國立臺灣師範大學
資訊教育研究所碩士論文

指導教授：　葉耀明　博士

陳柏琳　博士

數據擬合與分群方法於強健語音特徵擷取之研究

Exploring the Use of Data Fitting and Clustering Techniques for Robust Speech Recognition

研究生：　林士翔　撰

中華民國　九十六　年　七　月

摘　　要

語音長久以來一直是人類最自然且最容易使用的溝通媒介。無庸至疑地，語音也勢必會扮演著未來人類與各種智慧型電子設備間最主要的人機互動媒介，因此自動語音辨識(Automatic Speech Recognition, ASR)技術將會是扮演其中最關鍵且重要的角色。目前大部份的自動語音辨識系統在語音訊號不受干擾的理想乾淨實驗室環境下，可獲得非常不錯的辨識效果；但若應用至現實環境中，語音辨識率卻往往會因為環境中複雜因素的影響，造成訓練環境與測試環境存在的不匹配(Mismatch)的問題存在，使得系統辨識效能大幅度地降低。因此，語音強健(Robustness)技術就顯得格外重要與受到重視。

目前有關語音強健方法的研究若以其處理對象而言，大致上可從二種不同層面討論：從語音特徵值本身為出發，或是從統計分布出發，此二類研究各有其優缺點。本論文嘗試結合上述二種層面的優點，並且利用數據擬合(Data-fitting)技術來增進語音辨識系統的辨識效能。吾人首先提出了群集式為基礎之多項式擬合統計圖法(Cluster-based Polynomial-fit Histogram Equalization, CPHEQ)，利用統計圖等化法(Histogram Equalization)的概念與雙聲源訓練語料(Stereo Training Speech Data)的使用求得多項式轉換函數。再者，吾人將此方法做一些假設及延伸，進而衍生出二種不同方法，其一是以多項式擬合統計圖等化法(Polynomial-fit Histogram Equalization, PHEQ)來改良傳統統計圖等化法需要耗費較多記憶體空間與處理器運算時間的缺點；另一個則是配合遺失特徵理論(Missing Feature Theorem)的選擇性群集式為基礎之多項式擬合統計圖等化法(Selective Cluster-based Polynomial-fit Histogram Equalization, SCPHEQ)來進行語音特徵參數的重建。語音辨識實驗是以 Aurora-2 語料庫為研究題材；實驗結果顯示，在乾淨語料訓練模式下，吾人所提出的方法相較於基礎實驗結果能顯著地降低詞錯誤率，並且其成效也較其它傳統語音強健方法來的好。

Abstract

Speech is the primary and the most convenient means of communication between individuals. It is also expected that automatic speech recognition (ASR) will play a more active role and will serve as the major human-machine interface for the interaction between people and different kinds of intelligent electronic devices in the near future. Most of the current state-of-the-art ASR systems can achieve quite high recognition performance levels in controlled laboratory environments. However, as the systems are moved out of the laboratory environments and deployed into real-world applications, the performance of the systems often degrade dramatically due to the reason that varying environmental effects will lead to a mismatch between the acoustic conditions of the training and test speech data. Therefore, robustness techniques have received great importance and attention in recent years.

Robustness techniques in general fall into two aspects according to whether the methods' orientation is either from feature domain or from their corresponding probability distributions. Methods of each have their own superiority and limitations. In this thesis, several attempts were made to integrate these two distinguishing information to improve the current speech robustness methods by using a novel data-fitting scheme. Firstly, cluster-based polynomial-fit histogram equalization (CPHEQ), based on histogram equalization and polynomial regression, was proposed to directly characterize the relationship between the speech feature vectors and their corresponding probability distributions by utilizing stereo speech training data. Moreover, we extended the idea of CPHEQ with some elaborate assumptions, and two different methods were derived as well, namely, polynomial-fit histogram equalization (PHEQ) and selective cluster-based polynomial-fit histogram equalization (SCPHEQ). PHEQ uses polynomial regression to efficiently approximate

the inverse of the cumulative density functions of speech feature vectors for HEQ. It can avoid the need of high computation cost and large disk storage consumption caused by traditional HEQ methods. SCPHEQ is based on the missing feature theory and use polynomial regression to reconstruct unreliable feature components. All experiments were carried out on the Aurora-2 database and task. Experimental results shown that for clean-condition training, our method achieved a considerable word error rate reduction over the baseline system and also significantly outperformed the other robustness methods.

