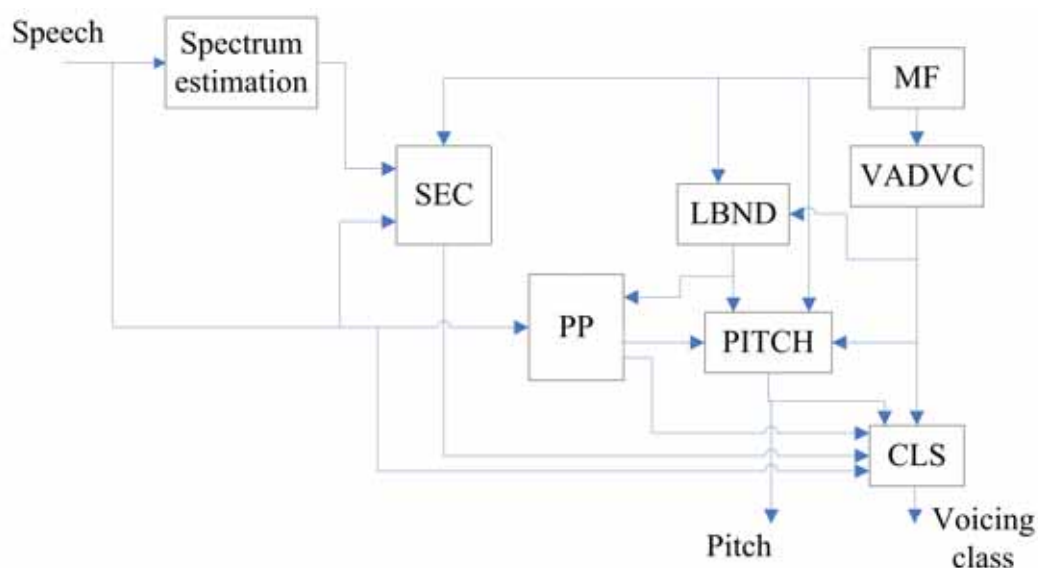


圖 3-1：分散式語音辨識前級估計基頻與語音狀態資訊系統架構

在歐洲電信標準協會編號 202 212 V1.1.1 標準中，基頻軌跡的追蹤是在分散式語音辨識系統的後級處理，如此可以補償一些由於傳輸錯誤所造成的基頻資訊錯誤。

在語音辨識器中，我們使用了連續隱藏式馬可夫模型 (Continuous HMM)，必須將沒有基頻值的語音信號補一個非零的值，這樣才能夠避免基頻參數的觀察機率發生不連續性的現象。所以本論文中將由 ETSI 202 212 V1.1.1 的 DSR 架構中所求得的基頻參數，取其對數 ( $\log-F_0$ )。接著再利用指數函數 (Exponential function)，將補償 (Interpolation) 介於兩段語音中間屬於無聲音 (Unvoiced) 的音框 (Frame)，以及每一個句子頭尾兩段沒有語音的音框 [6]。補償第  $n$  個音框非語音的基頻值，式子如下所示：

$$\log(f_0[n]) = \text{MAX}\left(\log(f_0[b]) \cdot e^{-\alpha(n-b)}, \log(f_0[g]) \cdot e^{-\alpha(g-n)}\right) \quad (\text{式 3-1})$$

其中  $b$  是代表上一個有語音的音框編號， $g$  是下一個有語音的音框編號， $\alpha$  是衰減係數，在本論文當中設定  $\alpha = 0.95$ 。

舉個例子來看經過補償後的基頻值，這個例子是加上地下鐵的環境雜訊，且訊噪比設定在 20dB 的條件下。由圖 3-2 可看出基頻 (F0) 在經過指數函數補償前與補償後的區別。

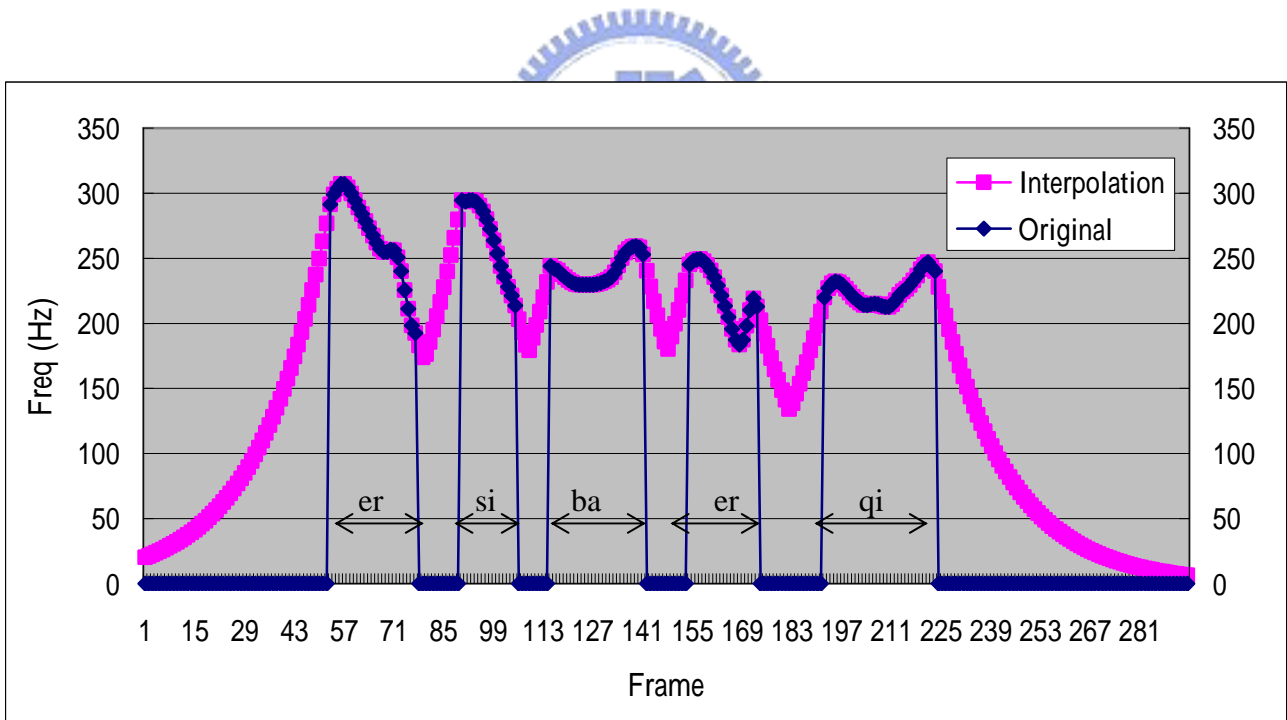


圖 3-2：經過指數函數補償的基頻

## 3.2 在不同環境雜訊之基頻偵測分析

為了要瞭解究竟在不同的環境雜訊以及在不同的訊噪比的條件之下

，對分散式語音辨識系統之基頻參數的抽取有什麼影響。表 3-1(a)到表 3-1(h)中，分別表示八種不同的環境雜訊在六種訊噪比之下與乾淨環境下基頻資訊的差異分析比較：在同一環境雜訊底下，六種不同的訊噪比分別與乾淨語料所求出的基頻做的比較。其中，將基頻錯誤分成了三類，其中前兩種是偵測到有聲音 (Voiced) 和無聲音 (Unvoiced) 的分類錯誤；第三類才是在有聲音的狀態下，基頻參數求取的錯誤。

1. U->V 判別錯誤：乾淨語音所求得的基頻屬於無聲音 (pitch = 0)，而加上環境雜訊後的語音所求得的卻是有聲音的 (pitch 有值)。
2. V->U 判別錯誤：乾淨語音所求得的基頻是有聲音的 (pitch 有值)，而加上環境雜訊後的語音所求得的基頻卻是屬於無聲音 (pitch = 0)。
3. V->V 相對錯誤：是指乾淨語音所求的基頻值 (pitch ≠ 0)，與加上環境雜訊後的語音所求得的基頻值 (pitch ≠ 0) 的相對錯誤。以均方誤差 (Mean square error; MSE) 表示兩者的相對錯誤，計算公式為

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N \left( \frac{P_i - P'_i}{P_i} \right)^2 \quad (\text{式 3-1})$$

其中， $P$  為乾淨語音所求得的基頻值； $P'$  為加上環境雜訊後的語音所求得的基頻值； $i$  是音框編號； $N$  是在這段語音的音框數。

表 3-1(a)：基頻在地下鐵環境下不同訊噪比之分析

訊噪比 (dB)	U->V	V->U	V->V 相對錯誤
20	1.84%	7.84%	0.087
15	1.15%	11.48%	0.119
10	0.38%	19.41%	0.200
5	0.25%	38.85%	0.404
0	0.09%	70.05%	0.719
-5	0.09%	94.10%	0.947

表 3-1(b)：基頻在嘈雜的人聲環境下不同訊噪比之分析

訊噪比 (dB)	U->V	V->U	V->V 相對錯誤
20	3.60%	9.07%	0.093
15	4.61%	12.80%	0.141
10	4.41%	21.91%	0.242
5	5.65%	41.92%	0.457
0	5.48%	70.08%	0.729
-5	4.80%	89.27%	0.905

表 3-1(c)：基頻在汽車環境下不同訊噪比之分析

訊噪比 (dB)	U->V	V->U	V->V 相對錯誤
20	2.29%	10.44%	0.108
15	2.89%	16.30%	0.182
10	2.56%	29.06%	0.336
5	2.36%	53.70%	0.604
0	2.46%	78.93%	0.839
-5	2.39%	94.00%	0.958

表 3-1(d)：基頻在展覽會場環境下不同訊噪比之分析

訊噪比 (dB)	U->V	V->U	V->V 相對錯誤
20	2.35%	7.44%	0.081
15	1.38%	10.18%	0.111
10	0.82%	15.40%	0.168
5	0.47%	30.78%	0.322
0	0.43%	62.68%	0.648
-5	0.48%	93.57%	0.942

表 3-1(e)：基頻在餐廳環境下不同訊噪比之分析

訊噪比 (dB)	U->V	V->U	V->V 相對錯誤
20	4.27%	7.90%	0.084
15	4.99%	11.43%	0.121
10	5.16%	18.89%	0.200
5	6.66%	36.64%	0.396
0	6.30%	63.95%	0.671
-5	6.38%	87.07%	0.892

表 3-1(f)：基頻在街道環境下不同訊噪比之分析

訊噪比 (dB)	U->V	V->U	V->V 相對錯誤
20	2.88%	8.47%	0.094
15	3.71%	13.10%	0.144
10	4.08%	21.95%	0.244
5	4.22%	40.53%	0.450
0	3.40%	70.86%	0.739
-5	3.01%	86.80%	0.894

表 3-1(g)：基頻在機場環境下不同訊噪比之分析

訊噪比 (dB)	U->V	V->U	V->V 相對錯誤
20	5.78%	8.92%	0.105
15	4.87%	14.65%	0.162
10	5.61%	25.07%	0.272
5	6.31%	47.85%	0.526
0	6.43%	73.61%	0.784
-5	8.89%	87.47%	0.924

表 3-1(h)：基頻在火車環境下不同訊噪比之分析

訊噪比 (dB)	U->V	V->U	V->V 相對錯誤
20	6.08%	9.62%	0.107
15	8.03%	15.60%	0.186
10	10.48%	24.54%	0.328
5	11.84%	43.55%	0.548
0	11.47%	67.07%	0.759
-5	13.46%	78.25%	0.864

從表 3-1 可以看到隨著訊噪比越低，V->U、U->V 的錯誤率會越來越高，且 V->V 基頻相對錯誤也會越來越大；在三種錯誤分析中，又以 V->U 的錯誤機率隨著訊噪比下降，上昇的十分快，顯示出語音信號的週期特性非常容易受到環境雜訊的破壞。也就是說訊噪比越低，基頻受到環境雜訊的影響會越大，所求得的基頻也會越來越不可靠。在八種環境雜訊中，又以地下鐵、汽車、展覽會場受到的影響最大。

