國 立 交 通 大 學

電信工程學系

碩 士 論 文

以特徵參數正規化為基礎之強健性語音辨認

Robust Speech Recognition Based on Feature Normalization

研 究 生：高世哲

指導教授：陳信宏　博士

中華民國九十五年八月

以特徵參數正規化為基礎之強健性語音辨認
Robust Speech Recognition Based on Feature Normalization

研究生：高世哲　　Student：Shyh-Jer Kao

指導教授:陳信宏 博士　　Advisor: Sin-Horng Chean

國立交通大學

電信工程學系碩士班

碩士論文

A Thesis

Submitted to Institute of Commmunication Engineering

College of Electrical Engineering and Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master of Science

in

Electrical Engineering

August 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年八月

以特徵參數正規化為基礎之強健性語音辨認

研究生：高世哲　　指導教授：陳信宏 博士

國立交通大學電信工程學系碩士班

中文摘要

在本論文中，主要是針對強健性語音特徵參數作深入的探討，將現有的倒頻譜正規化法及分佈等化法做些許的改進。我們將分佈等化法加上 ARMA 濾波器後，經由國語數字串辨認實驗，辨識率從 80.08%提升到 82.03%。另外，我們提出的分兩群式 MVA 系統，也在經過改良後，辨識率由傳統 MVA 系統的 81.31%提升到 82.26%，同時，我們也做了理想分群 MVA 系統實驗，得知若準確分群，辨識率可提升至 83.63%。最後，我們利用正確的基頻將語音再多分一群，理想三群式 MVA 系統實驗結果顯示，辨識率可達 86.25%。

關鍵詞： 倒頻譜正規化法、分佈等化法、ARMA 濾波器、MVA 系統，基頻

Robust Speech Recognition Based on Feature Normalization

Student： Shyh-Jer Kao Advisor: Dr. Sin-Horng Chean

Department of Communication Engineering

National Chiao Tung University

Abstract

In this thesis. Some robust speech feature processing algorithms were proposed, in order to improve the speech recognition performance under the noisy environments . First, the well-known robust speech feature processing algorithms such as mean variance normalization(MVN) and histogram equalization(HEQ) was implemented in a Mandarin AURORA-like system database as the base-line system. Then, the class-based MVA was proposed to further implement the speech recognition performance. The class-based MVA algorithm was first categorized the signal into speech and non-speech parts and applied MVAs to each class separately. A 82.26% recognition rate can be achieved comparing to 81.31% in traditional MVA. Final, a Three-class voiced, unvoiced and non-speech MVA was investigated. A 86.25% recognition rate can be achieved under the ideal category of voiced/unvoiced/non-speech case.

Keywords: mean and variance normalization(MVN), Histogram(HEQ),AURORA,MVA

誌謝

在這研究所的兩年，非常感謝陳信宏老師及王逸如老師，身為陳老師的指導學生，真的是非常榮幸，老師的氣度及寬宏大量，讓我非常敬佩也非常感動;而王老師也讓我學習到非常多的東西，您的苦口婆心，一對一的單獨指導，真的讓我得到許多教訓及學到許多做事的方法，最後，再一次謝謝兩位老師的調教，讓我在讀碩士的生涯中，成長非常多。

接下來要感謝實驗室的學長，智合和性獸學長，非常謝謝你們從我一進來，就開始幫忙我及照顧我，並幫我解決許多問題，至於愛把妹阿德學長，也非常感謝你常常搞笑讓我的心情放鬆不少，而斯文的希群學長，也謝謝你兩年的照顧，還有坐我隔壁的長輩輝哥，也非常感謝你平時的叮嚀，讓我了解許多事情，最後要感謝之前畢業的學長們，謝謝你們一年的照顧，特別是柏宣學長，真的幫了我非常多的事，也讓我學到許多東西。

而在一起打拼兩年多的同學，包括國興、東毅、見偟、鴻彥、振豐、世帆、阿 Paul、家勇，謝謝你的平時的幫忙及照顧，幸虧有你們的陪伴，我才能撐過這兩年，希望在將來，還是有機會可以互相勉勵，在社會中一起打拼;而可愛的學弟們，也謝謝你們在我最後一年裡，陪我玩耍、打球及修課。

在這兩年中，也要感謝科技管理所的老師及同學，還有電信所的俊傑，謝謝你們的幫忙及協助，讓我可以順利的拿到了科管的輔所及創業大賽第三名。

最後，要感謝我親愛的父母、家人、朋友，還有我最可愛的女朋友，由於你們的支持，我才能努力到現在，謝謝你們!!

目錄

表目錄

圖目錄

第一章　導論

1.1 研究動機

人類在幾千年的演化過程中，智慧不斷的累積傳承，因此過去文明變遷和人類演化的步伐是一致的。而如今科技進化的速度，卻早已大大的超越了人類演化的速度。日常生活中可以使用的多媒體影音資訊越來越多，例如廣播電視節目，語音信件，演講錄影和數位典藏等。這些多媒體資訊可以從網路上大量地取得，成為傳統文字資訊外社會大眾廣泛使用的資訊來源。顯而易見的是，在上述的絕大部份多媒體中，語音可以說是最具語意的主要內涵之一。

語音是人類間最自然的傳播和接收訊息的媒介，若能以語音作為我們和電腦、手機或是其他電子產品之間的溝通方式，相信它可以成為一個有效且友善的人機介面。長久以來，卡通影片中的機器人多半是以語音控制的，反映出這個夢想一直存在人們的心中，而近年來許多學者及業界對語音辨識的理論及實務的貢獻已使得這個夢想逐步實現，有些手機已有語音撥號的功能，語音輸入的電腦軟體也不斷推陳出新。

語音辨識技術的應用是否會被大家廣泛的接受，其中一個很重要的因素就是辨識正確率的高低。使辨識下降的最重要原因之一，就是環境不匹配 (Environmental Mismatch)；當訓練聲學模型的環境和實際進行語音辨識的環境差異性很大時，辨識正確率便會大大地降低。本論文即是在發展一套能將環境不匹配所產生的辨識率下降有效減少的方法，以提高語音辨識系統的強健性 (Robustness)。

1.2 研究方向

近年來，已有許多學者對於環境不匹配所造成的問題提出了許多解決的方法，主要可以分作三類[1]:

(一) 語音模型調適技術（Speech Model Adaptation）

(二) 語音強化技術（Speech Enhancement Techniques）

(三) 強健性語音特徵參數（Robust Speech Features）

本論文主要的研究方向是針對強健性語音特徵參數作深入的探討，像是目前普遍使用的倒頻譜平均消去法（Cepstral Mean Normalization，CMN）[2]、倒頻譜正規化法（Cepstral Mean and Variance Normalization，MVN）[3]、分佈等化法（Histogram Equalization，HEQ）[4][5]，以及將倒頻譜正規化法結合 ARMA (Auto-Regression Moving Average)濾波器的 MVA[6][7]，並且將歐洲電信標準協會(European Telecommunications Standards Institute，ETSI)所提出的分散式語音辨識系統(Distributed Speech Recognition，DSR)延伸進階參數抽取技術(Extended Advanced Front-end Feature Extraction，XAFE)[8]作相關的探討與比較。本論文同時也針對分佈等化法及 MVA 提出了改良的方法，藉此來提升在雜訊干擾下的辨識效能，而詳細的細節及研究成果將於第四章說明。

1.3 章節概要

第一章 導論

第二章 背景理論

第三章 基礎系統建立

第四章 分群式倒頻譜正規化法

第五章 結論與展望

第二章　背景理論

本章節將針對語音辨識中的強健式語音特徵參數技術，作一些介紹，而本論文在稍後的章節亦將對這些技術中的若干項提出改進之方法。本章節的內容如下：第一小節介紹倒頻譜平均消去法和倒頻譜正規化法，第二小節則介紹分佈等化法，第三小節介紹ARMA濾波器，第四小節則介紹ETSI的分散式語音辨識系統中的延伸前端參數處理流程。

2.1 倒頻譜平均消去法與倒頻譜正規化法

2.1.1 倒頻譜平均消去法 （Cepstrum Mean Subtraction, CMS）

倒頻譜平均消去法主要是針對特徵向量的第一動差（First Order Moment）做正規化。假設 $X^i=\{X_0^i, X_1^i, ...X_{T-1}^i\}$ 是從含雜訊噪音的語音訊號所擷取出來的語音倒頻譜特徵參數向量在第i維度所形成的序列。其中 X_t^i 代表的是位於時間t 所獲得的第i維度特徵參數。$\overline{X}^i$代表的是所有時間點上第i維度特徵參數的平均值。而經過本方法處理後的特徵參數$\mathring{X}_t^i$以及本方法的步驟，可用式 (2-1) 與(2-2) 表示之：

$$\overline{X}^i=\frac{1}{T}\sum_{t=0}^{T-1}X_t^i \tag{2-1}$$

$$\mathring{X}_t^i=X_t^i-\overline{X}^i \tag{2-2}$$

經過倒頻譜平均消去法處理後之各維特徵參數的平均值必為零，使特徵參數不受抽取時聲學環境影響，因此減低了環境不匹配對特徵參數影響的程度。

2.1.2 倒頻譜正規化法

倒頻譜正規化法除了減去倒頻譜特徵參數的平均值外，更針對特徵向量的第二動差（Second Order Moment）做正規化。假設σ^i是所有時間點上第i 維度特徵參數的標準差。而經過本方法處理後的特徵參數$\overset{\circ}{X}_t^{\,i}$，可用式 (2-3)，(2-4)，(2-5)表示之：

$$\overline{X}^{\,i} = \frac{1}{T}\sum_{t=0}^{T-1} X_t^{\,i} \tag{2-3}$$

$$\sigma^i = \sqrt{\frac{1}{T}\sum_{t=0}^{T-1}(X_t^{\,i} - \overline{X}^i)^2} \tag{2-4}$$

$$\overset{\circ}{X}_t^{\,i} = \frac{X_t^{\,i} - \overline{X}^i}{\sigma^i} \tag{2-5}$$

經過倒頻譜正規化法處理後之各維特徵參數不僅平均值必為零，其變異數亦必為 1，兩者皆不受特徵參數抽取時的聲學環境影響；和倒頻譜平均正規化法相較，此方法又可進一步減低環境不匹配對特徵參數影響的程度。

2.2 分佈等化法

雜訊通常會造成特徵參數一種非線性的破壞，如圖 2-1 所示，經雜訊干擾的語音，其參數分佈也會因此改變，所以一般線性的補償方法是無法解決此問題的。

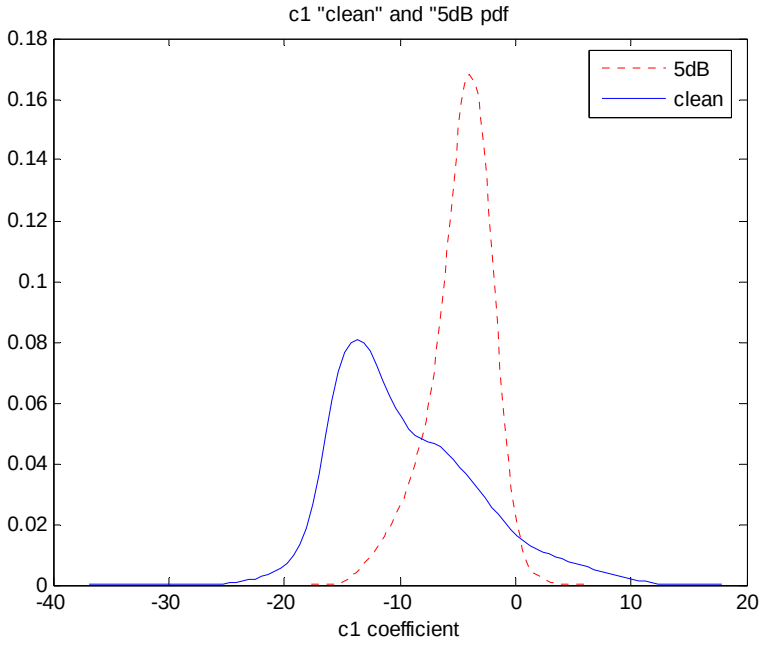

圖 2-1:乾淨及 5dB 測試語料的第一維倒頻譜係數機率分佈圖

分佈等化法是將參數的值做一種單調(monotonic)的轉換，經轉換的值，會在一固定範圍之內，且大小順序不變，可使乾淨及受雜訊干擾的倒頻譜係數機率分佈變為相同分佈，由圖 2-2 可知，經過分佈等化法後的分佈，其分佈與乾淨語音分佈最爲匹配。

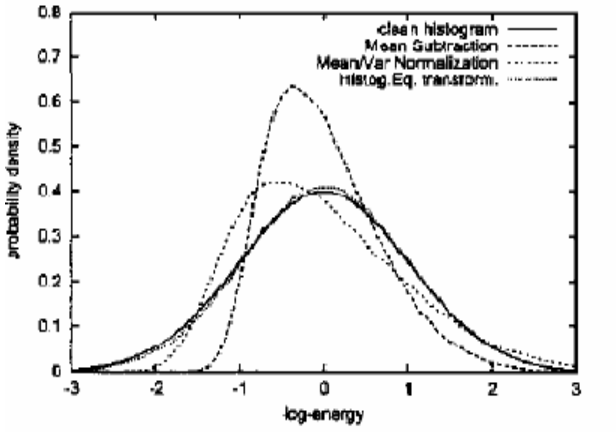

圖2-2:乾淨語音、及帶有雜訊語音分別經CMN、MVN、HEQ之log-energy係數機率分佈圖 (From [4])