誌謝

首先，要感謝父母親及家人的支持與陪伴，他們在我的求學過程中，總是毫無保留的支持與鼓勵我，尊重我所做的每一個決定，使得在此三年的碩士研究生涯中，能夠心無旁騖，潛心於知識的汲取與研究上。

再者，需感謝我的二位指導教授－葉耀明博士與陳柏琳博士，感謝老師們的諄諄教誨，讓學生在此三年中成長許多，不僅學問知識的獲得，更學到了對研究的態度與為人處事的道理。感謝葉耀明老師帶領我參與許多不同類型的研究與專案計畫，使我學習與了解到許多課堂上不會學到的事物，也感謝老師給予我許多的機會去嘗試與體驗先前從未接觸過的領域，讓我的見識能夠放得更廣與更遠；感謝陳柏琳老師總是不厭其煩的引導我研究方向，讓我學習到如何從不同的角度去思索與分析問題，且更進一步地解決問題，讓原本害怕數理統計的我，如今，亦能試著去解釋數學式背後的含意，也感謝老師讓我有數次機會出國參與國際研討會，增加自己的國際觀，且也能有機會與國外研究學者相互交流研究心得。同時，感謝口試委員王新民博士、陳永昇博士與洪志偉博士對於學生論文的愷切指正與許多寶貴意見，讓學生的論文能夠更臻完善。

同時，也要感謝實驗室的同學們，人瑋、志豪、成韋、士弘、怡婷、炫盛、燦輝、芳輝、鴻彬、庭瑋、斯涵、鴻欣、慶全、志彬、才業、智翔、鼎元及宜達，有你們的陪伴，使得研究生活變得不再那麼苦悶。在這段求學日子裡，我們曾經彼此相互學習、討論、砥礪與扶持過的日子，這所有的一切，都是美好的。

最後，要感謝亦純陪我渡過這段時間，感謝她的付出與體諒，讓我能有動力去迎接不同的挑戰。

謹將此論文獻給所有曾經幫助我的人。

士翔　謹誌

章節目錄

圖目錄

表目錄

第一章 序論

1.1 研究背景

語音長久以來一直是人類最自然且最容易使用的溝通媒介，無庸至疑地，語音勢必會扮演著未來人類與各種智慧型電子設備間最主要的人機互動媒介，因此自動語音辨識(Automatic Speech Recognition, ASR)技術將會是扮演關鍵且重要的角色[Juang and Frui 2000; Lee and Chen 2005]。自動語音辨識的挑戰在於給定一段語音訊號，如何讓電腦能夠快速且正確的辨識出語音訊號的內容，並將其轉換成一連串詞序列(Word Sequence)。然而現今語音辨識系統的辨識效能，根據語音辨識任務的複雜度不同，辨識效能亦有所差距，例如語者相關(Speaker-Dependent)的小詞彙獨立字詞辨識(Small Vocabulary Isolated Speech Recognition, SVISR)的辨識系統，若使用者能處在一個語音訊號不會受任何環境因素干擾的理想乾淨實驗室環境下，系統辨識錯誤率遠小於 10%；然而，若辨識任務是語者獨立(Speaker-Independent)的大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)，並且被使用在一個非常惡劣且吵雜的環境下，那麼辨識效能便會大幅度地下降。

目前已有少數的語音應用與商業化產品於日常生活當中，例如利用語音控制智慧型電子設備、銀行帳戶語音查詢系統或航空公司語音訂位系統等。此類系統的成功應用主要是來自於系統的辨識詞彙能力限制在數千或數百字詞以內，換句話說，現今的語音辨識系統在限定辨識詞彙個數的情況下，可獲得不錯的辨識效能。但是因為語言本身的複雜性，使得自動語音辨識要發展到實用性的階段，可以完全正確的辨識某一種語言的所有字彙，仍有許多尚待努力的空間。