平均觀察八種環境雜訊在六種訊噪比的條件下所做的分析，只觀察基頻在不同的訊噪比的不變化。結果在表 3-2 中。

表 3-2：基頻在不同訊噪比之分析

訊噪比 (dB)	U->V	V->U	V->V 相對錯誤
20	3.64%	8.71%	0.095
15	3.95%	13.19%	0.146
10	4.19%	22.03%	0.249
5	4.72%	41.73%	0.464
0	4.51%	69.65%	0.736
-5	4.94%	88.82%	0.916

從表 3-2 中可以更加清楚地看到在不同訊噪比時的基頻錯誤變化，發現到在訊噪比為 5dB、0dB 和 -5dB 時，V->U 的錯誤率幾乎是倍數在增加，也就是說基頻在這些條件下時，幾乎是訊噪比越低，基頻的資訊就會受到環境雜訊的影響而損失一半。



挑一句音檔當例子，分別比較在乾淨語音與汽車環境雜訊中訊噪比 0dB 的情形下，由分散式語音辨識系統之基頻偵測器所求出的基頻值，音檔內容為---24827。如圖 3-3 所示。從中可以觀察到因為分散式語音辨識系統後級有做 Pitch tracking，會造成大部分由加入汽車環境雜訊在訊噪比 0dB 時求取的基頻區段整段不見；而且其有求出基頻值之基頻區段的頻率會減半。

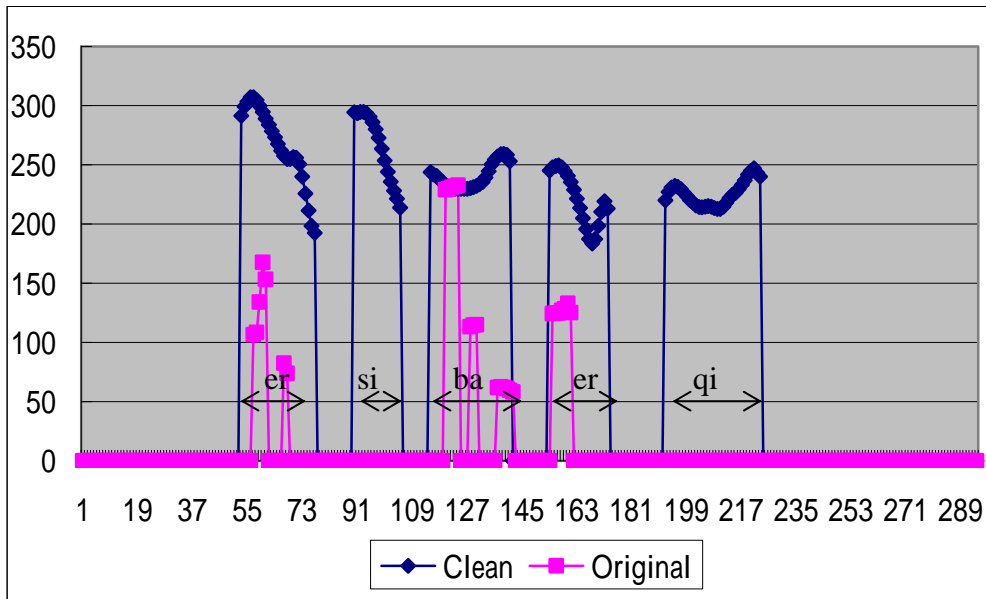


圖 3-3：比較在乾淨語音與汽車環境雜訊中訊噪比為 0dB 的基頻值

### 3.3 加入基頻參數之國語連續數字串之辨識

本章實驗是上一章實驗的延伸，主要是在所使用的辨認參數的不同，在這裡多加入了基頻相關參數的資訊。最後會將辨認結果會再與上一章的結果做比較。

#### 3.3.1 實驗設定

本實驗用的語料庫是國語連續數字串；實驗設定分為乾淨語料訓練以及複合情境訓練兩組模式，都與上一章實驗相同。

在分散式語音辨識系統前級中的參數抽取的各項參數設定，皆與分散式語音辨識系統前級的標準設定相同；而使用的語音參數除了 38 維梅爾倒頻譜係數之外，還加入了基頻參數，以及一維、二維的基頻參數變化量，總共使用 41 維的語音特徵向量。



隱藏式馬可夫語音辨認模型的建立則詳述如下：首先建立國語數字從 0 到 9 的聲學模型，每個聲學模型設定為 8 個狀態，每個狀態含有 16 個混合高斯數；除了國語數字的聲學模型外，還有兩個模型——靜音與停頓的聲學模型，是用來描述語音信號中靜音部分，其中靜音聲學模型是描述句首和句尾之靜音，設定為 3 個狀態，停頓聲學模型則用來描述字與字之間的靜音，設定為 1 個狀態，此狀態允許跳躍 (Skip)，並且與靜音模型的中間狀態合併 (Tying)，兩個聲學模型中每個狀態則含有 32 個混合高斯數。



### 3.3.2 實驗結果

表 3-3(a)與表 3-3(b)列出加入基頻參數後的國語連續數字串之乾淨語音訓練模式與複合情境訓練模式的辨識結果。

表 3-3(a)：加入基頻參數後的國語連續數字串之乾淨語音訓練模式辨識結果

乾淨語音訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	98.6%				
20	96.4%	97.7%	98.0%	97.4%	97.4%
15	93.9%	95.5%	96.6%	94.4%	95.1%
10	88.0%	90.5%	89.9%	89.3%	89.4%
5	67.3%	75.1%	70.1%	73.5%	71.5%
0	34.7%	40.8%	36.1%	38.5%	37.5%
-5	14.5%	12.3%	14.0%	12.3%	13.3%
平均值(20dB~0dB)	76.1%	79.9%	78.1%	78.6%	78.2%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	98.6%				
20	96.1%	97.4%	96.4%	97.4%	96.8%
15	93.9%	95.6%	94.9%	95.8%	95.1%
10	88.3%	88.9%	90.0%	90.7%	89.5%
5	70.9%	76.5%	74.3%	75.6%	74.3%
0	41.0%	38.6%	39.7%	47.4%	41.7%
-5	15.0%	17.8%	16.7%	25.2%	18.7%
平均值(20dB~0dB)	78.0%	79.4%	79.1%	81.4%	79.5%
八種環境雜訊及五種訊噪比的平均值					78.9%

表 3-3(b)：加入基頻參數後的國語連續數字串之複合情境訓練模式辨識結果

複合情境訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	96.3%				
20	97.7%	97.2%	98.3%	97.7%	97.7%
15	96.1%	96.7%	97.4%	97.8%	97.0%
10	92.4%	93.8%	94.1%	92.8%	93.3%
5	78.7%	83.2%	80.8%	81.9%	81.2%
0	46.6%	52.8%	51.6%	49.5%	50.1%
-5	17.8%	19.3%	18.9%	15.4%	17.9%
平均值(20dB~0dB)	82.3%	84.7%	84.4%	83.9%	83.8%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	96.3%				
20	96.4%	98.0%	95.6%	97.5%	96.9%
15	95.3%	96.7%	95.8%	96.9%	96.2%
10	92.4%	91.3%	91.7%	92.4%	92.0%
5	77.9%	84.4%	81.9%	82.6%	81.7%
0	50.5%	48.8%	52.0%	60.6%	53.0%
-5	20.9%	25.7%	23.5%	37.9%	27.0%
平均值(20dB~0dB)	82.5%	83.8%	83.4%	86.0%	83.9%
八種環境雜訊及五種訊噪比的平均值					83.9%

從實驗結果中，可以發現：

- (1) 比較乾淨語音訓練與複合情境訓練模式，在加入基頻資訊之後的實驗結果：與沒有加入基頻資訊的實驗結果類似，一樣只有在沒有加入任何環境雜訊的測試情形下，乾淨語音訓練模式的辨識率比複合情境訓練模式的辨識率好；其他組的測試情況，都是乾淨語音訓練模式的辨識率比複合情境訓練模式的辨識率還差。

- (2) 比較在不同的環境雜訊之下，兩種訓練模式的差異。從乾淨語料訓練模式到複合情境訓練模式，辨識率進步最多的是在加了地下鐵環境雜訊的情況，次之的是加了汽車環境雜訊的情況；進步最少的是加了機場環境雜訊的情況。這個結果與沒有加入基頻資訊的實驗作比較，發現差異很大，所以將在有加入基頻資訊的實驗結果與沒有加入基頻資訊的實驗結果中，將八種環境雜訊的進步程度詳細比較，列於表 3-4 中。我們可以發現：沒有加入基頻資訊前，進步最多的情形是在測試組合的 B 組，而前四名中有兩個 B 組的，後四名中有兩個 A 組的；加入基頻資訊之後，進步情形的前四名都是在測試組合 A 組當中，後四名則是在測試組合 B 組中，這跟預測的情況是相同的，因為 A 組是屬於與訓練情境匹配的環境雜訊，B 組是屬於與訓練情境不匹配的，所以 A 組進步的程度應該是要比 B 組多。
- (3) 個別觀察八種環境雜訊的情況：在乾淨語音訓練模式—只有在地下鐵、展覽會館跟餐廳這三種環境雜訊底下，有加入基頻資訊實驗的辨識率才會比沒有加入基頻資訊實驗的辨識率高；其他五種環境雜訊都是有加入基頻資訊實驗的辨識率比沒有加入基頻資訊實驗的辨識率還低，其中又以加入汽車環境雜訊的情況的辨識率下降最多，次之的是加入火車站環境雜訊的情況。在複合情境模式—在地下鐵、展覽會館跟街道這三種環境雜訊底下，有加入基頻資訊實驗的辨識率才會比沒有加入基頻資訊實驗的辨識率高；在餐廳這項環境雜訊的情況下是不變的；其他四種環境雜訊都是有加入基頻資訊實驗的辨識率比沒有加入基頻資訊實驗的辨識率還低，其中又以加入汽車環境雜訊的情況的辨識率下降最多，次之的是加入火車站環境雜訊的情況與乾淨語音訓練模式一樣。如圖 3-4 所示。
- (4) 訊噪比大於 5dB 時，加入基頻資訊的實驗結果幾乎都獲得比沒有加入基頻資訊的實驗結果還高的辨識率；而在訊噪比 5dB 以下時，加入基頻資訊實驗的辨識率幾乎都比沒有加入基頻資訊的辨識率還低，事實上與前一節基頻抽取器效能之分析結果一致，在訊噪比越低時，基頻參數受到環境雜訊

影響越大。整體的平均辨識率(20dB~0dB)都比沒有加入基頻資訊還要低。  
如圖 3-5 所示。

表 3-4 列出八種環境雜訊在兩個實驗中的進步情形從乾淨語音訓練模式到  
複合情境訓練模式的進步情形。

表 3-4：八種環境雜訊在兩個實驗中的進步情形

測試組合	A 組				B 組			
	地下鐵	人聲	汽車	展覽會館	餐廳	街道	機場	火車站
有基頻	8.15%	6.01%	8.07%	6.74%	5.77%	5.54%	5.44%	5.65%
沒有基頻	7.28%	7.46%	3.81%	8.74%	10.74%	2.96%	7.65%	4.75%

圖 3-4，圖 3-5 分別是表示八種環境雜訊與六種訊噪比在沒有加入基頻資訊  
的實驗 (non-pitch)、有加入基頻資訊的實驗 (pitch) 以及乾淨語音訓練模式  
(clean) 和複合情境模式 (multi) 間的辨識結果比較。