而分佈等化法的演算法，是假設 $X^{(i)}$ 是特徵向量中第 i 維特徵參數所形成之序列，其中 $X^{(i)}_t$ 是時間 t 所得到的第 i 維特徵參數，$CDF_{X^{(i)}}$ 為利用 $X^{(i)}$ 的分佈圖來近似產生的累積分布函數，$CDF_{\bar{X}^{(i)}}$ 為參考機率分佈(reference probability distribution)的累積分布函數，本文採用的參考分佈為標準高斯分佈(Standard Gauss Distribution)，而 $\bar{X}^{(i)}_t$ 為經分佈等化法處理後的特徵參數，且必須滿足下列數學式:

$$CDF_{\bar{X}^{(i)}}(\bar{X}^{(i)}_t) = CDF_{X^{(i)}}(X^{(i)}_t) \tag{2-6}$$

也就是:

$$\bar{X}^{(i)}_t = CDF^{-1}_{\bar{X}^{(i)}}[CDF_{X^{(i)}}(X^{(i)}_t)] \tag{2-7}$$

經過分佈等化法處理過的各維特徵參數，都會有著相同的分佈(histogram)，可以有效的非線性補償了訓練及辨識在不同的環境條件下，而且不僅將平均值與變異數相關的第一、第二動差正規化成零和1，其餘更高階的動差也均被正規化至定值，不受特徵參數抽取時的聲學環境影響，因此大大減低了環境不匹配對特徵參數影響的程度。

而有關於分佈等化法近來的研究，也是相當多，像是利用同步式的方法，來解決雜訊不穩定的現象[9][10];以及利用不同的語音特徵參數，像是時域頻域主成分特徵參數(Time-frequency principal components)[11]及最小變化失真響應特徵參數(Minimum Variance Distortionless Response ,MVDR)[12] ，做分佈等化法的處理;也有人將分佈等化法，對 13 維的倒頻譜係數差量及二階差量來共 39 維參數都做分佈等化法處理[13]，而在與其它種類方法結合上，像是與頻譜雜訊消去法(spectral noise reduction)所做的結合[14]，也有相當不錯的加成效果。

2.3 ARMA 濾波器

本論文所使用的ARMA濾波器是屬於低通濾波器的一種，近年來被用來在語音倒頻譜參數領域中，與倒頻譜正規化法相結合，並有不錯的加成效果，主要的功能

是用來緩和(smooth)特徵參數序列，消除參數變化過快的現象，以符合語音信號在短時間內變化較小的現象，其表示式如下

$$\hat{X}_{td}^{2} = \frac{\sum_{i=0}^{M} \overset{\circ}{X}_{(t-i)d} + \sum_{j=0}^{M} \overline{X}_{(t+j)d}}{2M+1} \tag{2-8}$$

其中 M 為濾波器階數(order)，階數愈高，所緩和的程度就愈大，適合在較低訊噪比的環境下使用，相對而言，也會造成一些辨識訊息的遺失，而圖 2-3 分別為二階及四階的頻率響應圖。

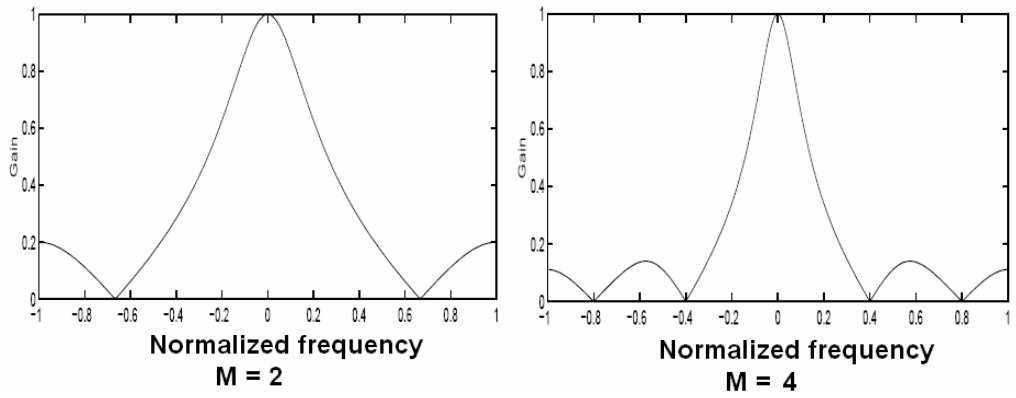

圖2-3:ARMA濾波器二階及四階之正規化頻率響應圖 (From [6])

2.4 分散式語音辨識之延伸進階前端處理

分散式語音辨識系統主要的構想是：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的不足。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端（DSR back-end）進行解碼以及辨識。本論文中之延伸進階前端參數處理是使用ETSI ES 202 212 V1.1.1之分散式語音辨識系統前級的標準，圖 2-4則是其系統的架構圖。

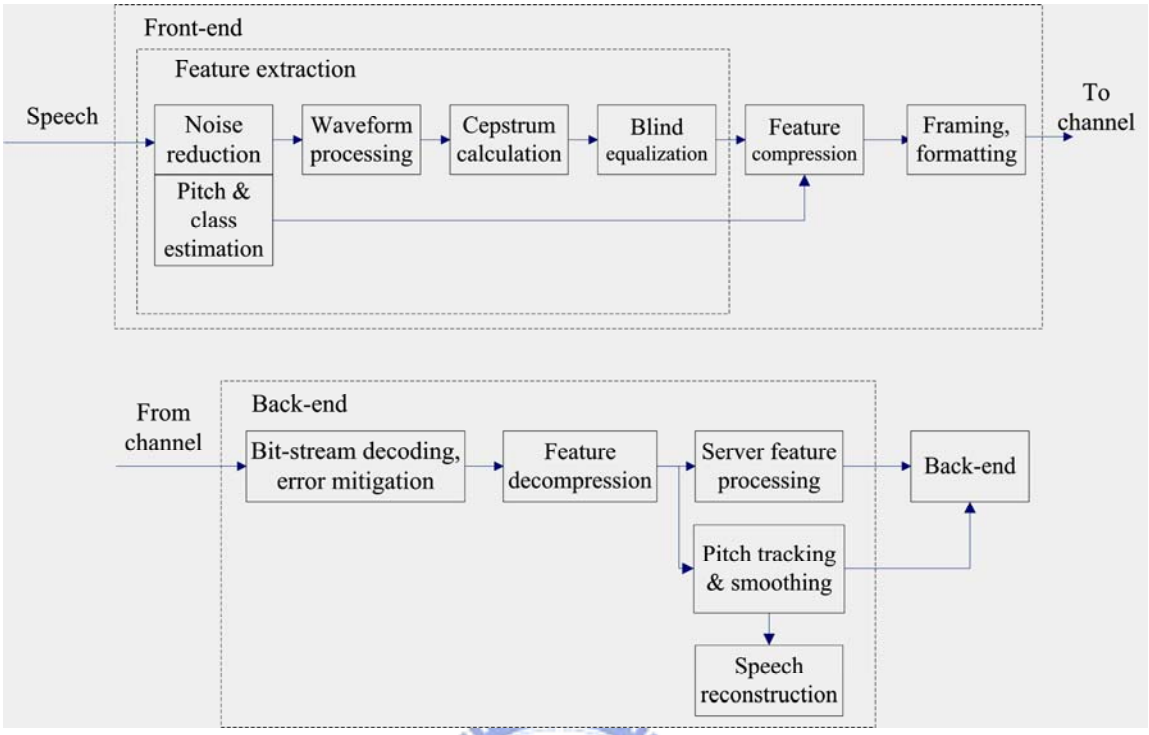

圖 2-4：分散式語音系統架構圖 (From [8])

當使用手持設備時，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的特徵參數抽取特別加入了降低雜訊（Noise reduction）的處理，其延伸進階前端參數處理流程如圖 2-5。

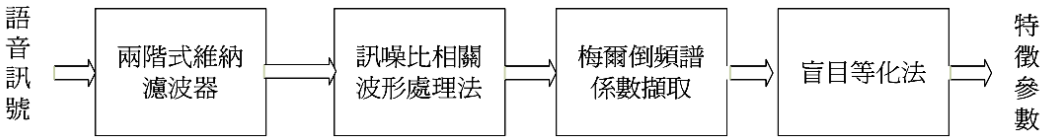

圖 2-5：延伸進階前端參數處理流程圖

其中兩階式維納濾波器可有效減小加成性雜訊，而訊噪比相關波形處理法，主要是用來強調語音波形訊號能量較高的部分，可讓語音整體的訊噪比提升，至於盲目等化法則是針對通道特性不匹配所做的補償，經由此流程所抽取的參數，較能夠抵抗雜訊的干擾。

第三章　基礎系統建立

本章敘述實驗所用到國語連續數字串語料庫、語音特徵參數的擷取、聲學模型及基礎系統之實驗結果，並提出分佈等化法結合ARMA濾波器的新系統。

3.1 國語連續數字串語料庫[16]

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3-1 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 3-1：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8 kHz 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 8 kHz。

3.1.1 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ETSI AURORA 2中提供的環境雜訊，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3-2 表示每個環境雜訊的音檔長度。

表 3-2：八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，也不一定會比乾淨語料還要長，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3-1 以圖示說明。在圖3-1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。

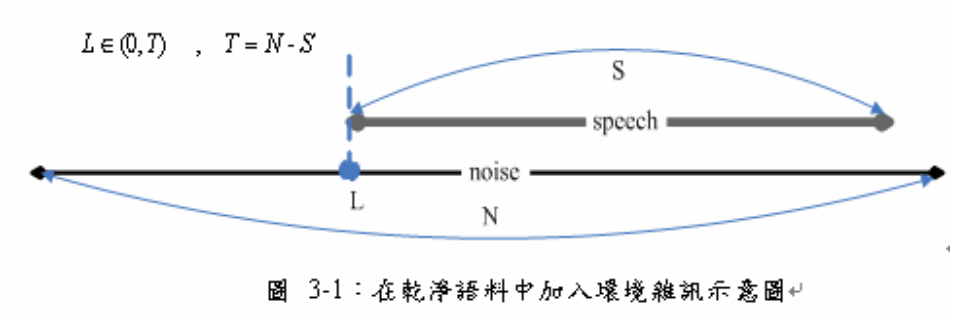

圖 3-1：在乾淨語料中加入環境雜訊示意圖

圖 3-1：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：

$$P=\frac{1}{M}\sum_{i=1}^{M}x^{2}(i) \tag{3-1}$$

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，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。

$$SNR=10\log(P_S)-10\log(P_N')\Rightarrow G=\left(\frac{P_S}{P_N}10^{\frac{-SNR}{10}}\right)^{\frac{1}{2}} \tag{3-2}$$

其中 $P_S = \frac{1}{M}\sum_{i=1}^{M} x_S^2(i)$ ， $P_N = \frac{1}{K}\sum_{i=1}^{K} x_S^2(i)$ ， $P_N^{'} = G^2 * P_N$ ， 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 3-2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 3-2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 3-3 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。

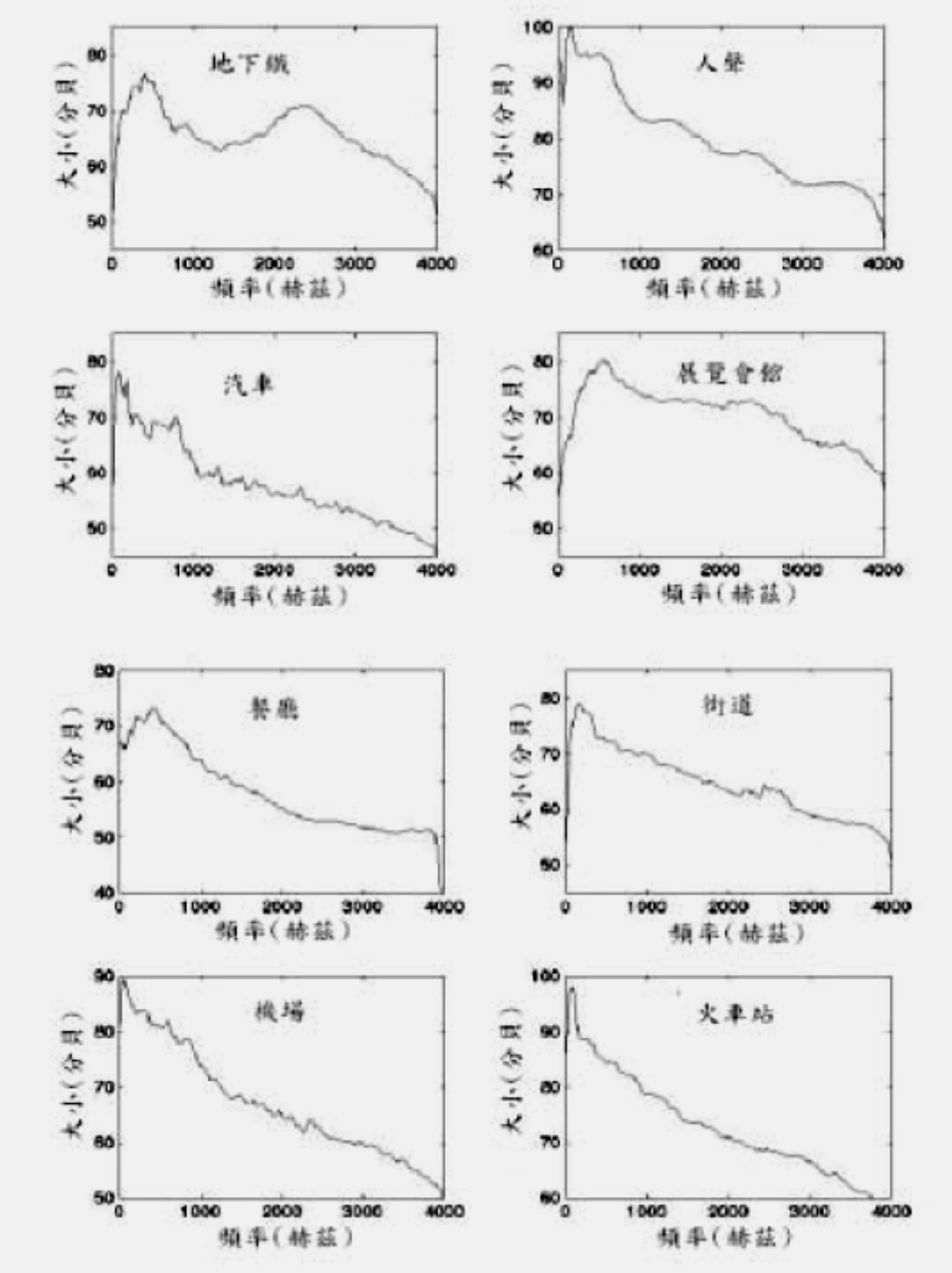

圖 3-2：八種環境雜訊的長時間頻譜 (from [15])