語音強健(Robustness)技術便是其中一個需解決的關鍵問題，因為自動語音辨識系統在語音訊號不受干擾的理想乾淨實驗室環境下，可獲得非常不錯的辨識

效能，但若實際應用至日常生活環境中，卻往往因環境中複雜因素的影響，造成訓練環境與測試環境存在環境不匹配(Environmental Mismatch) 的差異，使得辨識系統的辨識效能大幅度降低。干擾的因素可從許多不同層面探討[Acero 1990]，例如語者發音結構差異、訊號輸入源的差異、加成性噪音(Additive Noise)、頻譜傾斜(Spectral Tilt)或其他語者干擾等，正因如此，語音強健技術長久以來一直被視為重要的研究課題[Gong 1995; Junqua et al. 1996; Huang et al. 2001]，主要希望藉由對語音訊號本身、語音特徵參數或是聲學模型參數做適當的處理，以減緩雜訊干擾的影響情形、降低訓練環境與測試環境不匹配的情形，或是加強語音訊號或語音特徵參數本身的強健性，進而提高辨識系統的辨識效能。

1.2 統計式語音辨識

早期的研究學者嘗試將人類的發音過程與聽覺感知等生理現象，利用一些規則(Rules)來表示，但是因為語言本身的高複雜性，並無法窮舉出所有的規則可能，所以發現此舉是不可行的，因此後來研究便慢慢朝向用機率統計模型等機器學習(Machine Learning)的方式發展。統計式語音辨識(Statistical Speech Recognition)的目的在於給定一串語音特徵向量序列(Observation Sequence) $Y = y_1, y_2, \cdots, y_T$ 時，我們希望從所有可能詞序列組合集 $\overline{\mathbf{W}}_{All}$ 中，找出一串詞序列 $W_{Best} = w_1, w_2, \cdots w_N$ 當作是辨識結果，因此數學關係式可以表式成如下[Huang et al. 2001]：

$$W_{Best} = \arg\max_{W \in \overline{\mathbf{W}}_{all}} P(W \mid Y) \qquad \text{(式 1-1)}$$

$P(W \mid Y)$ 代表語音特徵向量序列 Y 發生時，產生某一串詞序列 W 的事後機率(Posterior Probability)，若進一步使用貝氏定理(Bayes' Theorem)展開，上式可改寫成：

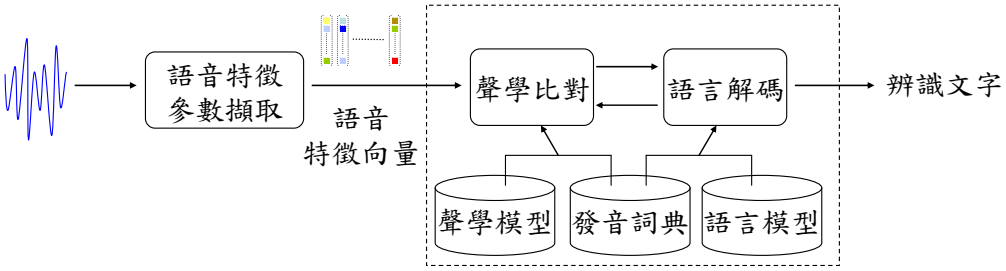


圖 1-1 統計式語音辨識流程

$$P(W \mid Y) = \frac{P(W)p(Y \mid W)}{p(Y)} \qquad \text{(式 1-2)}$$

其中 $p(Y \mid W)$ 表示詞序列 W 產生語音特徵向量序列 Y 的相似度(Likelihood)，一般會使用一些機率分布(Probability Distribution)模型來進行建模(Modeling)，因為此模型主要是用來決定與聲學相關的語音特徵向量序列 Y 所發生的機率，故稱其為聲學模型(Acoustic Model)； $P(W)$ 為詞序列 W 發生的事前機率(Prior Probability)，假設該詞序列 W 內含有 $w_1, w_2, \cdots w_N$ 個詞，則 $P(W)$ 的計算方式即等同計算該詞序列的聯合機率(Joint Probability) $P(w_1, w_2, \cdots, w_N)$，相同的，此聯合機率也會利用一些機率分布模型來描述，因為此模型是用來計算詞序列 W 發生的機率，所以又稱其為語言模型(Language Model)；而 $p(Y)$ 對找出最佳詞序列 W_{Best} 並不會有任何影響，故可加以省略。因此最後式 1-2 可改寫成：

$$W_{Best} = \arg\max_{W \in \overline{\mathbf{W}}_{all}} P(W)p(Y \mid W) \qquad \text{(式 1-3)}$$