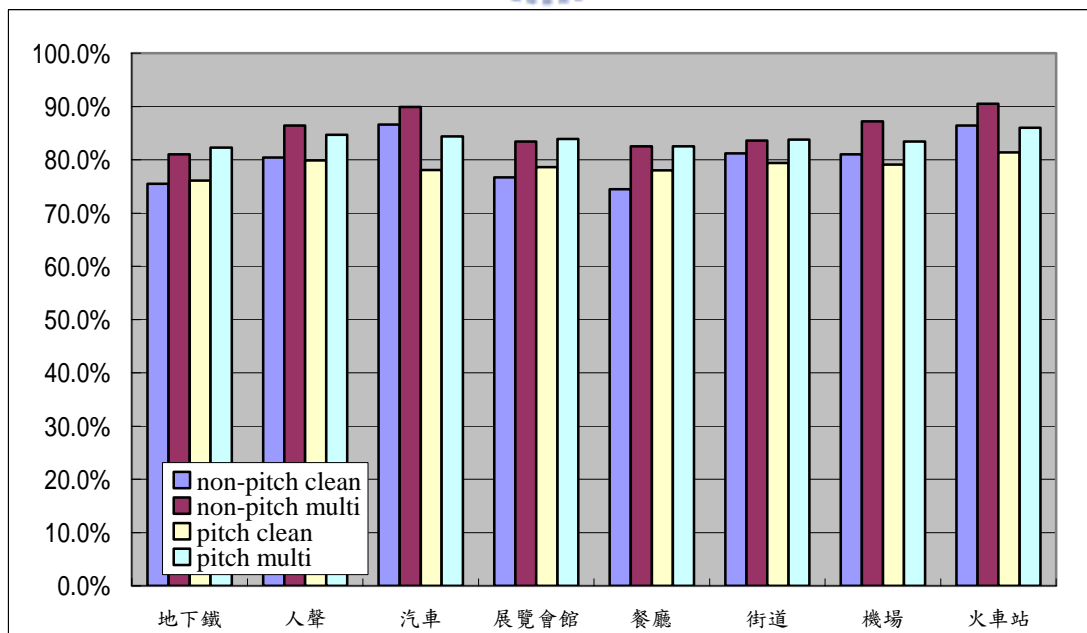


圖 3-4：八種環境雜訊的辨識結果比較

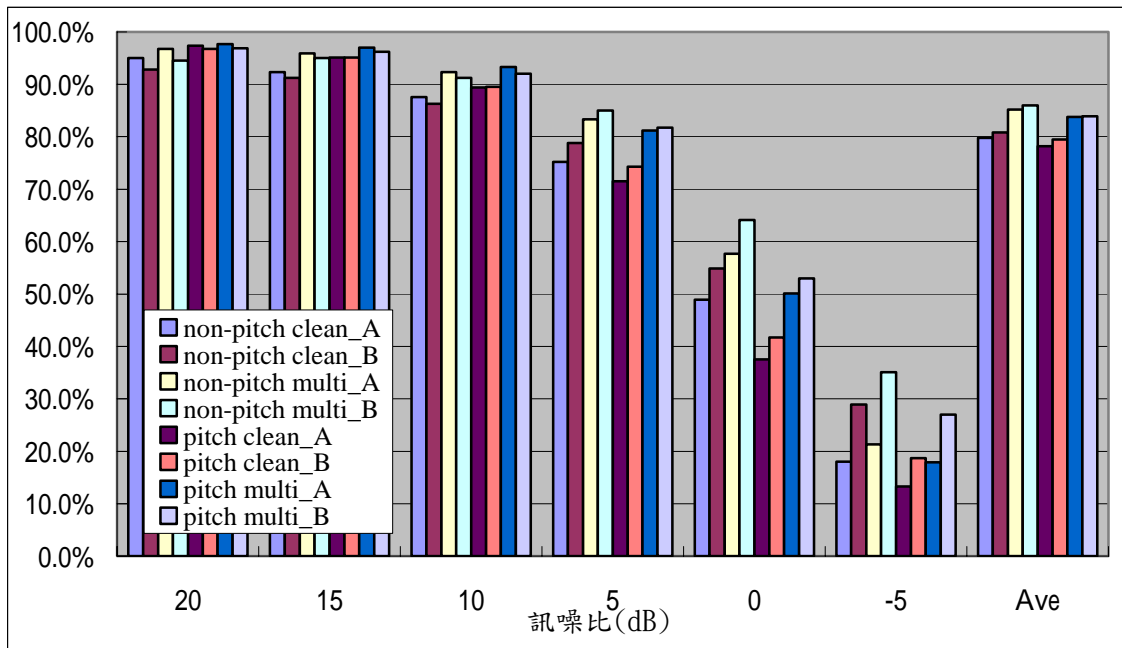


圖 3-5：不同的訊噪比的辨識結果比較



## 第四章 改良基頻參數抽取的方法

從前一章中可以看到歐洲電信標準協會編號 202 212 V1.1.1 標準中所提出的基頻偵測器在低訊噪比時，效能下降的很多，所以多加入了基頻這項資訊建立辨認器，雖然在訊噪比較高的時候，可以有效的提升辨識率；但是在訊噪比較低的時候，辨識率反而降低了。所以本章要介紹的是如何在小幅修改歐洲電信標準協會編號 202 212 V1.1.1 標準中的基頻偵測器之前提下，改進基頻參數抽取的方法，來增加辨識率。

### 4.1 DSR 基頻參數抽取之改良

原本分散式語音辨識系統抽取基頻參數的作法 (ETSI XAFE) 是先在前級接收到語音信號經過波形處理後，做基頻的估計，然後在後級再做一次基頻資訊的更正與軌跡的追蹤。可是在前級中，已經有針對降低環境雜訊做了兩次維納濾波器，以獲得較為乾淨的語音訊號。於是便想要利用經過降低環境處理過後的語音訊號來抽取基頻參數，如此一來在有環境雜訊時，應該可以得到更好的基頻參數。改良式的參數抽取作法 (Modified XAFE) 就是在接收到語音信號後，先做一次降低雜訊的動作，也就是先經過二階式維納濾波器降低雜訊的干擾後，再做基頻參數估計；在分散式語音辨識系統的後級的部分，則是沒有改變作法。圖 4-1 是改良式分散式語音辨識系統前級之基頻參數抽取的作法的架構。

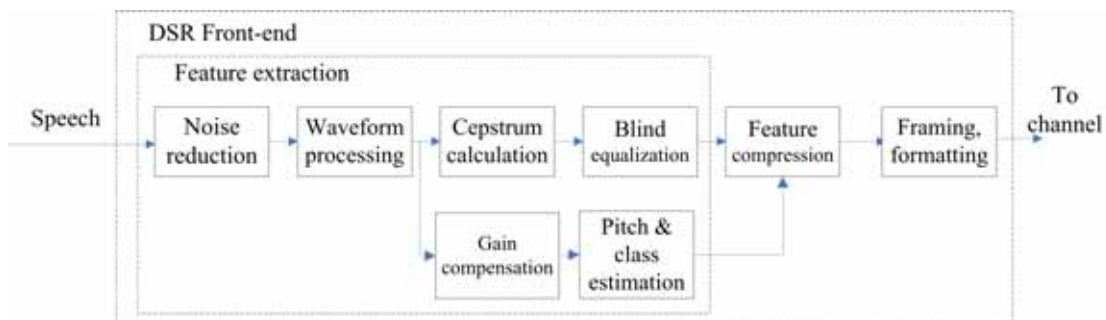


圖 4-1：改良式分散式語音辨識系統前級之基頻參數抽取的架構

其中加入信號補償方塊 (Gain compensation) 是因為分散式語音辨識系統中降低雜訊處理的維納濾波器，會減弱語音信號的能量，而原 XAFE 的基頻偵測機制中，會使用信號能量來做無聲音或有聲音的辨別。為了要補償能量的改變，所以對每個音框，將經過降低雜訊處理之語音信號的平均能量與輸入之語音信號的平均能量作比較，得到兩個平均能量的比值，將經過降低雜訊處理之語音信號的振幅乘以兩個平均能量的比值開跟號，如此即可補償能量的改變。所以在信號補償方塊中，信號振幅放大值  $g(i)$  如下式所示：

$$g(i) = \sqrt{\frac{P_{input}(i)}{P_{reduction}(i)}} \quad (4-1)$$

其中  $P_{input}(i)$  代表第  $i$  個音框輸入語音信號的平均能量， $P_{reduction}(i)$  代表經過降低雜訊處理之語音信號的平均能量。



## 4.2 改良式 DSR 之基頻參數抽取之效能分析

對改良式分散式語音辨識系統前級之基頻參數抽取器作效能分析，我們同樣利用 3-2 節所定義的三項錯誤分析來評估改良式分散式語音辨識系統前級之基頻參數抽取器的效能。



表 4-1：比較原本基頻參數抽取的作法與改進後基頻參數抽取的作法

訊噪比 (dB)	ETSI XAFE			Modified XAFE		
	U->V (%)	V->U (%)	V->V 相對錯誤	U->V (%)	V->U (%)	V->V 相對錯誤
20	3.64	8.71	0.095	2.68	8.74	0.097
15	3.95	13.19	0.146	2.78	13.20	0.144
10	4.19	22.03	0.249	3.15	21.41	0.240
5	4.72	41.73	0.464	3.98	39.05	0.433
0	4.51	69.65	0.736	4.21	63.96	0.679
-5	4.94	88.82	0.916	5.72	84.18	0.872

同樣也可以從表 4-1 觀察到：在 ETSI XAFE 中 V->V 的判別錯誤率也與 V->U 的錯誤率有相同的現象，隨著訊噪比下降而增加，尤其是在訊噪比在 10dB 以下，下降得更嚴重；改良後的基頻參數抽取方法在訊噪比 15dB 以後的相對錯誤，有比原本的基頻參數抽取方法下降的緩慢。所以我們僅對原本歐洲電信標準協會編號 202 212 V1.1.1 的分散式語音辨識系統前級的架構做小幅改良，即可大幅降低原本分散式語音辨識系統標準中基頻資訊所受環境雜訊的影響。

### 4.3 國語連續數字串之辨識---加入改良式分散式語音辨識系統抽取的基頻參數

經由上一節證實，本論文所提出的改良式分散式語音辨識系統前級之基頻參數抽取器，並沒有大幅更改原來分散式語音辨識系統前級的架構，由前一節之分析可以發現：改良式分散式語音辨識系統前級之基頻參數抽取器可以將基頻偵測的效能提高。接著將使用改良式分散式語音辨識系統前級之基頻參數抽取器應用於國語連續數字串之辨識器中。

### 4.3.1 實驗設定與訓練模型建立

本實驗用的語料庫都是國語連續數字串；實驗設定也都分為乾淨語料訓練以及複合情境訓練兩組模式。訓練模型也是有 12 個聲學模型，10 個國語數字的聲學模型，每個聲學模型設定為 8 個狀態，每個狀態含有 8 個混合高斯數；還有兩個聲學模型：靜音聲學模型 3 個狀態、停頓聲學模型 1 個狀態，每個狀態含有 16 個混合高斯數。



### 4.3.2 實驗結果

表 4-2(a)、表 4-2(b)分別是加入改良式分散式語音辨識系統抽取的基頻參數的國語連續數字串辨認實驗中的乾淨語音訓練模式以及複合情境訓練模式的辨識結果。

表 4-2(a)：加入改良式分散式語音辨識系統抽取之基頻參數的國語連續數字串辨認實驗中乾淨語音訓練模式之辨識結果

乾淨語音訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	98.4%				
20	95.3%	97.7%	97.7%	96.0%	96.7%
15	93.3%	93.9%	96.4%	93.6%	94.3%
10	86.9%	89.1%	92.2%	88.8%	89.3%
5	71.0%	75.2%	77.7%	75.4%	74.8%
0	44.1%	47.8%	51.9%	47.7%	47.9%
-5	15.0%	18.2%	17.8%	15.3%	16.6%
平均值(20dB~0dB)	78.1%	80.7%	83.2%	80.3%	80.6%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	98.4%				
20	92.4%	96.6%	93.8%	96.3%	94.8%
15	90.3%	95.5%	92.7%	94.2%	93.2%
10	80.2%	87.9%	88.3%	86.0%	85.6%
5	66.4%	78.2%	76.6%	73.8%	73.8%
0	43.2%	46.4%	50.9%	54.8%	48.8%
-5	20.3%	20.9%	18.9%	33.6%	23.4%
平均值(20dB~0dB)	74.5%	80.9%	80.5%	81.0%	79.2%
八種環境雜訊及五種訊噪比的平均值					79.9%