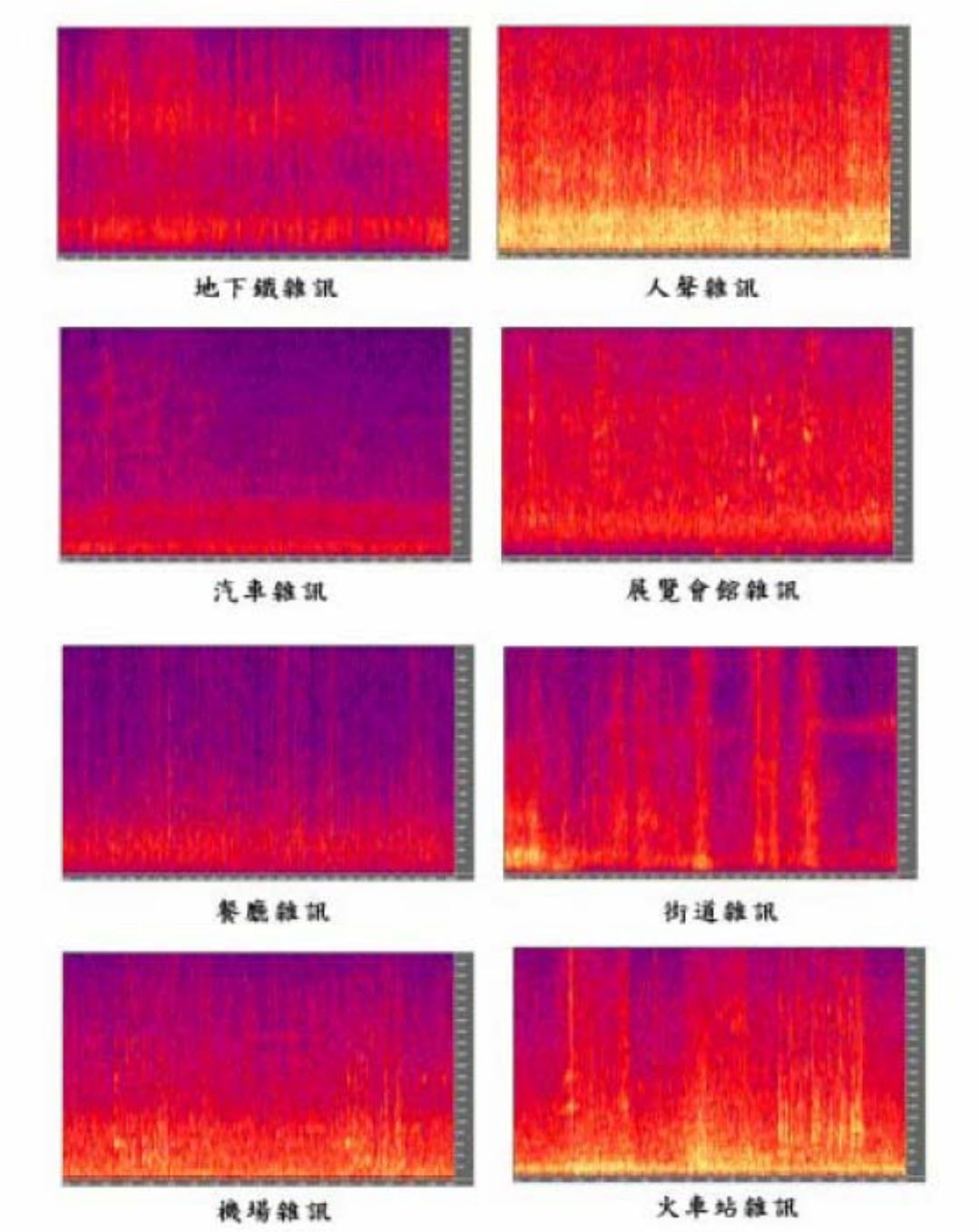

圖 3-3：八種環境雜訊的頻譜－時間圖（橫軸：時間；縱軸：頻率）(from [15])

3.2 基礎系統建立

3.2.1 隱藏式馬可夫模型之語音辨識器

在實驗中，基礎系統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 3-3 所示。

表 3-3：加上環境雜訊的國語連續數字串內容介紹

<table>
<tr><th colspan="4">國語連續數字串語料庫</th></tr>
<tr><td>取樣頻率</td><td colspan="3">8 kHz</td></tr>
<tr><td rowspan="2">訓練模式</td><td>乾淨語音訓練</td><td colspan="2">複合情境訓練</td></tr>
<tr><td>音段數：900
環境雜訊：
無</td><td colspan="2">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</td></tr>
<tr><td rowspan="3">測試組合</td><td colspan="2">A 組</td><td>B 組</td></tr>
<tr><td colspan="2">音段數 :2,800
環境雜訊 ：
地下鐵、人聲、汽車、展覽會館</td><td>音段數 ：2,800
環境雜訊 ：
餐廳、街道、機場、火車站</td></tr>
<tr><td colspan="3">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</td></tr>
</table>

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。

本實驗使用的語音辨認參數是 12 維梅爾倒頻譜係數(Mel Frequency Cepstral Coefficients, MFCC)，加上其一維與二維的變化量，以及能量及其一維與二維的變化量，共 39 維特徵向量。表 3-4 列出特徵參數抽取過程中各項參數設定。其中前五項是如同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 3-4：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)	$1-0.97z^{-1}$
梅爾濾波器組(Mel-frequency filter bank)	23 個濾波器
語音特徵向量(Speech feature vector)	39 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，而訓練聲學模型流程如圖 3-4。

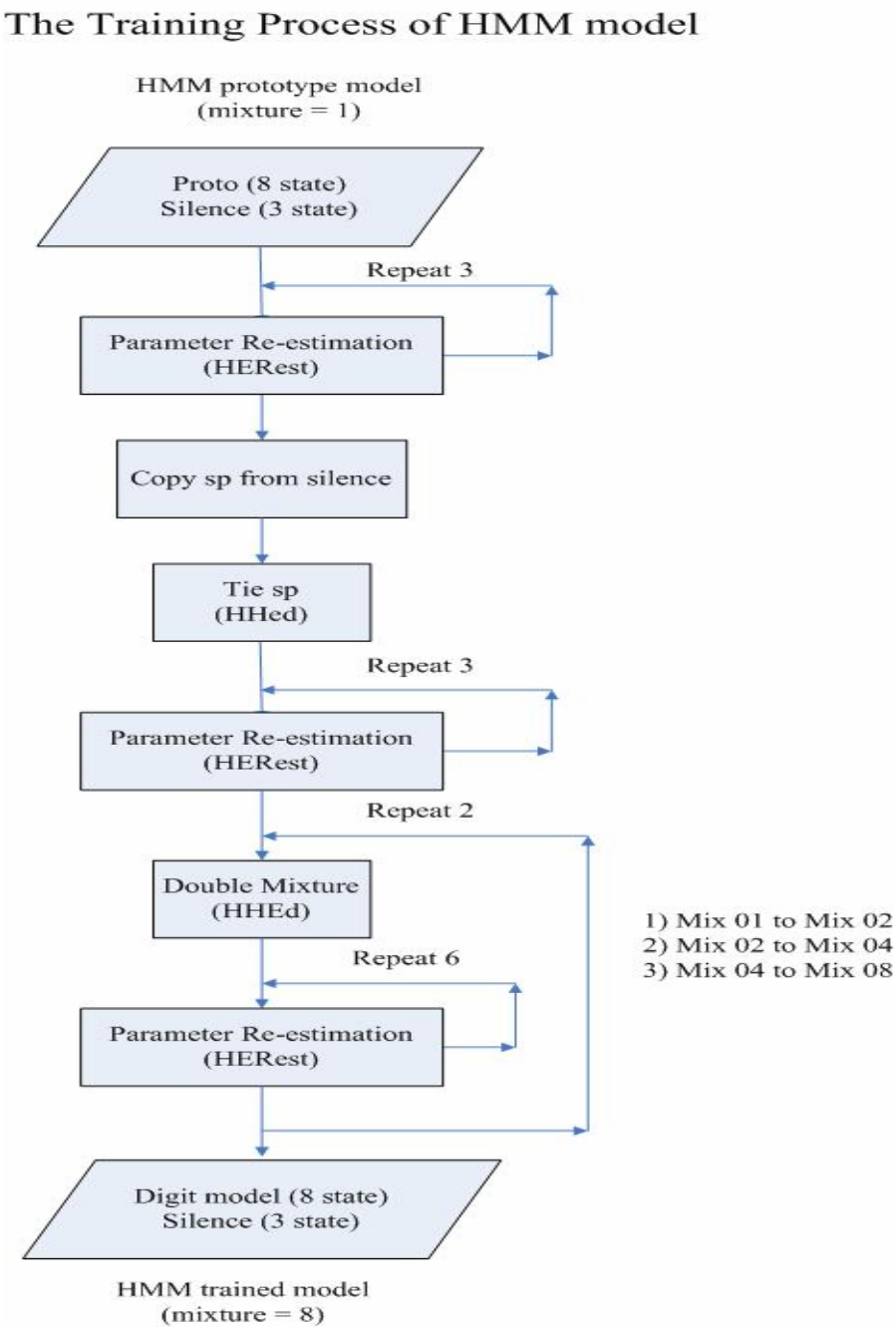

圖 3-4 聲學模型訓練流程

3.2.2 基礎系統實驗結果

3.2.2.1 分佈等化法

表 3-5 及表 3-6 分為梅爾倒頻譜係數經分佈等化法後，在乾淨語音訓練模式及複合情境訓練模式下的辨識結果。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 3-5：經分佈等化法之實驗結果，乾淨語音訓練模式

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	98.91				
20	93.3	96.42	96.42	94.86	95.25
15	90.03	93.77	94.39	89.25	91.86
10	82.87	85.2	90.81	83.49	85.59
5	67.45	75.08	79.91	64.95	71.84
0	48.91	47.2	57.63	43.77	49.37
-5	22.43	24.43	36.08	23.68	26.65
平均值(20dB～0dB)	76.51	79.53	83.83	75.26	78.78
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	98.91				
20	95.48	96.57	96.57	97.2	96.45
15	92.21	92.52	95.48	95.48	93.92
10	83.02	86.14	89.25	92.06	87.61
5	68.07	74.45	78.97	81.93	75.85
0	42.99	52.96	55.14	61.06	53.03
-5	22.28	23.42	24.86	37.52	27.02
平均值(20dB～0dB)	76.35	80.52	83.08	85.54	81.37
八種環境雜訊及五種訊噪比的平均值					80.08

表 3-6：經分佈等化法之實驗結果，複合情境語音訓練模式

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	97.51				
20	97.04	97.51	97.35	97.35	97.31
15	95.17	96.11	97.04	96.88	96.3
10	90.03	94.08	96.11	91.74	92.99
5	81	85.51	90.65	82.87	85.00
0	59.03	63.86	74.61	60.75	64.56
-5	31.15	31.15	43.46	27.41	33.29
平均值(20dB～0dB)	84.45	87.41	91.15	85.91	87.23
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	97.51				
20	97.2	98.13	96.11	97.51	97.23
15	96.26	96.42	96.26	97.35	96.57
10	92.21	92.37	93.15	95.79	93.38
5	77.57	85.98	87.69	90.19	85.35
0	52.96	63.55	70.4	71.03	64.48
-5	27.88	32.87	36.92	45.95	35.90
平均值(20dB～0dB)	83.24	87.29	88.72	90.37	87.40
八種環境雜訊及五種訊噪比的平均值					87.32

3.2.2.2 倒頻譜正規化法與 ARMA 濾波器之結合

圖 3-5 為倒頻譜正規化法與 ARMA 濾波器之結合之架構圖，其流程為先將梅爾倒頻譜參數先經過倒頻譜正規化法，再經過 ARMA 濾波器;表 3-7 及表 3-8 分為在乾淨語音訓練模式及複合情境訓練模式下的辨識結果。