目前常見的統計式語音辨識系統大致上都會包含三個部份，如圖 1-1 所示 [Huang et al. 2001]：

- 前端處理(Front-End Processing)：主要是將根據人類發音的特性，將語音訊號轉換成一連串的語音特徵向量序列。

- 聲學比對(Acoustic Matching)：聲學比對將已建立好的聲學模型，例如連續密度隱藏式馬可夫模型(Continuous Density Hidden Markov Models,

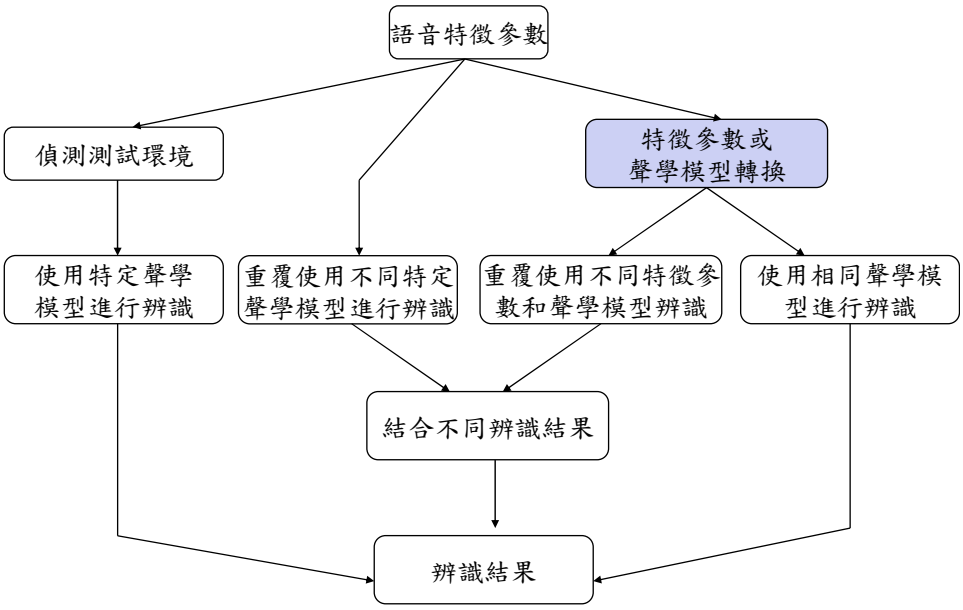


圖 1-2 解決訓練環境與測試環境不匹配的辨識方法

CDHMM)，與輸入語句中每一個可能是語音段落的語音特徵向量序列進行聲學比對，計算其發生的可能性。

- 搜尋解碼(Linguistic Decoding)：依據所有可能候選詞段落的聲學相似度與候選詞間的語言模型(如 *N* 連語言模型等)限制，進行解碼，找出機率最大的文句。

1.3 語音強健技術

為了解決環境不匹配問題，目前在語音強健技術研究上大致可分二個方向[Molau 2003]，第一個方向是針對不同測試環境，訓練出一組專用的聲學模型，另外一個研究方向是針對語音特徵參數或聲學模型做適當的處理，例如特徵參數轉換(Transformation)、增益(Enhancement)或聲學模型調適(Adaptation)等方法，以移除雜訊所造成訓練環境與測試環境不匹配的現象，此二種研究方向示意圖如圖 1-2 所示。