表 4-2(b)：加入改良式分散式語音辨識系統抽取之基頻參數的國語連續數字

串辨認實驗中複合情境訓練模式之辨識結果

複合情境訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	97.0%				
20	97.7%	98.1%	98.0%	97.7%	97.9%
15	96.4%	96.7%	98.0%	97.7%	97.2%
10	93.5%	94.2%	95.6%	93.8%	94.3%
5	81.2%	83.3%	86.0%	84.3%	83.7%
0	51.4%	59.4%	59.0%	61.5%	57.8%
-5	20.9%	21.0%	22.0%	21.0%	21.2%
平均值(20dB~0dB)	84.0%	86.3%	87.3%	87.0%	86.2%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	97.0%				
20	96.3%	97.7%	96.6%	97.0%	96.9%
15	94.7%	97.4%	95.6%	97.0%	96.2%
10	91.0%	92.7%	93.6%	93.8%	92.8%
5	79.3%	84.1%	82.9%	83.8%	82.5%
0	56.5%	56.1%	60.0%	64.5%	59.3%
-5	21.2%	27.7%	28.4%	39.3%	29.2%
平均值(20dB~0dB)	83.6%	85.6%	85.7%	87.2%	85.5%
八種環境雜訊及五種訊噪比的平均值					85.9%

比較沒有加入基頻參數、加入改良式分散式語音辨識系統抽取之基頻參數與加入分散式語音辨識系統抽取之基頻參數的實驗結果，將分別比較在乾淨語音訓練以及複合情境訓練模式，比較結果列於圖 4-2(a)、圖 4-2(b)。

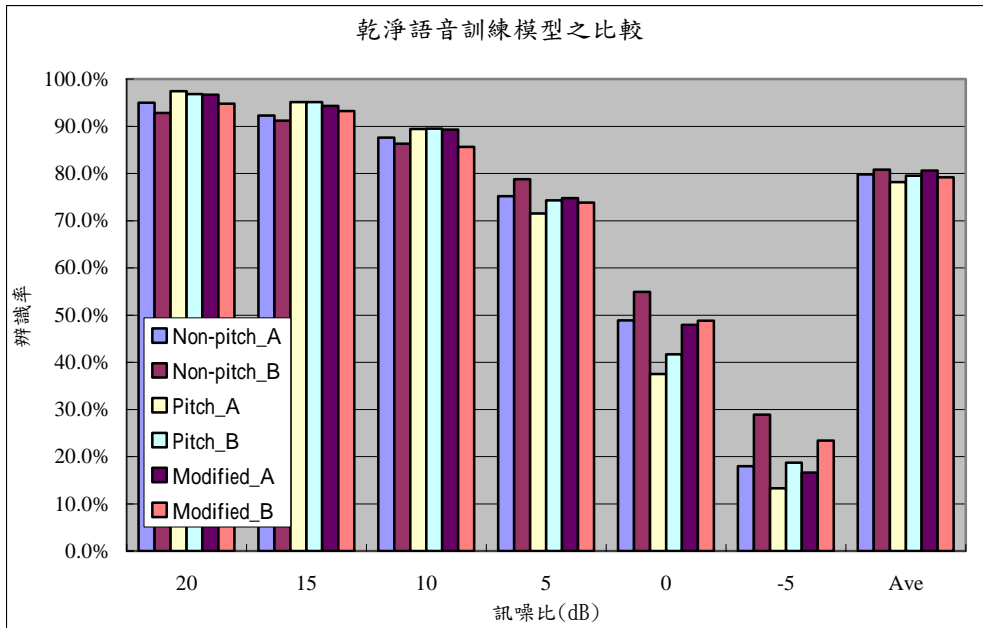


圖 4-2(a)：在乾淨語音訓練模式中比較沒有加入基頻參數、使用 DSR XAFE 與 Modified XAFE 抽取之基頻參數的辨識結果

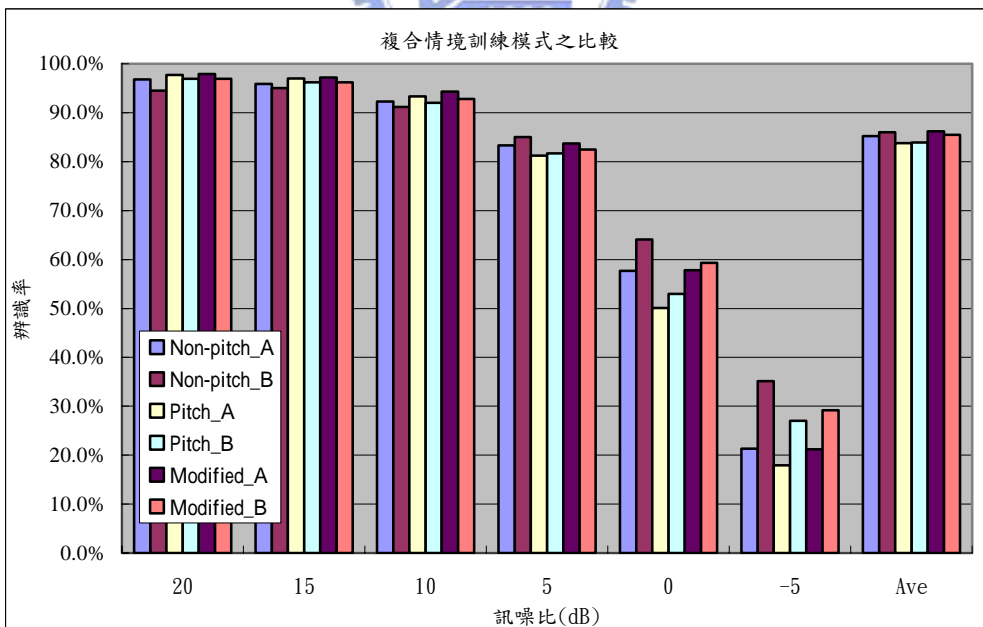


圖 4-2(b)：在複合情境訓練模式中比較使用沒有加入基頻參數、DSR XAFE 與 Modified XAFE 抽取之基頻參數的辨識結果

從實驗結果中，可以得到以下觀察：

- (1) 比較加入改良式分散式語音辨識系統抽取之基頻參數與加入原本分散式語音辨識系統抽取之基頻參數的實驗結果：在乾淨語料訓練模式中——測試組合 A 組在訊噪比 10dB 以上的辨識率幾乎都是加入改良式分散式語音辨識系統抽取之基頻參數比加入原本分散式語音辨識系統抽取之基頻參數還低，只有在汽車雜訊在訊噪比為 10dB 情形中，辨識率是加入改良式分散式語音辨識系統抽取之基頻參數比加入原本分散式語音辨識系統抽取之基頻參數還高；訊噪比在 5dB 以下的辨識率都是加入改良式分散式語音辨識系統抽取之基頻參數比加入原本分散式語音辨識系統抽取之基頻參數還高。測試組合 B 組中，不只在訊噪比 10dB 以上的實驗結果都是加入改良式分散式語音辨識系統抽取之基頻參數比加入原本分散式語音辨識系統抽取之基頻參數還低，還包括了在訊噪比為 5dB 時，在餐廳與火車站的環境雜訊中也是加入改良式分散式語音辨識系統抽取之基頻參數比加入原本分散式語音辨識系統抽取之基頻參數的辨識率還低；其餘的情況就都是加入改良式分散式語音辨識系統抽取之基頻參數比加入原本分散式語音辨識系統抽取之基頻參數的辨識率還高。在乾淨語音訓練模式情況中，整體的相對錯誤減少率 (Error reduction rate) 為 4.8%。

在複合情境訓練模式中——測試組合 A 組中，就只有汽車環境雜訊在訊噪比為 20dB 以及展覽會館在訊噪比為 15dB 時，辨識率是加入改良式分散式語音辨識系統抽取之基頻參數比加入原本分散式語音辨識系統抽取之基頻參數還低；其他的情形，包括乾淨語料測試，都是加入改良式分散式語音辨識系統抽取之基頻參數比加入原本分散式語音辨識系統抽取之基頻參數的辨識率還要高。在複合情境訓練模式情況中，整體的相對錯誤減少率為 12.4%。

綜合來說，加入改良式分散式語音辨識系統抽取之基頻參數的確是比加入原本分散式語音辨識系統抽取之基頻參數的辨識率高，相對錯誤減少率為 8.6%。

(2) 比較加入由改良式分散式語音辨識系統抽取之基頻參數與沒有加入基頻參數的實驗結果：乾淨語音訓練模式---測試組合 A 組在訊噪比 10dB 以上時都是加入由改良式分散式語音辨識系統抽取之基頻參數比沒有加入基頻參數的辨識率高；在訊噪比 5dB 以下時，則否，而 A 組的平均辨識率則是加入由改良式分散式語音辨識系統抽取之基頻參數比沒有加入基頻參數的實驗高。測試組合 B 組在訊噪比 15dB 以上時都是加入由改良式分散式語音辨識系統抽取之基頻參數比沒有加入基頻參數的辨識率高；在訊噪比 10dB 以下時，則否，而 B 組的平均辨識率則是加入由改良式分散式語音辨識系統抽取之基頻參數比沒有加入基頻參數的辨識率低。在乾淨語音訓練模式情況中，整體的辨識率是加入由改良式分散式語音辨識系統抽取之基頻參數比沒有加入基頻參數的辨識率低。在複合情境訓練模式情況中，整體的相對錯誤減少率為 1.6%。

複合情境訓練模式---測試組合 A 組除了在訊噪比-5dB 的情況，其餘的都是加入由改良式分散式語音辨識系統抽取之基頻參數比沒有加入基頻參數的辨識率高，而 A 組的平均辨識率則是加入由改良式分散式語音辨識系統抽取之基頻參數比沒有加入基頻參數的實驗高。測試組合 B 組在訊噪比 10dB 以上時都是加入由改良式分散式語音辨識系統抽取之基頻參數比沒有加入基頻參數的辨識率高；在訊噪比 5dB 以下時，則否，而 B 組的平均辨識率則是加入由改良式分散式語音辨識系統抽取之基頻參數比沒有加入基頻參數的實驗低。

上述結果是因為在高訊噪比時經過二階式維納濾波器的信號會改變信號波形進而破壞信號的週期特性；但在低訊噪比時，二階式維納濾波器移除較多的環境雜訊，而造成基頻偵測的改進。

綜合來說，加入改良式分散式語音辨識系統抽取之基頻參數的辨識結果還是比沒有加入基頻參數的辨識率低。

由上述兩點觀察結果，可以知道加入改良式的分散式語音辨識系統求取之基頻參數辨識器雖然其效益的確比加入原本的分散式語音辨識系統之基頻參數的辨識器要好，但是卻沒有比沒有加入基頻參數的辨識器好。因此我們便想到可以結合兩個辨識器的優點，成為一個新的辨識器，在訊噪比 10dB 以上的語音以加入改良式的分散式語音辨識系統求取之基頻參數辨識器為主；在訊噪比 5dB 以下，以沒有加入基頻參數的辨識器為主，這樣組合而成的辨識器應該會有較好的效能。