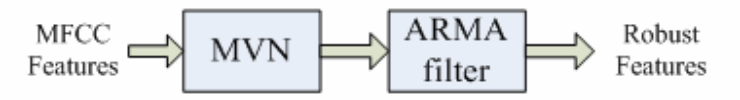

圖 3-5 倒頻譜正規化法與 ARMA 濾波器之結合之架構圖

表 3-7：倒頻譜正規化法+ARMA 濾波器實驗結果，乾淨語音訓練模式

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	98.44				
20	95.33	97.66	97.04	95.02	96.26
15	92.06	94.86	95.17	92.83	93.73
10	86.92	88.16	91.12	82.87	87.26
5	70.09	75.08	82.24	67.45	73.71
0	47.66	47.2	62.31	44.08	50.31
-5	26.01	21.34	31.62	19.31	24.57
平均值(20dB～0dB)	78.41	80.59	85.58	76.45	80.25
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	98.44				
20	96.88	97.04	97.51	98.13	97.39
15	92.68	94.24	94.24	96.88	94.51
10	82.58	87.23	90.5	93.3	88.40
5	67.13	77.88	81.46	83.96	77.60
0	43.15	53.58	56.07	62.77	53.89
-5	22.43	31	27.73	32.4	28.39
平均值(20dB～0dB)	76.48	81.99	83.96	87.00	82.36
八種環境雜訊及五種訊噪比的平均值					81.31

表 3-8：倒頻譜正規化法+ARMA 濾波器經分佈等化法，複合情境訓練模式

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	97.35				
20	97.2	98.29	97.82	97.04	97.58
15	95.02	97.2	97.66	95.02	96.22
10	91.9	93.46	96.11	91.28	93.18
5	80.53	83.8	88.94	81.46	83.68
0	58.72	65.73	71.03	58.57	63.51
-5	32.87	31.31	43.3	27.73	33.80
平均值(20dB～0dB)	84.67	87.69	90.31	84.67	86.83
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	97.35				
20	97.98	97.66	97.66	97.98	97.82
15	97.04	95.95	97.2	97.51	96.92
10	92.37	90.81	95.48	96.57	93.80
5	82.24	85.67	88.94	90.19	86.76
0	54.36	62.46	71.81	73.99	65.65
-5	28.57	38.32	38.63	49.07	38.64
平均值(20dB～0dB)	84.79	86.51	90.21	91.24	88.19
八種環境雜訊及五種訊噪比的平均值					87.51

3.2.2.3 分散式語音辨識系統之延伸進階前端參數處理

表 3-9 及表 3-10 分別為梅爾倒頻譜參數經過分散式語音辨識系統之延伸進階前端處理後，在乾淨語音訓練模式及複合情境訓練模式下的辨識結果。

表 3-9：延伸進階前端處理實驗結果，乾淨語音訓練模式

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 3-10：延伸進階前端處理實驗結果，複合情境訓練模式

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%

3.2.3 實驗結果討論

圖 3-5 為三種基礎系統在乾淨語音訓練模式及複合情境訓練模式下的 20 到 0dB 平均辨識率比較圖，可以觀察到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，主要是因為此時複合情境訓練模式所產生的聲學模型反而和乾淨測試語料存在較不匹配的關係，而 MVA 在兩種訓練模式的辨識率都表現最佳。

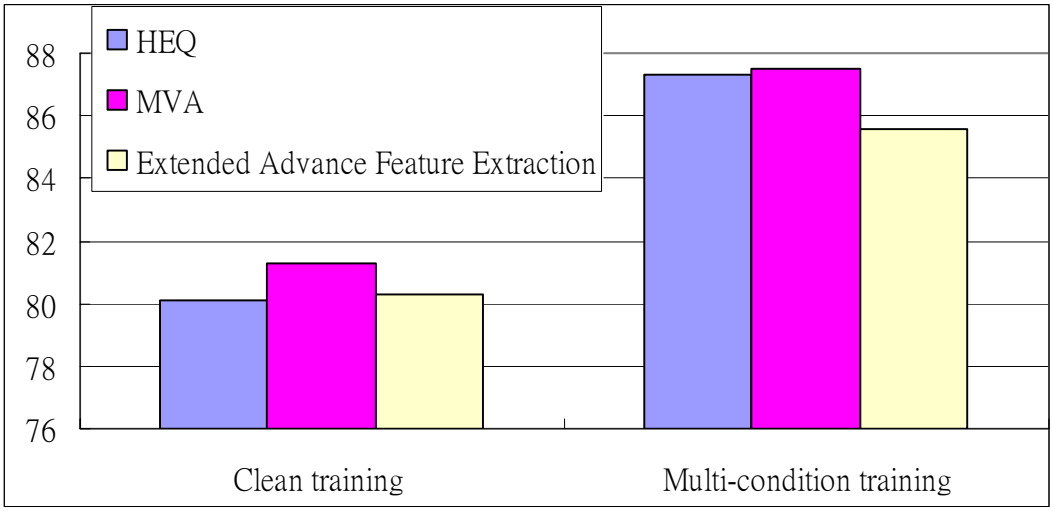

圖 3-6: 基礎系統 20-0dB 平均辨識率比較圖

3.3 分佈等化法與 ARMA 濾波器之結合

3.3.1 系統構想

由上一小節可以觀察倒頻譜正規化法加上 ARMA 濾波器後，經過實驗證實，有不錯的加成性效果，可有效的提升辨識率，特別是在低訊噪比的情況下。在先前的 2.1 小節有介紹到倒頻譜正規化法也就是將特徵向量的一階及二階動差正規化，而我們知道分佈等化法又為倒頻譜正規法的延伸，也就是各階動差都被正規化，因此，我們將分佈等化法與 ARMA 濾波器做結合，相信可以提升辨識率。

如圖 3-6 為分佈等化法與 ARMA 濾波器之結合之架構圖，我們將梅爾倒頻譜參數先經先經過分佈等化法後，再經過先前基礎系統 MVA 所使用的二階 ARMA 濾波器，來得到較強健的語音特徵參數。

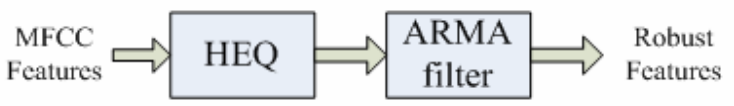

圖 3-7:分佈等化法與 ARMA 濾波器之結合之架構圖

3.3.2 實驗結果

表 3-11 及表 3-12 分別為梅爾倒頻譜參數經過分佈等化法結合 ARMA 濾波器後，在乾淨語音訓練模式及複合情境訓練模式下的辨識結果。

表 3-11:分佈等化法結合 ARMA 濾波器實驗結果，乾淨語音訓練模式

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	98.75				
20	95.33	96.42	96.88	94.86	95.87
15	91.9	94.7	94.86	90.5	92.99
10	84.58	88.94	93.61	85.46	88.14
5	70.87	76.01	84.27	67.76	74.72
0	51.71	49.69	68.07	44.7	53.54
-5	32.55	22.59	41.12	24.45	30.17
平均值(20dB～0dB)	78.87	81.15	87.53	76.65	81.05
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	98.75				
20	96.33	96.42	96.42	97.35	96.63
15	93.21	95.02	93.77	95.64	94.41
10	83.8	88.16	90.19	93.3	88.86
5	65.73	79.75	80.06	85.36	77.72
0	44.55	59.03	60.12	65.89	57.39
-5	23.96	33.49	28.82	44.24	32.62
平均值(20dB～0dB)	76.72	83.67	84.11	87.50	83.00
八種環境雜訊及五種訊噪比的平均值					82.03

表 3-12：分佈等化法結合 ARMA 濾波器，複合情境訓練模式

複合語音訓練					
訊噪比 (dB)	A 組				
	地下鐵	人聲	汽車	展覽會館	平均值
乾淨	97.51				
20	97.2	97.98	97.82	97.82	97.70
15	95.64	97.35	97.35	96.42	96.69
10	91.59	93.77	95.95	91.12	93.10
5	83.33	87.23	92.37	84.58	86.87
0	61.21	65.11	78.04	62.77	66.78
-5	36.14	34.27	53.74	34.11	39.56
平均值(20dB～0dB)	85.79	88.28	92.30	86.54	88.23
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	97.51				
20	97.04	97.51	96.73	97.82	97.27
15	95.95	96.57	96.42	96.88	96.45
10	91.74	92.99	94.24	96.57	93.88
5	78.66	87.85	87.07	90.5	86.02
0	54.98	67.13	73.52	73.99	67.40
-5	30.55	39.25	41.43	53.89	41.28
平均值(20dB～0dB)	83.67	88.41	89.59	91.15	88.20
八種環境雜訊及五種訊噪比的平均值					88.22

3.3.3 實驗結果討論

由圖 3-7 及圖 3-8 可以觀察到無論是乾淨語音訓練模式或複合情境訓練模式，分佈等化法結合 ARMA 濾波器後，在每一訊噪比情況下的辨識率，都有所提升，特別是在低訊噪比的情況下，而整體 20 到 0dB 的平均辨識率，也比 MVA 還要好，代表分佈等化法結合 ARMA 濾波器如同我們預測的，的確有不錯的加成效果，而複合情境訓練模式如先前的結果一樣，在有雜訊的情況下，辨識效能都比乾淨語音訓練模式的辨識率都還要好。

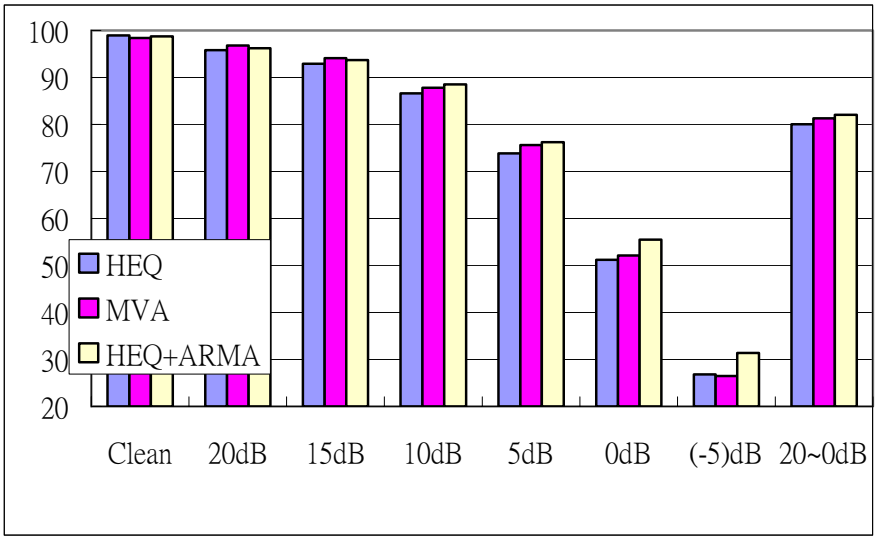

圖 3-8:乾淨語音訓練模式，HEQ、MVA、HEQ+ARMA 在各種訊噪比及 20~0dB 平均辨識率比較圖

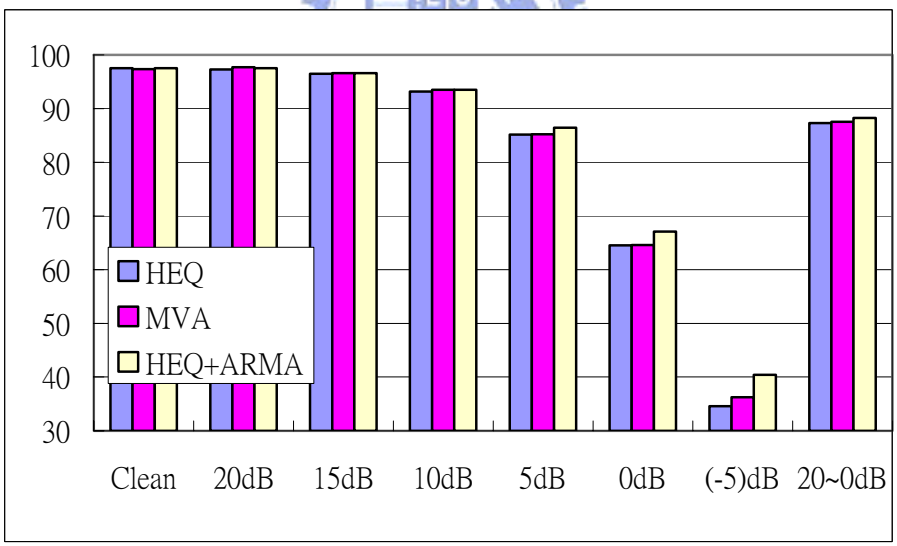

圖 3-9: 複合情境訓練模式，HEQ、MVA、HEQ+ARMA 在各種訊噪比及 20~0dB 平均辨識率比較圖

第四章　分群式倒頻譜正規化法

本章將介紹傳統倒頻譜正規化的潛在問題，並且提出分群式倒頻譜正規化法來解決，但我們所提出的分群式MVA系統在低訊噪比的情況下，辨識率將因分群效果不佳而下降，因此我們提出了改良式分群MVA系統來解決此問題，同時我們也進一步的利用已知的切割資訊，完成理想分群MVA實驗，來了解分群MVA系統的上限，最後並提出理想三群式MVA系統，來有效的提升辨識率。

4.1 傳統倒頻譜正規化法的潛在問題

圖4-1是對所有測試語料，梅爾倒頻譜第一維係數在乾淨及訊噪比5dB的機率密度函數，我們可以觀察到，當乾淨的語音加入雜訊後，其機率密度函數的分佈的狀況將有所改變，不僅分佈的平均值改變了，其分佈的變異數較相對的變小了。