(1) 偵測測試環境，使用特定聲學模型進行辨識

主要是訓練各種不同測試環境情況下的聲學模型，例如特定語者(Speaker-dependent)、特定性別(Gender-dependent)、特定噪音(Noise-dependent)、或特定通道(Channel-dependent)等，然後再利用與測試環境相符的聲學模型進行辨識。但此方法必須有個前提假設，在進行辨識時，必須能夠線上即時決定辨識環境是什麼環境，進而決定要使用哪一組特定聲學模型進行辨識。若沒有辦法事先決定測試環境為何時，可利用各個聲學模型進行一次辨識，再將各種聲學模型所辨識出的辨識結果結合，找出最佳的辨識結果[Fiscus 1997]。此研究方向的優點是如果測試環境可以正確地被識別(Identified)，用與測試環境相匹配的聲學模型進行辨識，那麼必可得到良好的辨識效能；但缺點是因為現實環境中影響因素太多，並無法考慮到所有的可能情況，在進行辨識時，若無法找到相對應環境的聲學模型，那麼辨識效能還是無法提昇。

(2) 語音特徵參數或聲學模型轉換

在此類研究方向中，不論是在聲學模型訓練階段或語音辨識階段都只利用一組通用(Universal)聲學模型，主要有三類作法：

(I) 語音強化技術(Speech Enhancement)

語音強化技術目的在於提升語音訊號本身的品質，通常是假設語音訊號與雜訊訊號二者在統計上是不相關(Uncorrelated)，希望能由觀察到的雜訊語音(Noisy Speech)重建還原出原本的乾淨語音(Clean Speech)訊號，常見的技術有頻譜消去法(Spectral Subtraction, SS)[Boll 1979]、維爾濾波器(Wiener Filter, WF)[Huang et al. 2001]或卡爾曼濾波器(Kalman Filter)[Koo et al. 1989]等。

(II) 強健性語音特徵(Robust Speech Features)

主要是從語音訊號中擷取出較不易受到環境變化干擾而失真的強健性語音特徵參數或是對語音特徵參數進行補償，目前的研究議題包括語音特

徵參數轉換法(Feature Transformation)、語音特徵參數補償法(Feature Compensation)或語音特徵參數重建法(Feature Reconstruction)。此類作法為本論文研究重點，吾人將在後面章節詳述與探討近年來廣被使用並且能有效提昇語音辨識效能之方法。

(III) 聲學模型調適(Model adaptation)

藉由少量在測試環境中收錄的調適語料(Adaptation Data)來調整通用聲學模型中的機率分布參數，期望調適後的模型可以適用於新的環境，以降低環境不匹配的現象。常見的技術有最大事後機率法則(Maximum a Posteriori, MAP)[Gauvain and Lee 1994; Huo et al. 1995]、最大相似度線性回歸法(Maximum Likelihood Linear Regression, MLLR)[Leggeter and Woodland 1995; Gales 1998]及平行模型合併法(Parallel Model Combination, PMC)[Gales and Young 1995; 1996; Hung et al. 2002]等。

1.4 研究內容與貢獻

綜觀過去研究結果顯示，在測試環境調適語料與對應的正確轉譯文字(Reference Transcription)可獲得的情況下，前小節所述之三種研究方向以聲學模型調適可獲得較佳的辨識效能，因為藉由調適語料與對應的正確轉譯文字的使用，直接調整聲學模型參數，以降低由雜訊所產生環境不匹配的不確定性(Uncertainty)統計特性。但是此作法相較於其他作法，必需有額外在測試環境下所收錄的語音與正確轉譯文字的調適資料，且需花費額外相當的運算時間以進行聲學模型參數調整。另一方面而言，就語音強化技術而言，其主要目的是消除雜訊並改善語音的品質，例如提昇語音的訊噪比(Signal-to-Noise Ratio, SNR)，但此舉通常並不一定保證能可以提高語音辨識效能。相較於前述作法，強健性語音特徵不僅可有效地提昇辨識效能，而且通常只需額外短暫的運算時間即可完成。因此在本論文主要探討強健性語音特徵技術，首先介紹一些現有強健性語音特徵技術的作法，包括語