#### 4.4 國語連續數字串之辨識---整合沒有加入基頻參數的辨識器與加入改良式分散式語音辨識系統抽取之基頻參數的辨識器



因為比較過沒有加入基頻參數的辨識結果與加入改良式的分散式語音辨識系統求取之基頻參數的辨識結果，發現訊噪比在 5dB 以下時，加入改良式的分散式語音辨識系統求取之基頻參數反而會使辨識率下降。所以想要將沒有加入基頻參數的辨識器與有加入經過降低雜訊干擾處理的辨識器整合成一個新的辨識器，希望能夠提升所有訊噪比的辨識率。



#### 4.4.1 實驗設定

首先分別將兩個辨認器辨識出來分數 (Log-likelihood scores)，依據所辨認句子不同的訊噪比，給相對應的比重係數 (Weight)，再將乘上比重係數後的分數相加，最後取分數最高的為辨識結果。當訊噪比越高，越信任所求得的基頻參數，也就是越信任有加入基頻參數的辨識答案；反之，訊噪比越低，越不信任所求得的基頻參數，也就是越信任沒有加入基頻參數的辨識答案。若是沒有相同的答案，則是依照訊噪比，當訊噪比在 10dB 以上時，就採用加入改良式的分散式語音辨識系統求取之基頻參數辨識器的辨識答案；當訊噪比在 5dB 以下，就採用沒有加入基頻參數的辨識器的辨識答案。以下面的式子表示：

$$S' = \omega \cdot S_{with\_pitch} + (1 - \omega) \cdot S_{without\_pitch} \quad (4-2)$$

其中  $S'$  代表整合後的辨識分數； $S_{with\_pitch}$  代表有加入基頻參數的辨識分數； $S_{without\_pitch}$  代表沒有加入基頻參數的辨識分數。而  $\omega$  是比重係數，其比重函式

(Weighting function) 則是使用 Sigmoid 函數[7]，如下式：

$$\omega(d) = \frac{1}{1 + \exp(-\gamma d + \theta)} \quad , \quad \gamma=2.5, \theta=19 \quad (4-3)$$

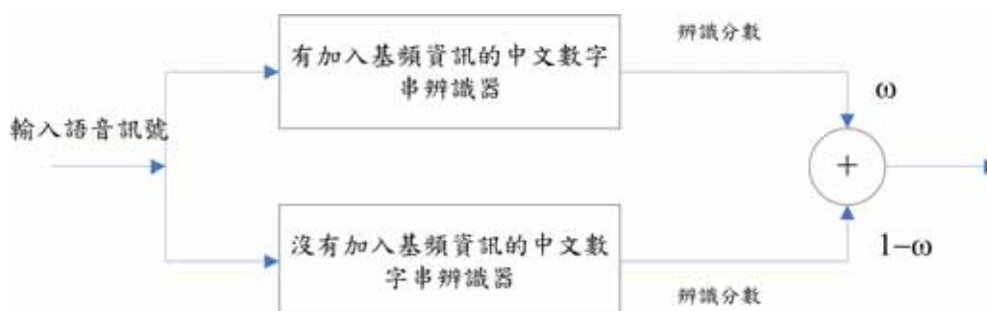


圖 4-3：整合含有基頻參數以及不含基頻參數的辨識器之系統方塊圖

#### 4.4.2 實驗結果

表 4-3(a)、表 4-3(b)分別是整合有由改良式的分散式語音辨識系統之基頻參數以及不含基頻參數的辨識實驗中的乾淨語音訓練模式以及複合情境訓練模式的辨識結果。

表 4-3(a)：整合有由改良式的分散式語音辨識系統抽取之基頻參數以及不含基頻參數的國語連續數字串辨識實驗中乾淨語音訓練模式之辨識結果

乾淨語音訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	98.4%				
20	96.0%	97.5%	97.8%	96.3%	96.9%
15	93.2%	94.7%	96.4%	93.6%	94.5%
10	87.7%	90.2%	93.8%	89.9%	90.4%
5	71.3%	78.7%	84.4%	73.8%	77.1%
0	43.6%	52.8%	60.1%	45.2%	50.4%
-5	15.1%	20.4%	19.0%	15.0%	17.4%
平均值(20dB~0dB)	78.4%	82.8%	86.5%	79.8%	81.9%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	98.4%				
20	92.5%	96.9%	93.5%	96.6%	94.9%
15	91.6%	95.5%	92.7%	95.0%	93.7%
10	83.0%	88.0%	89.6%	88.5%	87.3%
5	69.0%	83.0%	81.5%	83.5%	79.3%
0	50.8%	48.9%	57.0%	65.7%	55.6%
-5	22.1%	23.8%	29.6%	37.5%	28.3%
平均值(20dB~0dB)	77.4%	82.5%	82.9%	85.9%	82.2%
八種環境雜訊及五種訊噪比的平均值					82.1%

表 4-3(b)：整合有由改良式的分散式語音辨識系統抽取之基頻參數以及不含基頻參數的國語連續數字串辨識實驗中複合情境訓練模式之辨識結果

複合情境訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	97.0%				
20	97.4%	97.7%	98.3%	98.1%	97.9%
15	96.4%	96.4%	97.7%	96.9%	96.9%
10	93.3%	95.0%	96.1%	92.4%	94.2%
5	79.6%	85.4%	90.0%	82.1%	84.3%
0	48.6%	63.1%	67.3%	53.3%	58.1%
-5	17.8%	24.8%	24.5%	15.3%	20.6%
平均值(20dB~0dB)	83.1%	87.5%	89.9%	84.6%	86.3%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	97.0%				
20	96.1%	97.2%	96.0%	97.2%	96.6%
15	96.0%	97.4%	95.6%	96.7%	96.4%
10	91.7%	91.3%	94.4%	94.6%	93.0%
5	80.7%	86.1%	86.1%	90.7%	85.9%
0	58.7%	55.3%	67.5%	75.1%	64.2%
-5	24.9%	26.6%	38.6%	47.0%	34.3%
平均值(20dB~0dB)	84.6%	85.5%	87.9%	90.9%	87.2%
八種環境雜訊及五種訊噪比的平均值					86.8%

將此實驗結果與加入改良式的分散式語音辨識系統求取之基頻參數的實驗結果比較可觀察到在乾淨語音訓練模式中，所有訊噪比的情況整合兩個辨識器的實驗結果都比加入改良式的分散式語音辨識系統求取之基頻參數的實驗結果還要好，其錯誤減少率為 10.56%。在複合情境訓練模式中，整體的錯誤減少率為 6.22%。

## 4.5 國語連續數字串之辨識---使用乾淨語音的基頻參數之辨識器

這個實驗的目的是：假設當所求得的基頻參數沒有受到環境雜訊的影響時，即所求得的基頻參數可靠性相當高，可以將這個實驗結果當作是此方法的上限 (Upper bound)。

### 4.5.1 實驗設定

梅爾倒頻譜係數是從加了環境雜訊後的語音訊號抽取的，而基頻參數是從與加了環境雜訊後的語音訊號相對應的乾淨語料中抽取的。



### 4.5.2 實驗結果

表 4-4 是使用乾淨語音之基頻參數的辨識實驗中的乾淨語音訓練模式的辨識結果。從辨識結果中可以觀察到使用由乾淨語料中抽取的基頻參數之辨識器，其辨識結果的確是在本論文中五個國語連續數字串的實驗中最好的。因此我們可以將此實驗結果當作是加入基頻參數之辨識器的一個指標，也就是說如果我們所抽取的基頻參數可以完全不受到環境雜訊的影響，便有可能達到此實驗的效果。與表 4-2(a)比較，可以發現在低訊噪比時使用改良式分散式語音辨識系統求取之基頻參數的辨識結果，其辨識率還是大幅下降，所以可以知道在低訊噪比時，基頻偵測器還是有改善的空間。

表 4-4：使用乾淨語音之基頻參數的國語連續數字串辨識中乾淨語音訓練  
模式之辨識結果

乾淨語音訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	98.4%				
20	95.2%	96.4%	98.0%	95.5%	96.3%
15	91.6%	95.0%	96.1%	93.9%	94.2%
10	84.4%	90.3%	93.9%	87.9%	89.1%
5	71.2%	77.4%	89.1%	74.3%	78.0%
0	56.4%	59.7%	74.6%	53.9%	61.2%
-5	36.3%	36.8%	50.0%	24.6%	36.9%
平均值(20dB~0dB)	79.8%	83.8%	90.3%	81.1%	83.8%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	98.4%				
20	95.6%	96.0%	96.9%	97.2%	96.4%
15	94.7%	95.2%	94.1%	96.0%	95.0%
10	86.0%	87.7%	91.0%	94.4%	89.8%
5	75.6%	83.3%	85.2%	86.9%	82.8%
0	56.9%	62.9%	67.5%	75.6%	65.7%
-5	29.6%	42.8%	46.9%	59.7%	44.8%
平均值(20dB~0dB)	81.8%	85.0%	86.9%	90.0%	85.9%
八種環境雜訊及五種訊噪比的平均值					84.9%

## 第五章 大字彙國語連續語音辨認

前面都是討論在分散式語音辨識系統環境下，國語連續數字串的辨認，但對於大字彙國語連續語音的辨認系統呢？在國外的研究，在分散式語音辨識系統環境下，對 HUB4 測試語料之辨認率會下降至 20~30%。所以本章主要是要探討在分散式語音辨識系統環境下大字彙的國語連續語音辨識；藉以瞭解、分析在分散式語音辨識系統的架構中，大字彙國語語音辨識系統的效能並探討加入聲調資訊後對系統的改進程度。

### 5.1 語料庫介紹---TCC300

在本論文中所使用的國語大字彙語料庫是 TCC300。TCC300 是一套由台灣大學、成功大學，以及交通大學聯合錄製的麥克風語料，是一套大字彙連續國語語音的語料。表 5-1 列出此套語料的錄製方式，取樣頻率、訓練語料的句數、測試語料句數，以及語料統計特性。

表 5-1：大字彙連續國語語音的語料庫


錄製方式	麥克風
取樣頻率	16k Hz
編碼格式	16 位元 PCM
語料	男性語者 150 人和女性語者 150 人，共 27,337 句，332,267 個音節
統計特性	平均每句含有 12 個音節

與國語連續數字串一樣，為了符合一般大眾使用的 GSM 手機的取樣頻率—8K Hz 的標準，所以將 TCC300 語料降頻至 8K Hz。

## 5.2 大字彙國語連續語音之辨識---沒有加入基頻參數的辨識器

分散式語音辨識系統興起的主要原因是想要以語音輸入介面取代傳統式的鍵盤、按鍵輸入介面，因此我們絕對需要建立在分散式語音辨識系統環境下之大字彙連續語音辨認器。在本論文中接要介紹的就是在分散式語音辨識系統環境下支大字彙國語連續語音辨認。

### 5.2.1 實驗設定



實驗設定是使用如表 5-2 所示的設定。在實驗中採用隱馬可夫模型語音辨識器。隱馬可夫模型的產生也可以分成只用乾淨語料訓練，或是用加入不同的環境雜訊、以及不同訊噪比的語料做訓練，分別對應到「乾淨語料訓練」和「複合情境訓練」這兩種訓練模式；而且依照各種訊噪比加上 8 種不同的環境雜訊，按照所加環境雜訊的種類，分成 A、B 兩種測試組合，其中 A 組所加入的環境雜訊是與訓練語料所加入之環境雜訊匹配，B 組所加入的環境雜訊與是訓練語料所加入之環境雜訊不匹配。