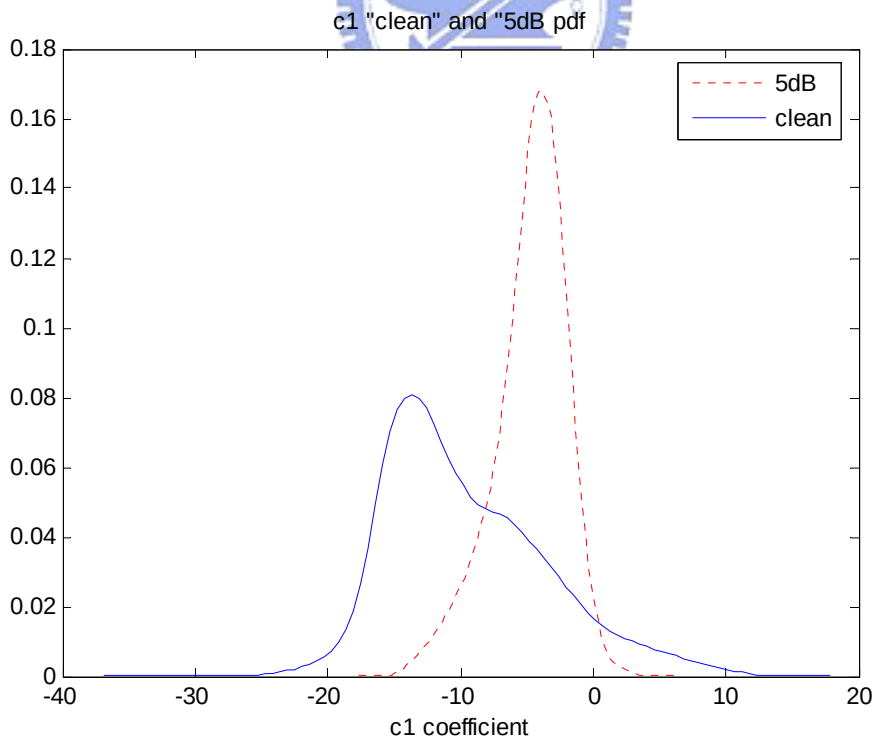

圖 4-1:乾淨及 5dB 測試語料的機率分佈

因此前人使用倒頻譜正規化法，來將乾淨語音與受雜訊干擾語音其機率密度函數的平均值及標準差正規化，也就是將兩者的機率密度函數轉換成平均值為 0，標準差為 1 的分佈，來減少它們的不匹配性，藉藉此改善辨識效能。

經倒頻譜正規化所處理過後的語料，如圖 4-2 所示，可觀察到乾淨與受雜訊干擾語料兩者的機率分佈，都變為平均值為 0，標準差為 1 的分佈，因此增加它們的匹配性，辨識效能也因此提升。

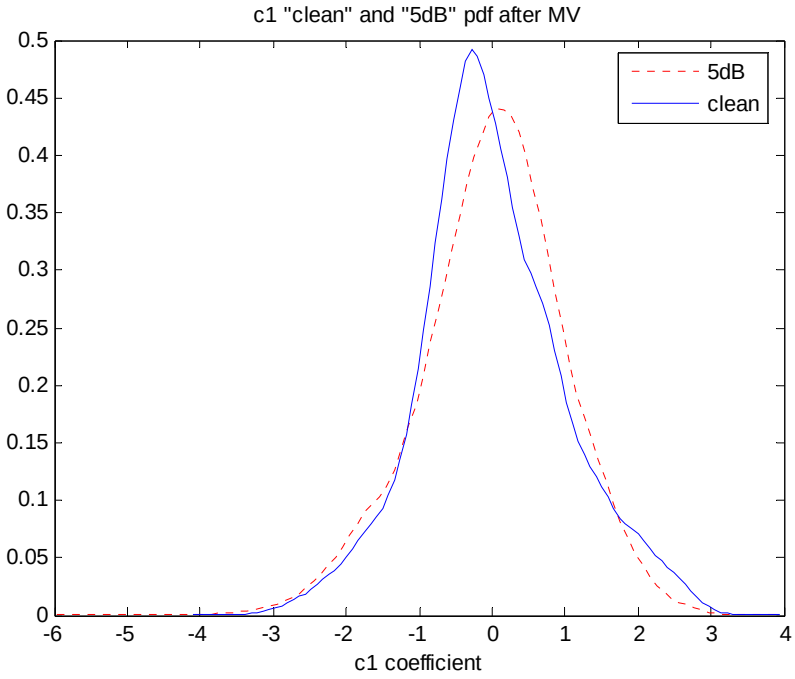

圖 4-2:經倒頻譜正規化處理後乾淨及 5dB 測試語料的機率分佈

而倒頻譜正規化法是將參數值做一種單調的(monotonic)轉換，也就是將參數的值，轉換成一固定範圍的值，但值的大小順序(order)依舊不變，然而事實上，語料的內容可以分為語音(speech)及非語音(non-speech)兩類，語音及非語音可能會因雜訊干擾，而造成一種參數值順序(order)的改變，因此造成倒頻譜正規化法的效果不彰。為了改善此情況，我們提出了一種想法，就是將語音及非語音分別做倒頻譜正規化法，相信可減少彼此間順序的干擾，藉此來提升辨識率。

4.2 分群倒頻譜正規化法

4.2.1 語音及非語音特性分析

如圖4-3，我們觀察乾淨語料的語音及非語音機率分佈，兩者的機率分佈不管是平均值或標準差，都因為語音特性的不同，有相當大的差距。

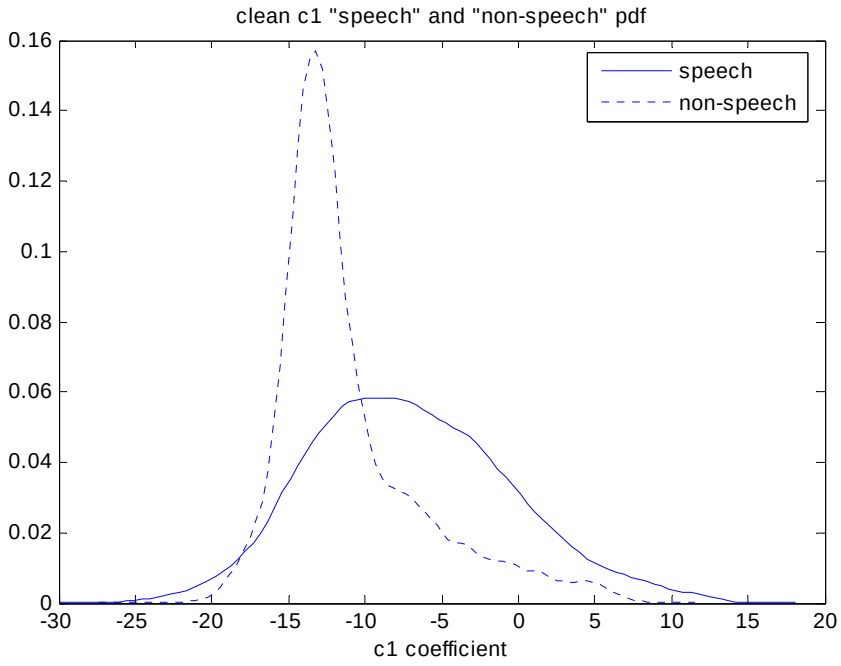

圖 4-3:乾淨語料語音及非語音的機率分佈

同樣的如圖4-4，我們觀察受雜訊干擾的語音及非語音機率分佈，可以發現到受雜訊干擾的機率分佈，不管是語音部分或非語音部分，其機率分佈的平均值都增加了，且機率分佈的標準差也有所改變，值得注意的是，我們可以從圖4-5得知語音及非語音機率分佈在雜訊的干擾下，其平均值與標準差的變化程度是有所差異的，且原本語音及非語音的順序(order)關係，也產生了變化，符合前述的說法我們在4.1所預測的狀況，會造成倒頻譜正規化法的效果不彰，因此我們提出了分群倒頻譜正規化法，來改善傳統倒頻譜正規化法的不足。

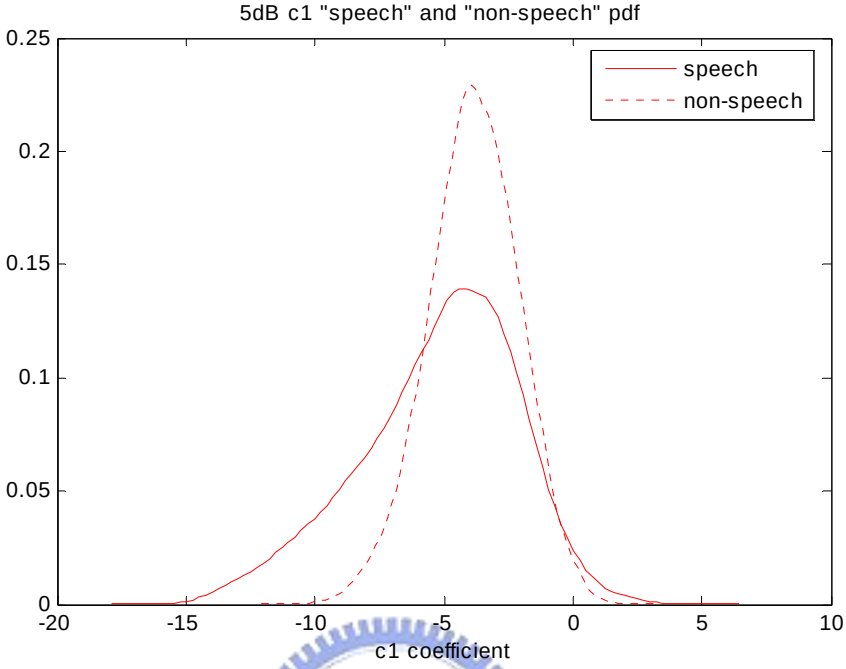

圖 4-4:測試語料在訊噪比 5dB 語音及非語音的分佈

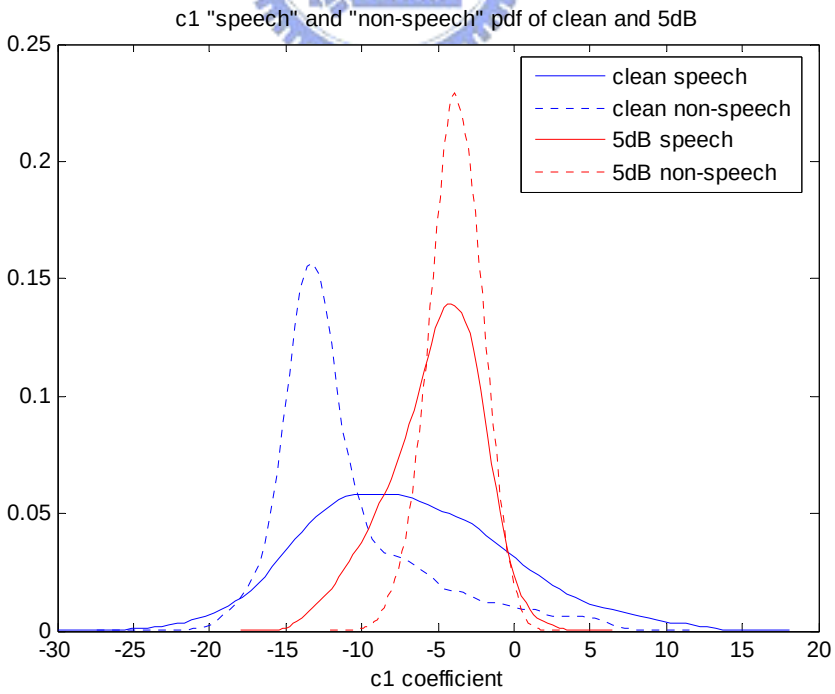

圖 4-5: 測試語料語音及非語音在乾淨與訊噪比 5dB 下的分佈

4.2.2 分群倒頻譜正規化法

類似之前所介紹的傳統式倒頻譜正規化法，所謂的分群倒頻譜正規化法，就是將語料內容分類為「語音」及「非語音」，再分別來做倒頻譜正規化法，其分群倒頻譜正規化法的表示式如下面的式子所示:

$$\hat{X}_s = \mu_{s,x} + (Y_s - \mu_{s,y})(\frac{\sigma_{s,x}}{\sigma_{s,y}})^{1/2} \quad (4\text{-}1)$$

$$\hat{X}_n = \mu_{n,x} + (Y_n - \mu_{n,y})(\frac{\sigma_{n,x}}{\sigma_{n,y}})^{1/2} \quad (4\text{-}2)$$

上面的式子中，$\mu_{s,x}$ 及 $\mu_{n,x}$ 分別代表的是所有訓練語料語音部分及非語音部分之平均值，$\sigma_{s,x}$ 及 $\sigma_{n,x}$ 分別代表的是所有訓練語料語音部分及非語音部分之標準差，Y_s 及 Y_n 分別代表的是測試語料語音部分及非語音部分之特徵參數，$\mu_{s,y}$ 及 $\mu_{n,y}$ 分別代表的是測試語料語音部分及非語音部分之平均值，$\sigma_{s,y}$ 及 $\sigma_{n,y}$ 分別代表的是測試語料語音部分及非語音部分之標準差，$\hat{X}_s$ 及 $\hat{X}_n$ 分別代表的是經分群倒頻譜正規化法後所求得的語音部分及非語音部分之特徵參數。

如圖 4-6 及圖 4-7 分別為經分群倒頻譜正規化法後的語音及非語音機密分佈在乾淨及受雜訊干擾情況下，我們可觀察到，經分群倒頻譜正規化法處理後的乾淨及受雜訊干擾情況的語音及非語音機密分佈，平均值及變異數已經都已補償的極為相似，語音及非語音也不再干擾彼此的順序關係，減少了彼此間不匹配性，相信可以提高辨識的效能。