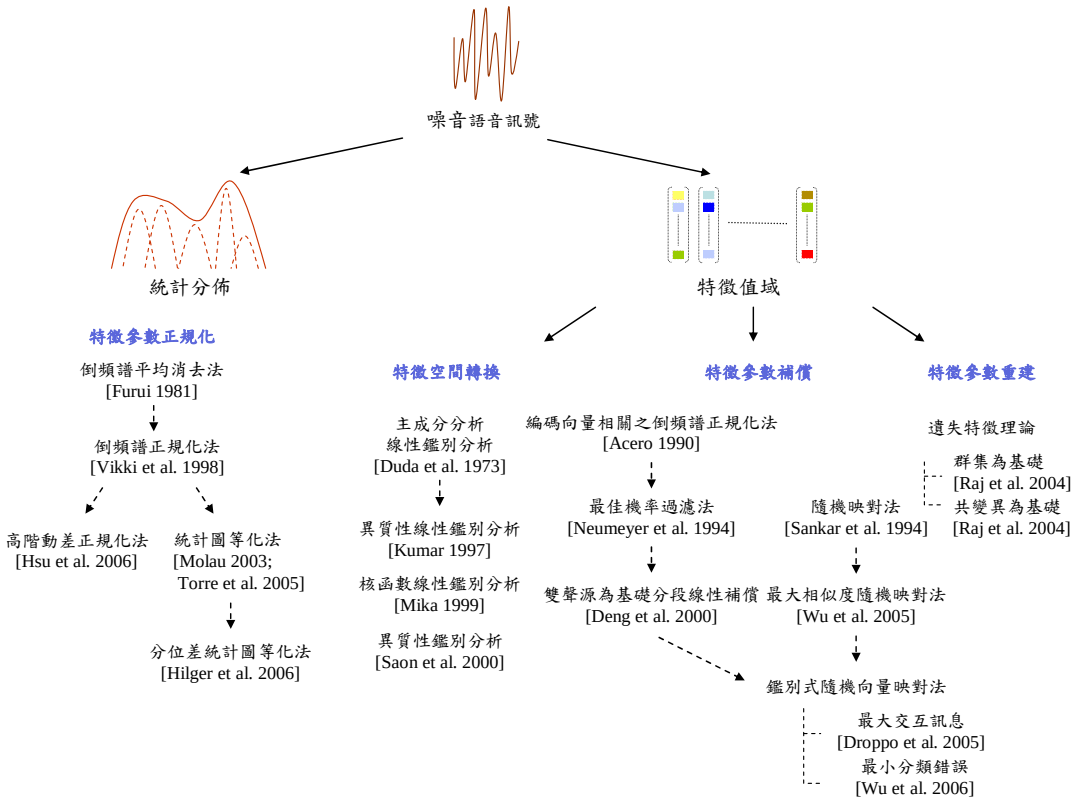


圖 1-3 強健性語音技術分類圖

音特徵參數轉換法、語音特徵參數補償法和語音特徵參數重建法。然而此三類研究法若依處理對象層面不同，又可再區分為以語音特徵值本身為出發或是考量語音特徵參數的統計分布為出發，分類圖如圖 1-3 所示，此二種以不同考量為出發點的強健性語音特徵技術有其優點，但卻也有所限制。

在本論文吾人嘗試結合上述二種研究方向的優點，且利用數據擬合(Data Fitting)技術來增進語音辨識系統的辨識效能。首先，吾人從最小均方誤差(Minimum Mean Square Error, MMSE)出發，改良現有雙聲源為基礎分段線性補償(Stereo-based Piecewise Linear Compensation, SPLICE)為基礎的語音特徵補償技術，嘗試利用多項式轉換函數取代傳統線性補償向量(Compensation Vector)，解決基礎分段線性補償只能做線性補償的缺點，此外吾人也利用較不易受雜訊乾擾的統計分布特性，進行語音特徵參數補償，而不是單單使用特徵參數本身的特徵值而已。再者，吾人將此方法做一些假設及延伸，進而衍生出二種不同作法，其

一是改良傳統統計圖等化法(Histogram Equalization)需耗費記憶體空間與處理器運算時間的缺點；另一個則是配合遺失特徵(Missing Feature) 理論，進行語音特徵參數的重建，實驗結果顯示吾人提出的方法，皆對提昇辨識系統的辨識效能有非常顯著的效果。