其中在乾淨語音訓練模式中，將語料庫的十分之九當訓練語料，共 24,742 句，300,856 個音節；在複合情境訓練模式中，將乾淨語音訓練模式的 24,742 句，平均分為 20 組，每組 1,237 句，或 1,238 句，而每組分別是加入不同環境雜訊、不同訊噪比的情境。在兩種訓練模式中，都是語料庫的另外十分之一當測試語料，共 2,595 句，31,411 個音節。因為測試語料太多所以我們將 2,595 句語料分為 8 組，每組 324 句，或 325 句，這 8 組分別加入不同的環境雜訊，然後

在每一組中所有的語料都是有 7 種不同的訊噪比 (20dB、15dB、10dB、5dB、0dB、-5dB 以及完全乾淨)，但是在這 8 組中每組的音節數不盡相同，將於表 5-3 中詳細列出 8 組中所包含的音節數。

表 5-2：加上環境雜訊的大字彙連續國語語音內容介紹

大字彙國語連續語音		
取樣頻率	8 kHz	
訓練模式	乾淨語音訓練	複合情境訓練
	音段數：24,742 環境雜訊： 無	音段數：24,742 環境雜訊： ● 種類：地下鐵、人聲、汽車、展覽會館 ● 訊噪比：20dB、15dB、10dB、5dB 和完全乾淨 ● 4 種雜訊乘以 5 種 SNR，共 20 種情境
測試組合	A 組	B 組
	音段數：9,086 環境雜訊： 地下鐵 人聲 汽車 展覽會館	音段數：9,079 環境雜訊： 餐廳 街道 機場 火車站
	對於上述的每種環境雜訊，訊噪比都控制在 20dB、15dB、10dB、5dB、0dB、-5dB 以及完全乾淨七種程度，並且對於每種雜訊的每依個訊噪比程度都計算一組辨識結果	

表 5-3：在測試語料中在 8 種環境雜訊下之音節數

環境雜訊	地下鐵	人聲	汽車	展覽會館	餐廳	街道	機場	火車站
音節數	3908	3986	3916	4034	3816	3931	3836	3984



沒有加入基頻資訊的大字彙國語語音辨認實驗所使用的辨認模型，就是使用國語的 411 音節的次音節模型 (Syllable) 所建立的，也就是 100 個韻母相關之聲母模型、40 個韻母模型；韻母相關之聲母模型使用 3 個狀態，韻母模型使用 5 個狀態。其中每個狀態都是 64 個混合高斯數；還有兩個模型——靜音與停頓的聲學模型，其中靜音聲學模型設定為 3 個狀態，停頓聲學模型設定為 1 個狀態，此狀態允許跳躍，並且與靜音模型的中間狀態合併，兩個聲學模型中每個狀態含有 128 個混合高斯數。



## 5.2.2 實驗結果

表 5-4(a)、表 5-4(b) 分別是沒有加入基頻參數的大字彙國語連續語音辨識實驗中的乾淨語音訓練模式以及複合情境訓練模式的辨識結果。

表 5-4(a)：沒有加入基頻參數的大字彙國語連續語音辨識實驗中的乾淨語音訓練模式之辨識結果

乾淨語音訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	61.2%	63.1%	62.2%	62.0%	62.1%
20	49.1%	56.2%	57.4%	49.9%	53.2%
15	39.0%	48.0%	51.6%	40.4%	44.8%
10	28.1%	36.1%	42.1%	27.3%	33.4%
5	15.2%	21.0%	28.3%	15.1%	19.9%
0	7.5%	8.6%	12.7%	6.5%	8.8%
-5	2.2%	2.8%	3.8%	2.2%	2.8%
平均值(20dB~0dB)	27.8%	34.0%	38.4%	27.8%	32.0%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	61.8%	63.2%	61.1%	61.9%	62.0%
20	51.2%	52.8%	55.6%	57.6%	54.3%
15	41.8%	47.0%	49.2%	51.4%	47.4%
10	30.3%	35.7%	41.8%	43.1%	37.7%
5	16.9%	22.2%	26.1%	31.3%	24.1%
0	6.2%	12.0%	13.6%	18.4%	12.6%
-5	1.1%	4.0%	4.7%	5.9%	3.9%
平均值(20dB~0dB)	29.3%	33.9%	37.3%	40.4%	35.2%
八種環境雜訊及五種訊噪比的平均值					33.6%

表 5-4(b)：沒有加入基頻參數的大字彙國語連續語音辨識實驗中的複合情境訓練模式之辨識結果

複合情境訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	54.8%	56.1%	55.8%	57.9%	56.2%
20	43.2%	49.5%	51.1%	45.1%	47.2%
15	34.0%	42.2%	46.6%	37.2%	40.0%
10	25.5%	32.2%	37.8%	26.2%	30.4%
5	14.5%	18.0%	25.4%	14.8%	18.2%
0	7.0%	7.4%	12.6%	6.0%	8.3%
-5	2.0%	2.5%	3.9%	1.9%	2.6%
平均值(20dB~0dB)	24.8%	29.9%	34.7%	25.9%	28.8%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	54.9%	56.0%	54.9%	55.6%	55.4%
20	45.2%	46.9%	49.1%	51.6%	48.2%
15	36.9%	41.9%	43.9%	45.7%	42.1%
10	25.5%	31.7%	37.0%	38.7%	33.2%
5	13.5%	19.7%	23.5%	28.4%	21.3%
0	4.8%	10.8%	12.2%	16.7%	11.1%
-5	0.9%	3.2%	4.4%	4.9%	3.4%
平均值(20dB~0dB)	25.2%	30.2%	33.1%	36.2%	31.2%
八種環境雜訊及五種訊噪比的平均值					30.0%

由上面實驗可以發現雖然 DSR 前級已經使用二階式維納濾波器來濾除雜訊，但在語音與環境雜訊比越小，辨認率仍會下降。在訊噪比小於 10dB 時辨認率下降的十分之快，可以發現在大字彙國語語音辨識系統中，如果能使用匹配的訓練及測試語料當然可以提升辨識效能，但是使用多種環境雜訊及不同的訊噪比的語音來訓練辨認模型並不能夠改善系統對環境雜訊的抵抗能力。

## 5.3 大字彙國語連續語音之辨識---加入改良式的分散式語音辨識系統抽取之基頻參數的辨識器

在本論文中前面已經得到證實：使用基頻參數之辨識器其效能會比沒有使用基頻參數之辨識器的效能還要好，且使用本論文提出之改良式分散式語音辨識系統抽取的基頻參數效能會比使用原本分散式語音辨識之標準系統抽取之基頻參數更好，所以本節將介紹使用改良式的分散式語音辨識系統抽取之基頻參數的大字彙國語連續語音之辨識。

### 5.3.1 實驗設定

加入基頻資訊的實驗所使用的訓練模型：參考了新辭典、三民主義、國語日報的字典，以及微軟新注音輸入法，來統計每個音節聲調出現的情況；並且考慮到在讀的時候，會有變調的情形發生，也就是會有些三聲的音節會讀成二聲的音節，所以從上一步驟統計出的聲調出現的情況，將只有出現三聲的 411 音節也加入二聲。因此音節個數由 411 增加至 1,515 個，在表 5-5 中顯示在 1,515 音節中各聲調含有的音節數，聲調的 Perplexity 為 4.44；因為 Perplexity 小於 5，所以加入基頻參數將有助於 411 音節之辨認率。又因為韻母模型跟聲調的相關性較強，韻母相關之聲母模型則否，所以不增加韻母相關之聲母模型個數，還是 100 個聲母模型，韻母模型個數由 40 個增加至 177 個與聲調相關之韻母模型。其中韻母相關之聲母模型是 3 個狀態，韻母模型是 5 個狀態，其中每個狀態都是 64 個混合高斯數；還有兩個聲學模型：靜音模型與停頓模型的聲學模型，其中靜音聲學模型設定為 3 個狀態，停頓聲學模型設定為 1 個狀態，此狀態允許跳躍，並且與靜音模型的中間狀態合併，兩個聲學模型中每個狀態含有 128 個混合高斯數。

表 5-5：每個聲調的出現次數

聲調	出現次數
1	350
2	383
3	351
4	372
5	59

### 5.3.2 實驗結果

本實驗的實驗結果分為三組：1,515 音節之辨識率、聲調之辨識率以及 411 音節之辨識率。其中 1,515 音節之辨識率就是將辨識出來的答案與測試語料的 1,515 音節之標準答案作動態規劃 (Dynamic programming) 後，所得到的最佳辨識結果；聲調之辨識率是將所辨識出來的答案與測試語料的 1,515 音節之標準答案作動態規劃後，保留其與標準答案的相對應關係，去除音節只留下聲調資訊，在去計算辨識率；411 音節之辨識率是先將由 1,515 音節之聲學模式辨識出來的答案，去除聲調的資訊只留下 411 音節，再將此經過處理的辨識答案與測試語料的 411 音節之標準答案作動態規劃後，所得到的最佳辨識率。

表 5-6(a)、5-6(b)分別是加入改良式的分散式語音辨識系統抽取之基頻參數的大字彙國語連續語音辨識實驗中，乾淨語音訓練模式以及複合情境訓練模式中，1,515 個音節之辨識結果。由實驗結果可以發現在加入基頻資訊後，複合情境訓練模式的辨認率會優於乾淨語音訓練模式。

表 5-6(a)：加入改良式的分散式語音辨識系統抽取之基頻參數的大字彙國語連

續語音辨識實驗中的乾淨語音訓練模式 1,515 個音節之辨識結果

乾淨語音訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	44.0%	46.1%	43.8%	45.9%	45.0%
20	35.7%	42.2%	39.8%	37.7%	38.9%
15	28.3%	35.1%	34.3%	31.1%	32.2%
10	20.3%	27.6%	26.6%	20.5%	23.8%
5	10.8%	13.9%	14.7%	10.6%	12.5%
0	3.8%	4.5%	4.5%	3.3%	4.0%
-5	0.4%	0.5%	0.8%	0.4%	0.5%
平均值(20dB~0dB)	19.8%	24.7%	24.0%	20.6%	22.3%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	43.3%	44.8%	45.0%	44.2%	44.3%
20	37.5%	37.8%	41.5%	41.0%	39.5%
15	31.6%	32.7%	36.2%	35.3%	34.0%
10	22.3%	24.8%	28.6%	27.0%	25.7%
5	11.3%	13.4%	15.4%	16.0%	14.0%
0	3.6%	5.7%	5.0%	8.2%	5.6%
-5	0.7%	1.2%	1.1%	1.9%	1.2%
平均值(20dB~0dB)	21.3%	22.9%	25.3%	25.5%	23.8%
八種環境雜訊及五種訊噪比的平均值					23.1%

表 5-6(b)：加入改良式的分散式語音辨識系統抽取之基頻參數的大字彙國語連

續語音辨識實驗中的複合情境訓練模式 1,515 個音節之辨識結果

複合情境訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	40.8%	41.1%	40.1%	42.0%	41.0%
20	36.5%	39.7%	38.7%	37.8%	38.2%
15	31.0%	36.9%	35.8%	34.0%	34.4%
10	24.6%	30.7%	30.8%	27.3%	28.4%
5	15.4%	18.9%	18.2%	16.1%	17.2%
0	6.6%	6.9%	8.5%	6.6%	7.2%
-5	1.4%	2.1%	1.7%	1.3%	1.6%
平均值(20dB~0dB)	22.8%	26.6%	26.4%	24.4%	25.1%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	40.6%	41.0%	41.5%	41.5%	41.2%
20	35.7%	37.2%	40.3%	40.8%	38.5%
15	33.1%	33.5%	36.6%	35.7%	34.7%
10	24.9%	28.4%	30.4%	28.1%	28.0%
5	14.8%	17.3%	19.7%	19.2%	17.8%
0	5.4%	8.5%	8.1%	9.4%	7.9%
-5	1.0%	2.5%	2.3%	2.7%	2.1%
平均值(20dB~0dB)	22.8%	25.0%	27.0%	26.6%	25.4%
八種環境雜訊及五種訊噪比的平均值					25.3%