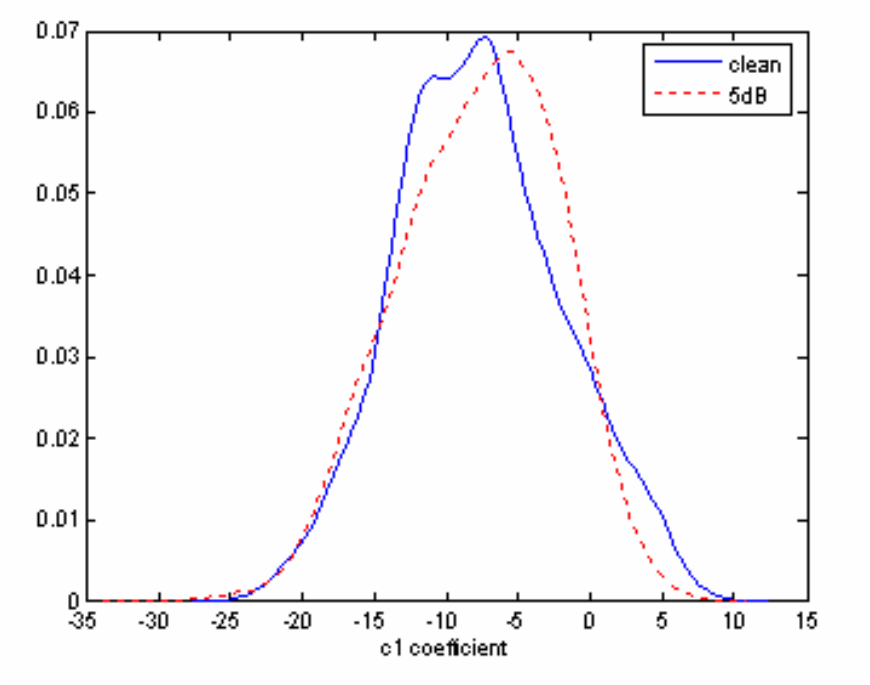

圖 4-6:語音經分群倒頻譜正規化法在乾淨與訊噪比 5dB 下的分佈

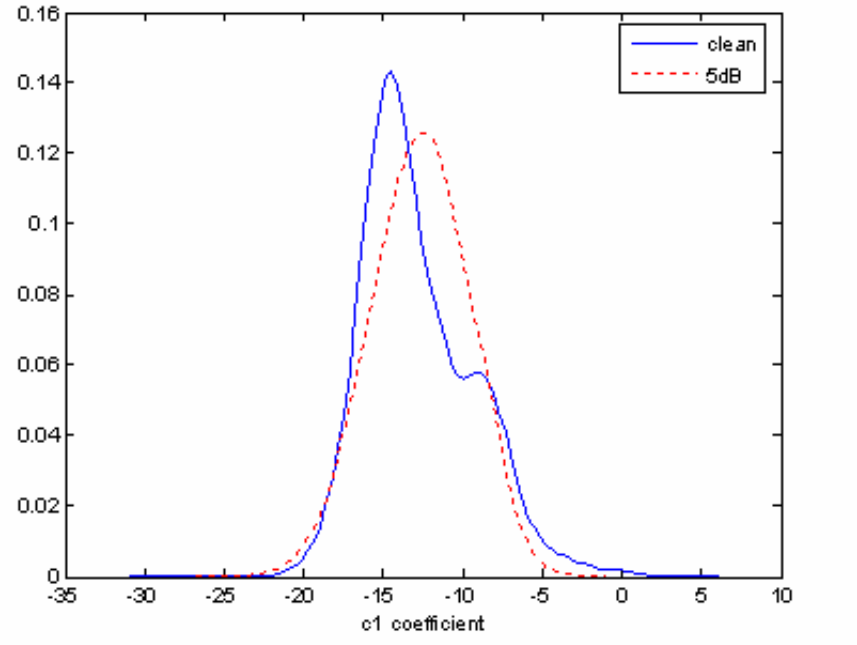

圖 4-7 非語音經分群倒頻譜正規化法在乾淨與訊噪比 5dB 下的分佈

最後，我們將經傳統式倒頻譜正規化法及分群倒頻譜正規化法的結果來做個比較，如圖 4-8 及圖 4-9，分別為傳統式倒頻譜正規化法及分群倒頻譜正規化法其語音及非語音在乾淨及受雜訊干擾情況下的機密分佈圖，可以觀察到，經分群倒頻譜正規化法後的機率分佈確實是比傳統式倒頻譜正規化法效果來的好，受雜訊干擾的語音及非語音參數值順序，回復與乾淨一致，並不像傳統倒頻譜正規化法一樣，參數值會互相干擾順序，因此辨識效能也應該會相對的較佳。

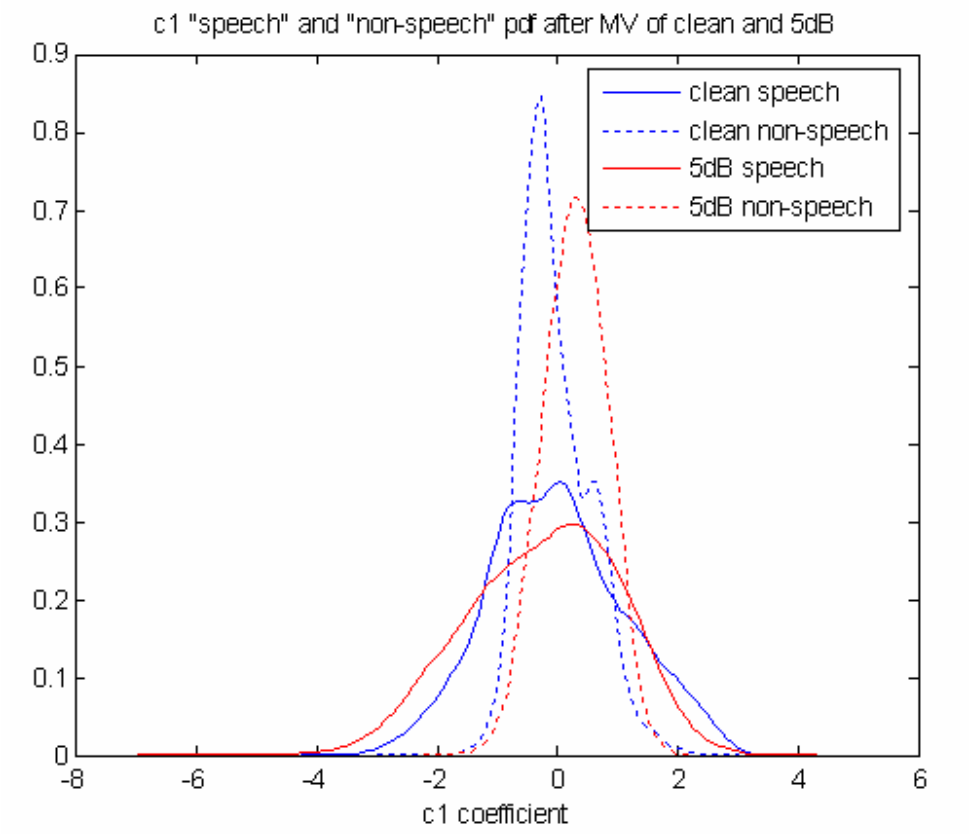

圖 4-8:經傳統倒頻譜正規化法之語音及非語音在乾淨與訊噪比 5dB 下的分佈

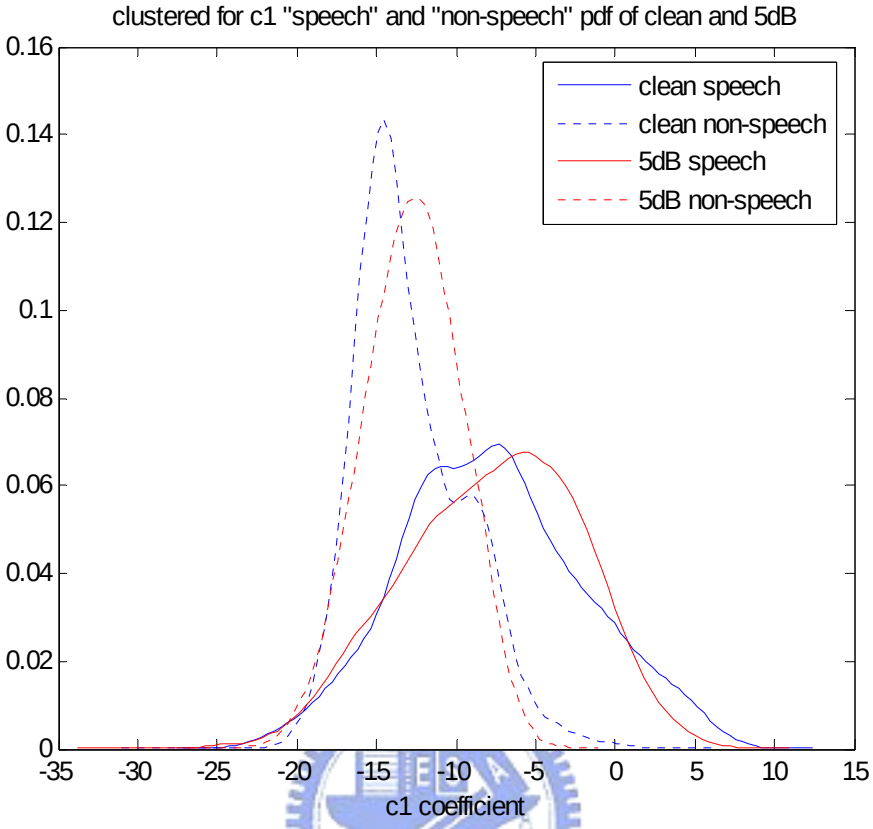

圖 4-9:經分群倒頻譜正規化法之語音及非語音在乾淨與訊噪比 5dB 下的分佈

4.2.3 實驗設定與流程

本實驗所用的語料庫、語音特徵參數的擷取、聲學模型及相關實驗設定，與基礎系統相同。而系統的流程如圖 4-10 所示，首先利用基礎系統 MVA 的聲學模型將訓練語料特徵參數做強迫對齊(Forced Alignment)，並對測試語料特徵參數做辨識(Recognize)，來得到語料內容的切割位置，利用這些切割位置來將特徵參數分為語音及非語音兩類，之後再統計分群式倒頻譜正規化法所需的語音及非語音特徵參數的平均值與變異數，利用這些資料及統計值來做分群式倒頻譜正規化法，最後為了能進一步提升辨識效能，我們將經過分群式倒頻譜正規化法的特徵參數，再經過 ARMA 濾波器，來消除雜訊在參數上所造成不屬於人聲的高頻成分，整個系統，簡稱為分群式(Clustered) MVA 系統。

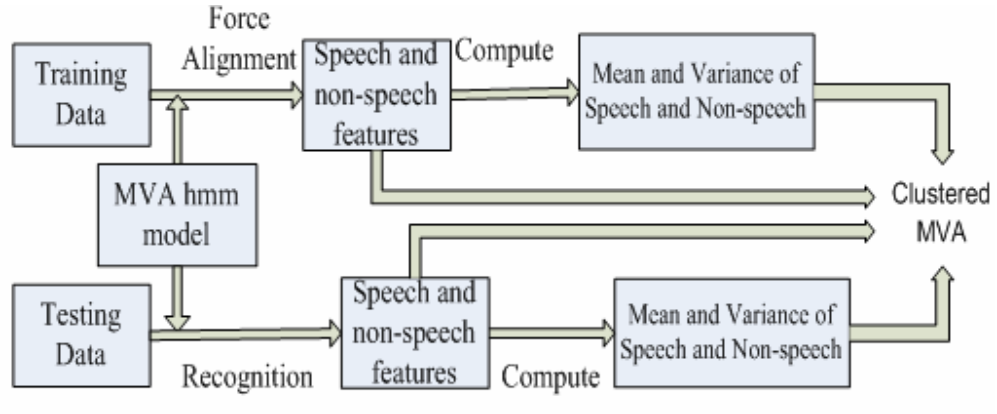

圖 4-10:分群倒頻譜正規化法實驗流程

4.2.4 實驗結果

表 4-1 是分群式 MVA 系統的國語連續數字串辨認實驗中的乾淨語音訓練模式辨識結果。

表 4-1:分群式 MVA 系統的國語連續數字串辨認實驗中的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	98.75%				
20	96.32	96.98	98.04	96.13	96.86
15	94.08	95.19	97	92.68	94.73
10	86.6	87.82	90.92	82.83	87.04
5	71.18	74.65	82.87	65.16	73.46
0	44.87	49.75	58.16	42.09	48.71
-5	22.64	24.6	29.53	18.89	23.91
平均值(20dB～0dB)	78.61	80.87	85.39	75.77	80.16
訊噪比	B 組				

(dB)	餐廳	街道	機場	火車站	平均值
乾淨	98.75%				
20	97.82	98.13	97.2	98.05	97.8
15	93.42	94.97	95.59	97.62	95.4
10	87.94	88.32	90.65	94.2	90.27
5	65.89	73.34	80.19	83.15	75.64
0	43.57	50	56.57	62.44	53.14
-5	22.76	26.48	27.71	34.86	27.95
平均值(20dB～0dB)	77.72	80.95	84.04	87.09	82.45
八種環境雜訊及五種訊噪比的平均值					81.3%

4.2.5 實驗結果討論

如圖 4-11 可觀察到，在訊噪比 10dB 以上，分群式 MVA 辨識率比起其它兩個系統都還好，證實分群式 MVA 確實有效，但在低訊噪比(5dB 以下)的情況下，辨識率卻表現較差，原因可能是因為在低訊噪比的情況下，分群的效果不佳而造成辨識率下降，若能改善，相信辨識率應該會有所提升。