接著我們只考慮國語聲調辨識率。由表 5-7(a)、5-7(b)分別是加入改良式的分散式語音辨識系統抽取之基頻參數的大字彙國語連續語音辨識實驗中，乾淨語音訓練模式以及複合情境訓練模式的聲調辨識結果。而其結果可以發現當訊噪比由無窮大（乾淨語料）掉到 5dB 時聲調辨認率會下降至原來的一半。

表 5-7(a)：加入改良式的分散式語音辨識系統抽取之基頻參數的大字彙國語連續語音辨識實驗中的乾淨語音訓練模式之聲調辨識結果

乾淨語音訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	63.3%	64.9%	63.5%	63.7%	63.8%
20	60.5%	63.5%	61.3%	61.2%	61.6%
15	56.3%	59.7%	58.2%	59.0%	58.3%
10	49.4%	56.9%	54.6%	50.7%	60.4%
5	34.9%	37.2%	36.7%	38.9%	36.9%
0	16.5%	17.7%	19.5%	20.4%	18.5%
-5	4.3%	4.7%	5.6%	4.1%	4.7%
平均值(20dB~0dB)	43.6%	47.0%	52.1%	46.0%	47.2%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	64.6%	64.0%	64.5%	63.1%	64.1%
20	62.1%	61.2%	62.4%	61.7%	61.8%
15	58.8%	58.7%	60.0%	58.1%	58.9%
10	51.2%	52.1%	53.0%	49.7%	51.5%
5	38.9%	38.0%	38.8%	38.2%	38.5%
0	21.7%	20.8%	20.3%	24.5%	21.8%
-5	9.6%	8.9%	9.7%	12.3%	10.1%
平均值(20dB~0dB)	46.5%	46.2%	46.9%	46.4%	46.5%
八種環境雜訊及五種訊噪比的平均值					46.8%



表 5-7(b)：加入改良式的分散式語音辨識系統抽取之基頻參數的大字彙國語連

續語音辨識實驗中的複合情境訓練模式之聲調辨識結果

複合情境訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	61.5%	63.3%	61.2%	61.5%	61.9%
20	60.2%	61.7%	61.5%	61.3%	61.2%
15	58.2%	61.1%	58.4%	58.5%	59.1%
10	54.6%	59.1%	86.5%	55.9%	64.0%
5	43.0%	43.1%	41.0%	45.7%	43.2%
0	22.2%	24.8%	26.4%	27.1%	25.1%
-5	9.3%	11.3%	9.9%	10.3%	10.2%
平均值(20dB~0dB)	47.6%	50.0%	54.8%	49.7%	50.5%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	63.0%	62.5%	63.5%	63.0%	63.0%
20	60.8%	61.3%	62.5%	62.1%	61.7%
15	58.4%	59.7%	60.0%	59.1%	59.3%
10	52.8%	54.7%	55.3%	51.6%	53.6%
5	42.8%	41.8%	43.6%	42.0%	42.5%
0	26.9%	25.7%	27.1%	28.8%	27.1%
-5	15.7%	14.0%	15.6%	17.3%	15.6%
平均值(20dB~0dB)	48.4%	48.7%	49.7%	48.7%	48.9%
八種環境雜訊及五種訊噪比的平均值					49.7%

接著，我們分析不考慮聲調的 411 音節辨認率。由分析表 5-8(a)、5-8(b) 分別是加入改良式的分散式語音辨識系統抽取之基頻參數的大字彙國語連續語音辨識實驗中，乾淨語音訓練模式以及複合情境訓練模式中，411 個音節之辨識結果，而結果可以發現當訊噪比由無窮大掉到 5dB 時音節辨認率會下降至原來的三分之一。

表 5-8(a)：加入改良式的分散式語音辨識系統抽取之基頻參數的大字彙國語連續語音辨識實驗中的乾淨語音訓練模式 411 個音節之辨識結果

乾淨語音訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	62.6%	65.2%	63.1%	64.7%	63.9%
20	51.7%	59.4%	58.4%	53.0%	55.6%
15	41.4%	51.1%	51.8%	43.9%	47.1%
10	30.8%	40.5%	45.0%	30.1%	36.6%
5	17.2%	23.7%	26.9%	16.6%	21.1%
0	6.4%	9.0%	10.1%	6.6%	8.0%
-5	1.7%	1.6%	2.8%	1.2%	1.8%
平均值(20dB~0dB)	29.5%	36.7%	38.4%	30.0%	33.7%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	62.9%	64.7%	62.8%	63.5%	63.5%
20	54.7%	55.6%	58.8%	59.6%	57.2%
15	46.9%	48.5%	52.3%	52.6%	50.1%
10	35.0%	37.7%	43.3%	43.7%	39.9%
5	19.8%	22.2%	26.3%	29.1%	24.4%
0	7.9%	10.6%	12.5%	17.2%	12.1%
-5	1.9%	3.0%	3.8%	5.4%	3.5%
平均值(20dB~0dB)	32.9%	34.9%	38.6%	40.4%	36.7%
八種環境雜訊及五種訊噪比的平均值					35.2%

表 5-8(b)：加入改良式的分散式語音辨識系統抽取之基頻參數的大字彙國語連

續語音辨識實驗中的複合情境訓練模式 411 個音節之辨識結果

複合情境訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	59.0%	59.8%	58.3%	60.1%	59.3%
20	52.6%	57.9%	56.1%	53.5%	55.0%
15	45.7%	53.2%	53.5%	48.7%	50.3%
10	37.6%	46.0%	46.9%	39.0%	42.4%
5	24.3%	31.1%	31.6%	24.6%	27.9%
0	11.5%	13.4%	15.4%	11.7%	13.0%
-5	3.7%	4.3%	5.2%	3.3%	4.1%
平均值(20dB~0dB)	34.3%	40.3%	40.7%	35.5%	37.7%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	59.1%	59.7%	58.7%	59.2%	59.2%
20	53.4%	54.7%	57.0%	58.0%	55.8%
15	49.0%	51.0%	52.7%	53.3%	51.5%
10	38.3%	42.9%	46.7%	45.7%	43.4%
5	24.4%	28.5%	32.3%	34.5%	29.9%
0	10.3%	14.6%	16.3%	19.2%	15.1%
-5	3.0%	5.4%	6.2%	6.7%	5.3%
平均值(20dB~0dB)	35.1%	38.3%	41.0%	42.1%	39.1%
八種環境雜訊及五種訊噪比的平均值					38.4%

由此實驗結果可以發現：

- (1). 比較沒有加入基頻參數的大字彙國語語音辨識結果與有加入基頻參數的大字彙國語語音辨識結果，可以看出基頻參數對環境雜訊的抵抗能力會較頻譜參數好，所以在複合情境訓練模式與乾淨語音訓練模式下都可以得到較佳的辨識結果。
- (2). 在 1,515 音節之辨識率、聲調之辨識率以及 411 音節之辨識率三組辨識結果

中，複合情境訓練模式的辨識結果都比乾淨語音訓練模式的辨識結果還要好。其相對錯誤減少率分別為：2.86%、5.45%與 4.94%，從這裡也可以看出加入基頻參數後的確對辨識率之提升有很好的效能。

#### 5.4 大字彙國語連續語音之辨識---整合沒有加入基頻參數的辨識器與加入改良式分散式語音辨識系統抽取之基頻參數的辨識器

在國語連續數字串中做了整合沒有加入基頻參數的辨識器與加入改良式分散式語音辨識系統抽取之基頻參數的辨識器，發現其效能會比沒有加入基頻參數的辨識器與加入改良式分散式語音辨識系統抽取之基頻參數的辨識器都還要好一點，所以在大字彙國語連續。

##### 5.4.1 實驗設定

將沒有加入基頻參數的辨識器與有加入經過降低雜訊干擾處理的辨識器整合成一個新的辨識器，希望能夠提升所有訊噪比的辨識率。依照訊噪比判斷那個辨識器的答案比較具有可靠性：當訊噪比在 10dB 以上時，就選擇加入改良式的分散式語音辨識系統求取之基頻參數辨識器的辨識答案；當訊噪比在 5dB 以下，就選擇沒有加入基頻參數的辨識器的辨識答案。

## 5.4.2 實驗結果

表 5-9 是整合有由改良式的分散式語音辨識系統抽取之基頻參數以及不含基頻參數的大字彙國語連續語音辨識實驗中乾淨語音訓練模式之辨識結果。

表 5-9：整合有由改良式的分散式語音辨識系統抽取之基頻參數以及不含基頻參數的大字彙國語連續語音辨識實驗中乾淨語音訓練模式之辨識結果

乾淨語音訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	62.8%	65.2%	63.1%	64.8%	64.0%
20	51.9%	59.5%	58.5%	53.2%	55.8%
15	41.8%	51.3%	52.0%	44.6%	47.4%
10	31.1%	40.6%	45.2%	30.3%	36.8%
5	16.4%	22.4%	28.6%	15.4%	20.7%
0	7.5%	8.7%	13.0%	6.6%	9.0%
-5	2.2%	2.8%	4.4%	2.8%	3.1%
平均值(20dB~0dB)	29.7%	36.5%	39.5%	30.0%	33.9%
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	63.0%	64.9%	63.1%	63.5%	63.6%
20	54.9%	55.7%	59.3%	60.0%	57.5%
15	47.1%	49.0%	52.5%	52.8%	50.4%
10	35.4%	37.8%	43.3%	44.2%	40.2%
5	18.0%	22.3%	28.6%	31.7%	25.2%
0	7.5%	11.5%	13.8%	18.6%	12.9%
-5	2.3%	4.1%	5.0%	6.3%	4.4%
平均值(20dB~0dB)	32.6%	35.3%	39.5%	41.5%	37.2%
八種環境雜訊及五種訊噪比的平均值					35.6%

比較表 5-4(a)、表 5-8(a)與表 5-9，可以發現當訊噪比在 10dB 以上時，有加入基頻參數之辨識實驗結果比沒有加入基頻參數之辨識實驗結果還好，而整合兩者後的辨識器效能是最好的；在訊噪比 5dB 以下時，因為並不全都是沒有加入基頻參數之辨識實驗結果比有加入基頻參數之辨識實驗結果還好，反而在 5dB 時只有在汽車、街道以及火車站三種環境雜訊下，沒有加入基頻參數之辨識實驗結果才會比有加入基頻參數之辨識實驗結果還好，並且連在 0dB 時都有三組（在人聲、展覽會館及餐廳的環境雜訊下）是有加入基頻參數之辨識實驗結果比沒有加入基頻參數之辨識實驗結果還好。這導致了在上述的八組情況（有加入基頻參數之辨識實驗結果比沒有加入基頻參數之辨識實驗結果還好）中，有六組在整合兩個辨識器後的新辨識器都沒有比有使用基頻參數之辨識器好，但是其他情況下的辨識率都有提升了，因此整體的辨識率還是有小幅提升。



## 5.5 加入語言模型至使用改良式分散式語音辨識系統抽取之基頻參數的大字彙國語連續語音辨識

在語音辨識作法中，聲學模型的比對是屬於較低層次的作法，因為其未包含任何有關的語言資訊，所以，一般而言，一部較佳的國語辨識器要能接受一連串的聲音訊號輸入，並輸出較為合理的口語句子。實際上的作法，就是根據一部有限詞彙的詞典去猜測聲音訊號，比較可能是什麼詞彙，最後就輸出有可能出現的句子。

上述所說的猜測的動作，即是依據語言的統計模型；所謂的語言模型一具有其獨特的文法規則，及語言特性，所求得一個機率模型，簡稱 LM (Language Model)，在辨識時，除了聲學模型外，若能加入語言模型的參考，通常能大幅提