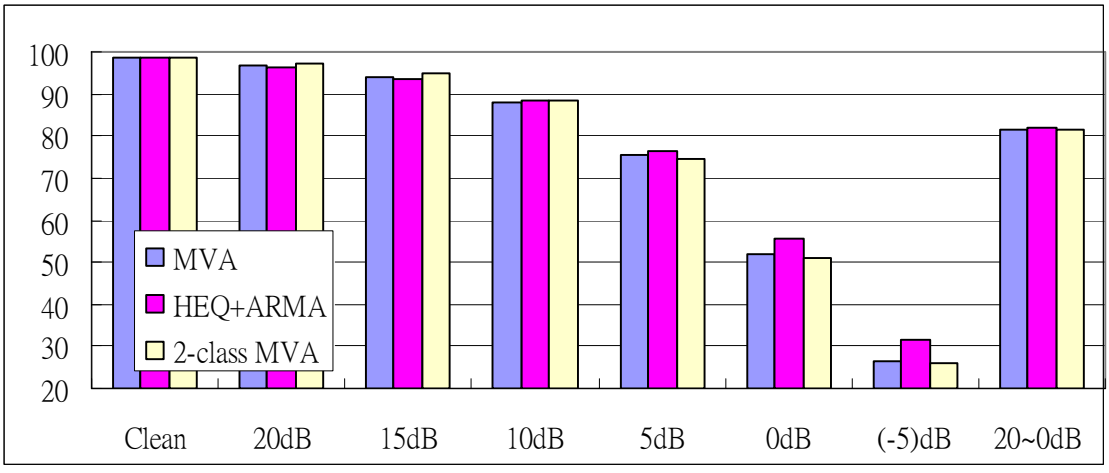

圖 4-11:MVA、HEQ+ARMA、分群式 MVA 在各種訊噪比的辨識率比較圖

4.3 改良式分群倒頻譜正規化法

4.3.1 分群倒頻譜正規化法的潛在問題

為了改善原始分群式倒頻譜正規化在低訊噪比辨識率下降的問題，如表 4-2、4-3 所示，我們統計「人聲」雜訊測試語料及所有訓練語料經由辨識及強迫對齊所求出來的切割資訊，發覺「語音」所佔的比例會因訊噪比不同而有所變化，當在低訊噪比的情況下，音框判斷為語音的音框數目下降許多，代表因雜訊的影響，使得許多「語音」的音框被判斷成「非語音」音框，使「語音」音框所佔的百分比下降許多，而當「語音」被判斷為「非語音」音框時，我們知道將影響分群式倒頻譜正規化法辨識效能，因此為了改善此問題，我們提出了改良式分群式 MVA 系統。

表 4-2: 「人聲」雜訊測試語料經由辨識所求出來的切割資訊

「人聲」雜訊測試語料		
	語音的音框	百分比
Clean	20604	66.37 %
20dB	20559	66.22 %
15dB	20443	65.82 %
10dB	20273	65.30 %
5dB	19363	62.37 %
0dB	18992	61.17 %
-5dB	19888	64.06 %

表 4-3: 訓練語料經由強迫對齊(force alignment)所求出來的切割資訊

訓練語料		
	音框	百分比
語音	198693	65.164 %
非語音	106219	34.836 %

4.3.2 改良式分群式 MVA 系統

為了改善在低噪比下，「語音」音框比例下降的問題，我們將辨識為「非語音」的機率降低，來改善因雜訊影響所造成「非語音」比例上昇的問題，而調整的方式為，當聲學模型的轉移機率都為一致時，我們將「語音」轉移至「非語音」聲學模型的轉移機率調整略低，來降低「非語音」音框的比例。

4.3.3 實驗結果

表 4-4 是改良式分群 MVA 系統的國語連續數字串辨認實驗中的乾淨語音訓練模式辨識結果。

表 4-4: 改良式分群 MVA 在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	98.44%				
20	96.32	97.14	98.04	96.45	96.98
15	94.55	95.19	97	93.47	95.05
10	87.4	88.3	90.92	84.52	87.78
5	71.82	76.85	83.8	72.98	76.36
0	54.07	48.71	61.9	45.1	52.44
-5	24.55	23.44	31.31	19.86	24.79
平均值(20dB～0dB)	80.83	81.23	86.33	78.50	81.72
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	98.44%				
20	97.82	97.98	96.88	98.05	97.68
15	92.94	95.13	94.96	97.78	95.20
10	84.41	88.32	90.34	94.14	89.30
5	67.86	77.97	79.87	84.1	77.45
0	45.33	51.09	56.1	64.6	54.28
-5	23.38	28.64	28.57	36.1	29.17
平均值(20dB～0dB)	77.67	82.09	83.63	87.73	82.78

八種環境雜訊及五種訊噪比的平均值	82.26%

4.3.4 實驗結果討論

由圖 4-12 可以比較改良式分群 MVA 與先前所提出的分群式 MVA，可以觀察到，只有在低訊噪比的情況下，辨識率才有較多的提升，可能因為在低訊噪比的情況下，「語音」的音框數才會明顯下降，改良式的 MVA 才會因此發揮較多的效果，

而改良式分群 MVA 相對於 MVA 而言，每一訊噪比及整體的辨識率也都有所提升。

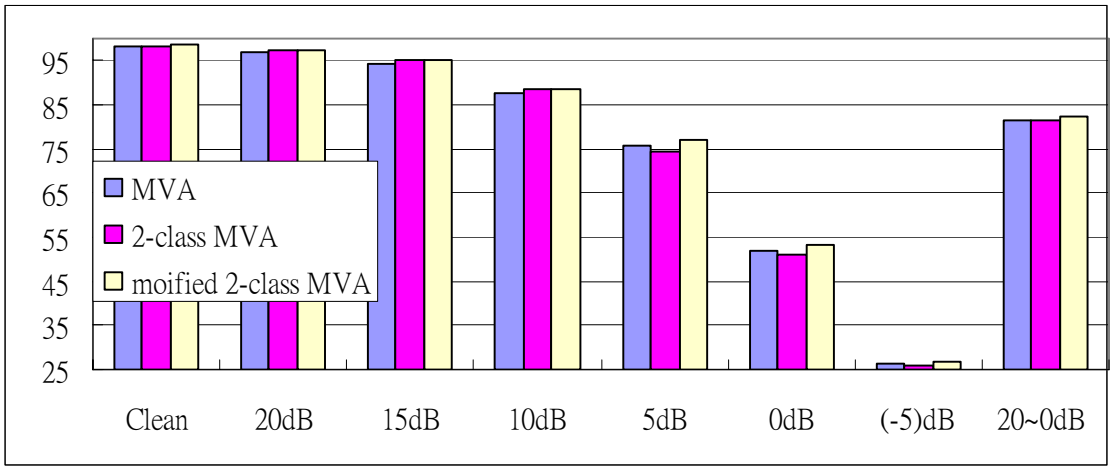

圖 4-12:MVA、分群式 MVA、改良式分群 MVA 在各種訊噪比下的辨識率比較圖

4.4 理想分群 MVA 系統

4.4.1 實驗目的與設定

證明若能準確分群，分群 MVA 系統是能夠有效的抵抗環境雜訊的干擾，並且增加辨識率，而分群的方式為先對未加上環境雜訊之測試語料進行強迫對齊(force alignment)，利用切割的結果來判斷「語音」或「非語音」。

4.4.2 實驗結果

表 4-5 是理想分群 MVA 系統的國語連續數字串辨認實驗中的乾淨語音訓練模式辨識結果。

表 4-5: 理想分群 MVA 在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	98.75%				
20	96.55	97.33	97.8	95.91	96.89
15	93.87	95.55	96.23	93.71	95.81
10	87.58	89.31	91.19	85.59	88.16
5	73.75	78.93	85.91	70.6	77.05
0	52.36	55.72	65.31	50.79	55.55
-5	29.2	27.21	37.12	24.17	29.42
平均值(20dB～0dB)	80.81	83.16	86.88	79.12	82.59
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	98.75%				
20	97.8	97.8	97.65	97.96	97.8
15	95.65	95.55	96.23	97.58	95.95
10	86.32	89.15	91.98	95.5	90.58
5	71.55	79.09	83.33	85.69	79.91
0	52.83	57.86	60.53	67.55	59.66

-5	28.57	37.69	37.28	52.51	39.01
平均值(20dB～0dB)	80.62	83.86	85.95	88.61	85.76
八種環境雜訊及五種訊噪比的平均值					83.63%

4.4.3 實驗結果討論

由圖 4-13 可以觀察到，理想分群 MVA 系統，確實可以有效提升辨識率，特別是在低訊噪比的環境下，而理想分群 MVA 系統在 0dB 及-5dB 的辨識率明顯比改良式 MVA 系統高出許多，代表改良式 MVA 系統還有些許的進步空間，是我們未來可以努力的研究方向。

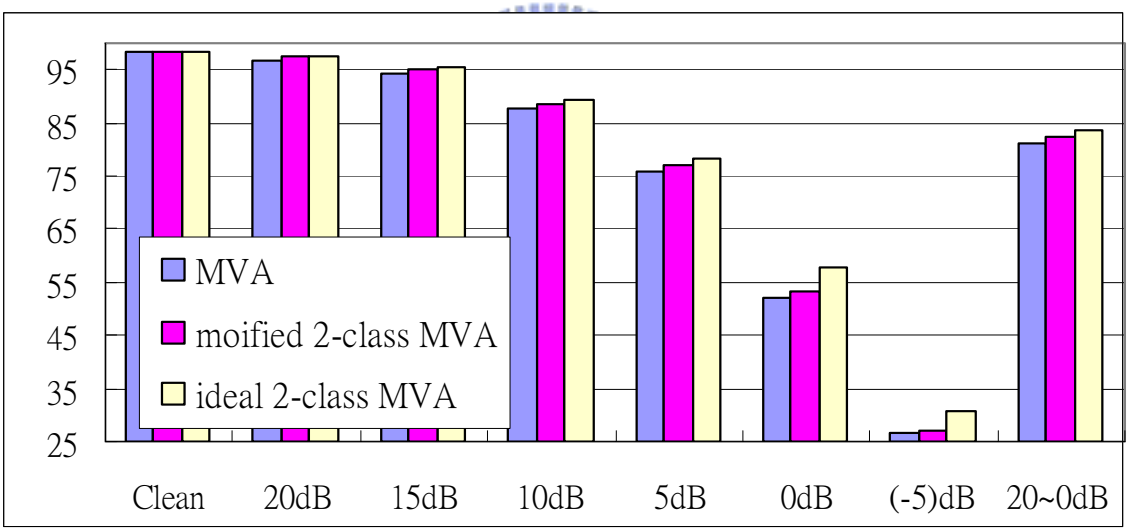

圖 4-13:MVA、改良式分群式 MVA、理想分群 MVA 在各種訊噪比下的辨識率比較圖

4.5 理想三群式 MVA 系統

4.5.1 實驗目的

由之前的實驗結果可得知，若能將「語音」及「非語音」準確分群，分群式 MVA 卻實可有效提升辨識率，為了能更進一步提升辨識效能，我們利用基頻(pitch)進一步將測試語料中「語音」的部分，分為 voiced 及 un-voiced，觀察它們在乾淨及受雜訊干擾下的機率分佈，圖 4-14 及 4-15 分別為測試語料中「語音」的 voiced 及 un-voiced 語音在乾淨及訊噪比 5dB 情況下的機率分佈，可以觀察到整個分佈的平均值偏移非常多，代表 un-voiced 語音的參數值嚴重的受雜訊影響，而使得與 voiced 語音參數值的順序(order)關係被破壞掉了，將可能造成「語音」部分做倒頻譜正規化法的效果不佳，因此我們進一步將語料「語音」中的 voiced 及 un-voiced 語音，還有「非語音」這三類分別做倒頻譜正規化法，其數學表示式如下

$$\hat{X} = \mu_{s\_voice,x} + (Y - \mu_{s\_voice,y})(\frac{\sigma_{s\_voice,x}}{\sigma_{s\_voice,y}})^{1/2} \quad (4\text{-}4)$$

$$\hat{X} = \mu_{s\_un-voice,x} + (Y - \mu_{s\_un-voice,y})(\frac{\sigma_{s\_un-voice,x}}{\sigma_{s\_un-voice,y}})^{1/2} \quad (4\text{-}5)$$

$$\hat{X} = \mu_{n,x} + (Y - \mu_{n,y})(\frac{\sigma_{n,x}}{\sigma_{n,y}})^{1/2} \quad (4\text{-}6)$$

$\mu_{s\_voice,x}$、$\mu_{s\_un-voice,x}$、$\sigma_{s\_voice,x}$、$\sigma_{s\_un-voice,x}$ 為訓練語料中「語音」有基頻及無基頻部分所統出來的平均值及變異數，$\mu_{s\_voice,y}$、$\mu_{s\_un-voice,y}$、$\sigma_{s\_voice,y}$、$\sigma_{s\_un-voice,y}$ 為測試語料中「語音」有基頻及無基頻部分所統出來的平均值及變異數，$\mu_{n,x}$、$\mu_{n,y}$、$\sigma_{n,x}$、$\sigma_{n,y}$ 為訓練語料及測試語料中「非語音」部分所統出來的平均值及變異數。

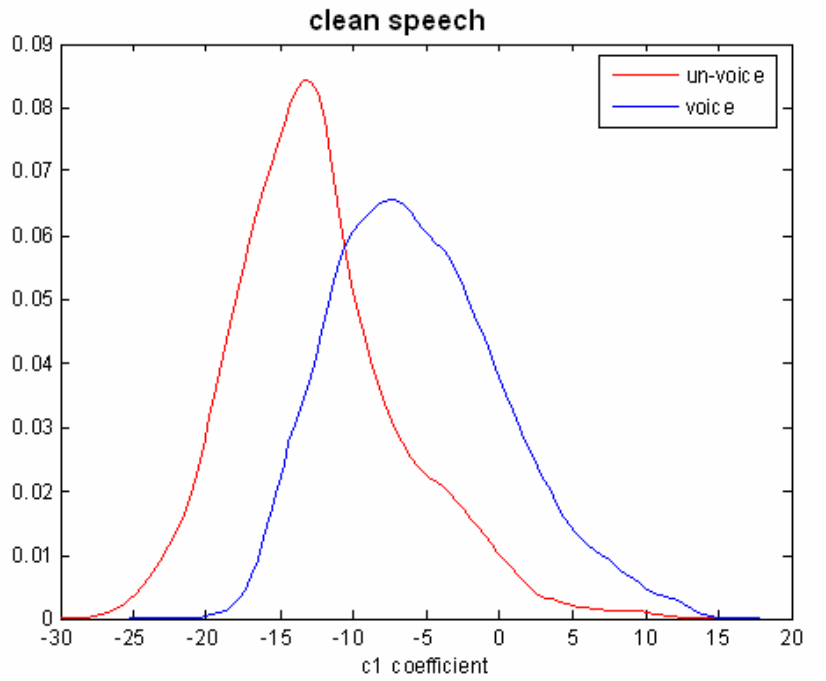

圖 4-14：測試語料語音的 voice 及 un-voice 在乾淨情況下的機率分佈

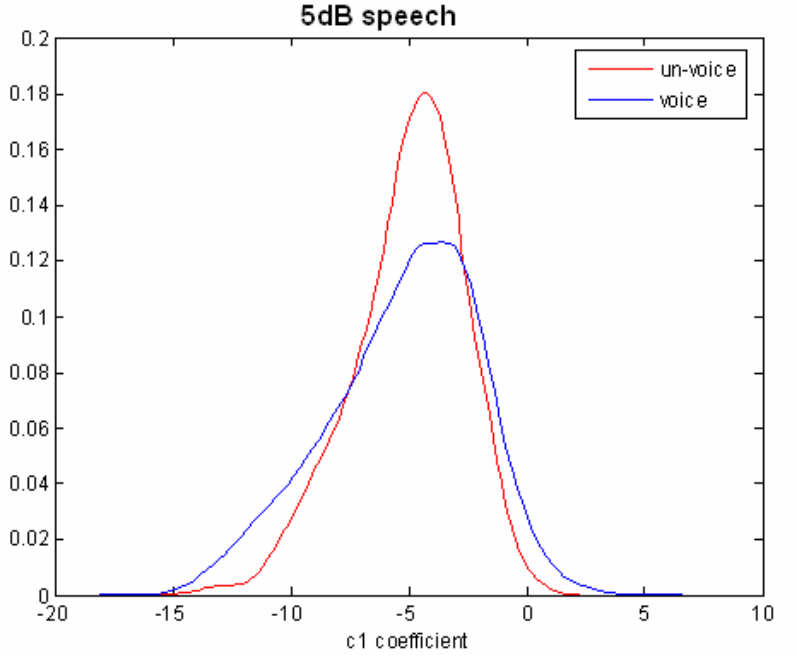

圖 4-15：測試語料語音的 voiced 及 un-voiced 語音在訊噪比 5dB 情況下的機率分佈

4.5.2 實驗設定

分群的方式為先對未加上環境雜訊之測試語料進行強迫對齊 (force alignment)，利用切割的結果來判斷「語音」或「非語音」，再利用ESPS(Entropic Corp.)來求得未加上環境雜訊測試語料的基頻(pitch)資訊，利用是否有基頻(pitch)來判斷「語音」為voiced或者是un-voiced。

4.5.3 實驗結果

表 4-6 是理想三群式 MVA 系統的國語連續數字串辨認實驗中的乾淨語音訓練模式辨識結果。

表 4-6:理想三群 MVA 在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	98.75%				
20	96.91	97.17	97.33	95.91	96.83
15	93.71	95.55	96.23	93.17	94.66
10	87.89	91.19	93.24	82.23	88.63
5	77.52	82.23	86.16	70.6	79.12
0	58.96	66.67	70.91	56.29	63.20
-5	38.99	42.45	49.06	34.28	41.19
平均值(20dB～0dB)	82.99	86.56	88.77	79.64	84.49
訊噪比 (dB)	B 組				
	餐廳	街道	機場	火車站	平均值
乾淨	98.75%				
20	97.33	97.65	97.8	97.96	97.68
15	95.65	96.23	97.58	96.63	96.52
10	87.89	90.88	92.61	94.65	91.50
5	79.87	83.81	87.42	88.05	84.78

0	61.95	66.35	72.48	77.36	69.53
-5	39.62	45.75	52.2	58.96	49.13
平均值(20dB～0dB)	84.53	86.98	89.57	90.93	88.00
八種環境雜訊及五種訊噪比的平均值					86.25%

4.5.4 實驗結果討論

圖 4-16 分別為 MVA、分佈等化法加 ARMA 濾波器、理想兩群式 MVA、理想三群式 MVA 在各訊噪比辨識率比較圖，可觀察到，若能夠準確的分出三群分別來做 MVA，在低訊噪比的辨識率將大幅提升，而整體的平均辨識率也提升許多，效果明顯比理想兩群式 MVA 還要好上許多，代表 voiced 及 un-voiced 參數值順序(order)，較容易受雜訊干擾，且嚴重影響辨識率好壞，因此若能有效改善此問題，將大幅提升辨識率，但實際上要準確的分出三群，是有困難性的，是未來可研究的方向之一。

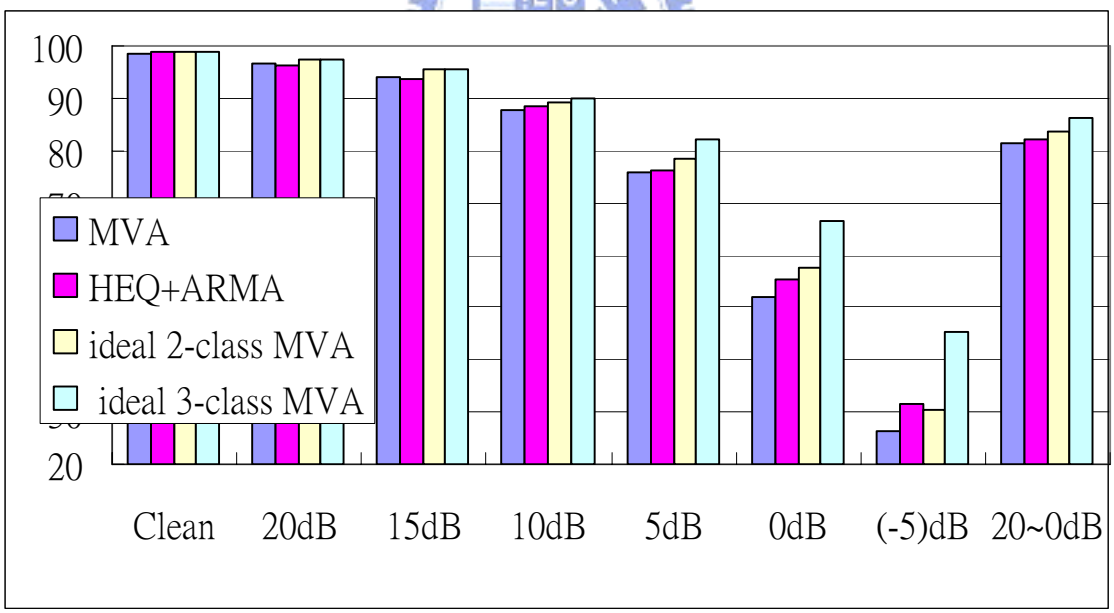

圖 4-16:MVA、HEQ+ARMA、理想二群式 MVA、理想三群式 MVA 各訊噪比辨識率比較圖

第五章 結論與未來展望

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。

5.1 結論

本論文主要的研究內容是針對強健性語音特徵參數作深入的探討，將現有的倒頻譜正規化法及分佈等化法做些許的改進。包括將分佈等化法加上 ARMA 濾波器，經由實驗結果證實，辨識率在每一種訊噪比的環境下都有所提升，特別是在低訊噪比的情況下;而我們所提出的另一個分群式 MVA 系統，在訊噪比 10dB 以上，辨識率都相對於 MVA 系統而言，都有所提升，但在訊噪比 5dB 以下，因分群效果不佳使得辨識率下降，因此我們提出了改良式分群 MVA 系統，進一步的改善了分群式 MVA 在訊噪比 5dB 以下「語音」音框百分比明顯下降的問題，也因此提升了訊噪比 5dB 以下的辨識率，使得每一訊噪比的辨識率，都比 MVA 系統來的好，代表我們所提出的改良式分群 MVA 系統，確實比 MVA 系統來的好，且可以實現的，我們也做了理想分群 MVA 系統實驗，讓我們知道分群 MVA 系統辨識率的上限，以及還有多少改進的空間。

最後，我們進一步的利用正確的基頻(pitch)將「語音」分為 voiced 及 unvoiced，將 voiced、unvoiced 語音及「非語音」分別作倒頻譜正規化，也就是理想三群式 MVA 系統，與理想兩群式 MVA 系統比較，發覺理想三群式 MVA 系統辨識率提升更多，特別是在低噪訊比的情況下，代表若能準確的分出三群，辨識率將提升更多。

5.2 未來展望

本論文的改良式分群 MVA 系統，在低噪訊比的辨識率，要達到上限還有一段

距離，代表若能進一步的改善低訊噪比「語音」及「非語音」的分群方法，像是利用頻譜熵值(Entropy)[17]或是利用特徵空間旋轉法(Feature Space Rotation)[18]等相關的方法來幫忙分群，相信辨識率還會再提升。

而在有雜訊的情況下如何將「語音」的基頻求出，好讓我們利用基頻來判斷是 voiced 還是 unvoiced，讓我們理想三群式 MVA 系統可以變成實際的系統，來對抗雜訊對辨識率的影響，也是另外一個值得研究的方向。

最後，本論文的語料庫都為中文連續數字串語料庫，字彙量較少，未來應將這些系統應用在大字彙辨識的工作上，以觀察字彙數量大小是否影響本論文中所提到多種系統對強健性的提升。

參考文獻

[1]. Y. Gong, “Speech Recognition in Noisy Environments: A Survey” , Speech Communication. 16,1995.

[2]. A. E. Rosenberg, C.-H Lee, and F. k. Soong, “Cepstral Channel Normalization Techniques for HMM-based Speaker Verification” , ICSLP, 1992.

[3]. O. Vikki and K. Laurila, “Noise Robust HMM-based Speech Recognition Using Segmental Cepstral Feature Normalization” , in ESCA NATO Workshop Robust Speech Recognition Unknown Communication Channels. France , 1997.

[4]. A. de la Torre, J. C. Segura, C. BENitez, A. M. Peinado, and A. J. Rubio, “Non-linear Transformations of the Feature Space for Robust Speech Recognition” , ICASSP, 2002.

[5]. A. de la Torre, A. M.Peinado, J. C. Segura, J. L. P. Cordoba, M. C. Benitez and A. J. Rubio, “Histogram equalization of speech recognition for robust speech recognition” , IEEE Trans. On Speech and Audio Processing, vol. 13, no. 3, May 2005,pp.355-366

[6]. Chia-Ping Chen, Jeff Bilmes, and D. Ellis, "Speech Feature Smoothing for Robust ASR", Proceedings of ICASSP 2005 pp.525-528

[7]. Chia-Ping Chen, Jeff Bilmes, and Katrin Kirchhoff, "Low-Resource Noise-Robust Feature Post-Processing on Aurora 2.0", Proceedings of ICSLP 2002 pp.2445-2448

[8]. 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.

[9]. JC Segura, C. Benítez, A. de la Torre, AJ Rubio and J. Ramírez, Cepstral Domain Segmental Nonlinear Feature Transformations for Robust Speech Recognition, IEEE Signal Processing Letters, 11(5), May 2004.

[10]. Shang-nien Tsai, Lin-shan Lee."A New Feature Extraction Front-end for Robust Speech Recognition using Progressive Histogram Equalization and Multi-Eigenvector Temporal Filtering" , ICSLP 2004

[11]. Shang-nien Tsai and Lin-shan Lee, **"**Improved Robust Features for Speech Recognition by Integrating Time-Frequency Principal Components (TFPC) and Histogram Equalization(HEQ)**,"** *IEEE 8th Automatic Speech Recognition and Understanding Workshop*, PP.297-302, St. Thomas, US Virgin Islands, USA, Dec. 2003.

[12]. Yi Chen, Lin-shan Lee, **"**Robust Features for Speech Recognition Using Minimum Variance Distortionless Response (MVDR) Spectrum Estimation and Feature Normalization Techniques**,"** *International Symposium on Chinese Spoken Language Processing*, PP.101-104, Hong Kong, Dec. 2004.

[13]. Y. Obuchi and RM Stern,"Normalization of Time-Derivative Parameters Using Histogram Equalization", Eurospeech 2003.

[14]. J.C.Sequra,C. Benitez , A. de la Torre, and A.Rubio, "Feature extraction combining spectral noise reduction and cepstral histogram equalization for robust ASR" ,ISCLP 2002.

[15]. Hans-Giinter 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.

[16].魯柏暄，”使用基頻資訊之國語分散式語音辨識系統”，交通大學碩士論文，2005 .

[17].Jia-lin Shen , Jeih-weih Hung ,Lin-shan Lee ;”Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environment”, *International Conference on Spoken Language Processing*, Sydney, Nov. 1998.

[18].Sirko Molau, Daniel Keysers, And Hermann Ney, “ Matching Training and Test Data Distributions for Robust Speech Recognition” , Speech Communication 41 , 579-601, ELSEVIER 2003.