高辨識系統的辨識率。

在本章節將建立一個內容較為廣泛及文字語料較多的，所訓練而得的語言模型——通用語言模型。

### 5.5.1 建立語言模型

本節將介紹我們是如何訓練語言模型，其流程如圖 5-1，

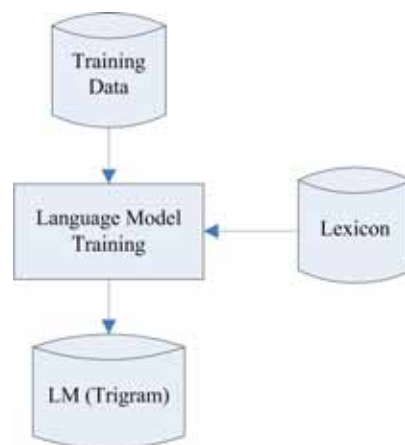


圖 5-1：LM 訓練流程圖

#### 5.5.1.1 訓練語料及詞典(lexicon)

建立語言模型必須要準備的兩樣資料—訓練語料及詞典，下面將介紹其用途，及本論文中所使用之訓練語料、詞典為何。

建立語言模型必須要有大量的文字資料庫，才可分析其語言規則，對於不同種類的訓練語料所分析出的語言規則也必定不同，本論文中所採用的訓練語料有

為：包含光華雜誌 (Sinorama)、NTCIR 和中研院的平衡語料庫，下面將稱之為通用語料庫。其中光華雜誌內容為一般雜誌文章，總共蒐集了 1976 年至 2000 年的資料；而 NTCIR( NACSIS Test Collections for IR )是一個建立檢索系統的標竿測試集，內容包含數種不同的科學領域；平衡語料庫是由中研院所錄製的，內容包含多種主題，目的在於研究語言分析，這三種語料庫的內容皆是文字性質，我們可藉此訓練出具有文字性質語言規則的語言模型。

有了語料庫我們即可做其語言上的分析，在漢語中文( Mandarin )下，以詞為單元來做分析是較符合語言規則的，所以必須將語料庫由原本以音節為單位轉換成以詞為單位，這時便需要詞典來做轉換，下面將對於本論文所使用之詞典其來源做介紹。

詞典的來源，是由交大電信所語音實驗室的詞典和台灣師大資工所做聯集以及後處理動作而得到的新詞典，此即為本論文中所使用之詞典，對於詞典中詞長分佈統計於表 5-10。

表 5-10：詞典中之詞長分佈

詞長	1	2	3	4	5	6	7	8	總合
數量	9,821	34,188	9,452	5,912	231	128	33	22	59,787
百分比	16.43	57.18	15.81	9.89	0.39	0.21	0.06	0.04	

根據論文中所使用的詞典，對語料庫作斷詞後的結果，其資料統計於



表 5-11。

表 5-11：通用語料庫之詞數表

訓練語料	詞數 (Word)	字數(Character)
光華雜誌	9, 870, 430	16, 406, 485
NTCIR	124, 442, 861	206, 847, 107
平衡語料庫	4, 796, 163	7, 972, 113
合計	139, 109, 455	231, 225, 705

### 5.5.1.2 訓練語言模型的方法

藉由訓練語料與詞典，本論文中我們是要訓練出 Trigram 的語言模型，因此要求出 Unigram、Bigram、Trigram 的機率，分別為  $P(w_i | w_{i-1})$ 、 $P(w_i | w_{i-1}, w_{i-2})$  及  $P(w_i | w_{i-1}, w_{i-2}, w_{i-3})$ ，下面將介紹，求取 n-gram 機率的方法，假設有一個詞串 (Word sequence) 或句子 (Sentence)，其內容以詞 (Word) 為單位為「 $w_1, w_2, \dots, w_m$ 」，則此詞串對應的機率為：

$$\begin{aligned}
 P(w_1, w_2, \dots, w_m) &= P(w_1)P(w_2 | w_1)P(w_3 | w_1, w_2) \cdots P(w_n | w_1, w_2, \dots, w_{n-1}) \\
 &= \prod_{i=1}^m P(w_i | w_1, \dots, w_{i-1})
 \end{aligned}
 \tag{式 5-1}$$

由於要求得所有詞的條件機率是不可能的，所以我們可以使用 n-gram 的機率去趨近所有詞的條件機率。

$$P(w_1, w_2, \dots, w_m) = \prod_{i=1}^m P(w_i | w_{i-n+1}, \dots, w_{i-1})
 \tag{式 5-2}$$

其中每個 n-gram 的機率如下式所示：

$$P(w_i | w_{i-n+1}, \dots, w_{i-1}) = \frac{\text{Count}(w_{i-n+1}, \dots, w_i)}{\text{Count}(w_{i-n+1}, \dots, w_{i-1})}
 \tag{式 5-3}$$

其中， $Count(.)$  表示為詞串出現的次數。在求得所有詞串  $n$ -gram 的機率後，我們即可得到所需求的語言模型了。

### 5.5.2 基本辨識器加入語言模型之辨識分析

要將語言模型加入辨識系統中，我們還需將之轉換為 Word-net，因為 Word-net 才是清楚的描述詞跟詞的轉移關係，由於 HTK[8] 中轉換上的問題，我們只使用到 Bigram 和 Unigram 的機率，其轉換流程如圖 5-2。

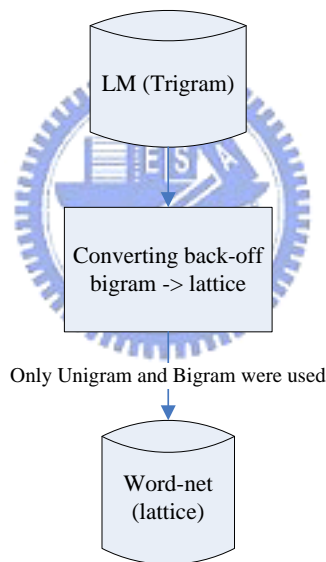


圖 5-2：LM 轉 Word-Net 之流程圖

有了 Word-net，相當於文法規則，之後便可將此文法加入基本的辨識系統中，而加入了語言模型，在辨識時我們除了會得到聲學模型的分數外，還會再得到語言模型的分數。實驗中我們的基本辨識系統為 5.3.1 節所介紹過的 1,515 音節中的聲母與韻母的聲學模型、靜音模型與停頓模型。

### 5.5.3 實驗結果

表 5-12 列出在複合情境訓練環境下，測試語料是加入了人聲環境雜訊在各種訊噪比之下的辨識率。可以看出加入了語言模型之後，音節的辨識率的確已經提高了許多；但是在低訊噪比時，還是受環境雜訊影響的想當嚴重。

表 5-12 在複合情境訓練、測試—人聲環境雜訊

訊噪比	乾淨語料	20	15	10	5	0	-5	平均值
Word	72.4	67.5	65.7	60.9	33.3	4.6	1.2	46.4
Syllable	87.3	90.2	87.7	78.7	54.0	16.0	10.0	65.3



## 第六章 結論與展望


在本論文的最後一章，我們將把本論文的貢獻做一次更加完整的說明；並且檢討本論文的不足，展望未來，提出可以加以補強以及延伸的研究方向。

### 6.1 結論

本論文的主要研究方向是：在分散式語音辨識系統環境下，藉由基頻資訊來對抗環境雜訊的干擾。在第二章中，做了沒有使用基頻資訊的國語連續數字串辨識實驗，發現了隨著訊噪比越低，辨識率也會隨著下降；尤其是訊噪比在 5dB 以下時，辨識率更是呈現大幅度的下降。而我們又發現在歐洲電信標準協會編號 202 212 V1.1.1 之分散式語音辨識系統的標準中，有現成的基頻偵測器，於是在第三章就是針對其基頻偵測器所抽取的基頻參數作分析，結果顯示出語音信號的週期特性非常容易受到環境雜訊的破壞，也就是說訊噪比越低，基頻受到環境雜訊的影響會越大，所求得的基頻也會越來越不可靠。在八種環境雜訊中，又以地下鐵、汽車、展覽會場受到的影響最大。接著做了使用分散式語音辨識系統抽取之基頻參數的國語連續數字串辨識實驗，並與沒有加入基頻參數的國語連續數字串實驗做比較，發現在訊噪比越低時，基頻參數受到環境雜訊影響越大，與之前的分析結果一致。

為了要改善在低訊噪比時，語音的週期特性非常容易受到環境雜訊影響的特性，於是想到可以利用歐洲電信標準協會編號 202 212 V1.1.1 之分散式語音辨識系統的前級標準中的降低雜訊處理——二階式維納濾波器。因為在原本歐洲電信標準協會編號 202 212 V1.1.1 之分散式語音辨識系統標準中基頻偵測器在偵

測基頻前，語音訊號並沒有經過二階式維納濾波器過濾雜訊之處理，所以我們就修改了分散式語音辨識系統標準中的前級抽取基頻參數的架構，使語音訊號在被抽取基頻參數前，先經過二階式維納濾波器做降低雜訊之處理，此方法我們稱為——改良式的分散式語音辨識系統之基頻參數抽取。之後也做了原來分散式語音辨識系統抽取之基頻參數與改良式的分散式語音辨識系統抽取之基頻參數的分析比較，和加入改良式的分散式語音辨識系統抽取之基頻參數的國語連續數字串之辨識，都可以發現改良式的分散式語音辨識系統基頻參數之抽取的效能，在低訊噪比的環境下的確有了大幅度的改進，但是仍然不夠好。同時也發現到一個現象：在高訊噪比時以加入改良式的分散式語音辨識系統抽取之基頻參數的辨識結果最好；而低訊噪比以沒有加入基頻參數的辨識結果最好。所以想到可以結合兩個辨識器的優點，合併為一個在所有訊噪比下都有最好的辨識率之語音辨識器。依據此想法而建立的語音辨識器之辨識結果，確實是比其他語音辨識器還要好。



為了能夠以語音輸入介面取代傳統式的鍵盤、按鍵之輸入介面為出發點，所以建立在分散式語音辨識系統環境下之大字彙國語連續語音辨認器是相當重要工作。第五章主要就是在說明如何建立一個在分散式語音辨識系統環境下使用基頻參數的大字彙國語連續語音辨識器。

## 6.2 展望

在本論文中提出了在高訊噪比時，能夠有效的利用基頻參數抵抗環境雜訊的干擾；雖然也成功降低了在低訊噪比時環境雜訊對語音的週期特性的干擾，但還是不盡理想。未來應該要朝再提升低訊噪比環境下之辨識率的方向努力。

## 參考文獻

- [1]. ETSI standard document, “Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end reconstruction algorithm”, ETSI Standard ES 202 212, Nov., 2003.
- [2]. Hans-Günter Hirsch, David Pearce, “The AURORA Experimental Framework for The Performance Evaluation of Speech Recognition Systems Under Noisy Conditions”, ISCA ITRW ASR2000, Paris, France, September 18-20, 2000.
- [3]. AURORA Database, <http://www.elda.org/article20.html>.
- [4]. WIKIPEDIA, [http://en.wikipedia.org/wiki/Signal-to-noise\\_ratio](http://en.wikipedia.org/wiki/Signal-to-noise_ratio).
- [5]. “DSR Front-end Extension for Tonal-language Recognition and Speech Reconstruction.” *Aurora Group Meeting*, April 2003, by IBM & Motorola, [http://portal.etsi.org/stq/DSR\\_Presentations/Presentation.pps](http://portal.etsi.org/stq/DSR_Presentations/Presentation.pps).
- [6]. Dau-Cheng Lyu, et al, "Large Vocabulary Taiwanese (Min-nan) Speech Recognition Using Tone Features and Statistical Pronunciation Modeling" In Proc. EuroSpeech, Switzerland, 2003.
- [7]. MathWorld, <http://mathworld.wolfram.com/SigmoidFunction.html>.
- [8]. S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, “The HTK Book ( for HTK Version 3.2.1 )”.