1.5 論文章節安排

本論後續共分五個章節，各章節編排如下：

第二章 首先先回顧語音特徵參數擷取流程並探討雜訊對語音訊號干擾情形，接著再回顧近年來較被廣泛及討論的語音強健技術，主要包含三個研究主軸：(一)語音特徵參數轉換法、(二)語音特徵參數補償法與(三)語音特徵參數重建法。

第三章 介紹本論文使用的實驗語料庫(Corpus)以及相關實驗設定與相關基礎實驗的實驗結果。

第四章 描述吾人所提出之群集式為基礎之多項式擬合統計圖等化法，並包括實驗參數的設定與相關實驗結果的討論。

第五章 描述由群集式為基礎之多項式擬合統計圖等化法所延伸推導出的二種方法，包括多項式擬合統計圖等化法與群集式為基礎之選擇性多項式擬合統計圖等化法，同時亦探討實驗參數的設定與相關實驗結果。

第六章 總結本論文的研究內容並探討未來可繼續研究之方向。

第二章 文獻回顧

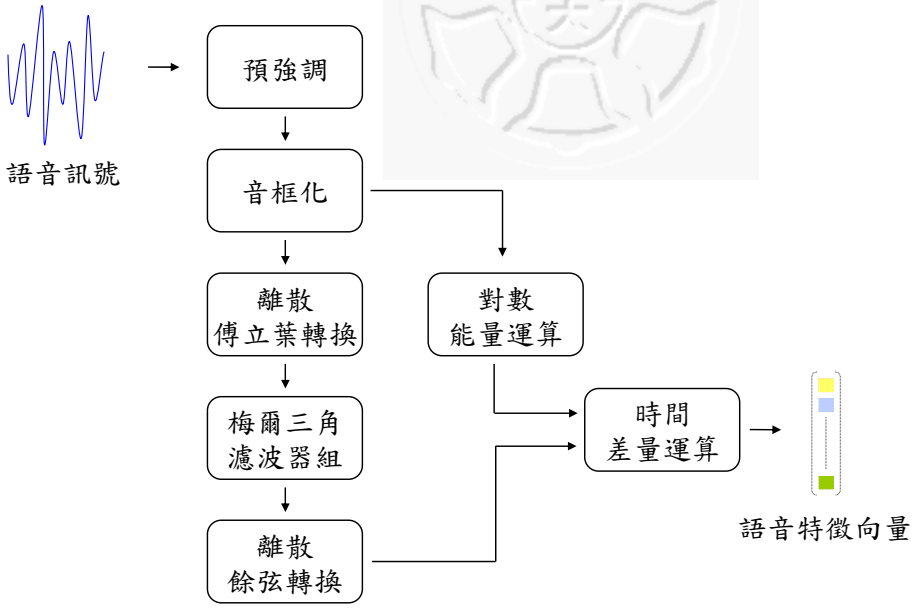


圖 2-1 梅爾倒頻譜係數特徵擷取流程圖

2.1 語音特徵參數擷取

語音特徵參數擷取主要的作用是將輸入的語音訊號轉換成一連串的語音特徵向量，同時也能達到降低維度(Dimension Reduction)的功用，以擷取出較能帶有聲學資訊的語音特徵向量。目前常見的技術大多是以考量人耳聽覺感知出發的梅爾倒頻譜係數(Mel-frequency Cepstral Coefficients, MFCC)[Davis and Mermelstein 1980]與感知線性預測係數(Perceptual Linear Prediction Coefficients, PLPC)[Hermansky 1991]，本論文所有實驗主要是採用梅爾倒頻譜係數為語音特徵參數為基礎。梅爾倒頻譜係數的擷取過程包含一系列步驟，有預強調(Pre-emphasis)、音框化(Windowing)、離散傅立葉轉換(Discrete Fourier Transform, DFT)、梅爾三角濾波器組處理(Mel-Scaled Triangular Filterbank)、離散餘弦轉換(Discrete Cosine Transform, DCT)、對數能量(Logarithm Energy)運算及時間差量(Time Derivation)運算等程序，擷取流程如圖 2-1 所示，下列將詳述梅爾倒頻譜係

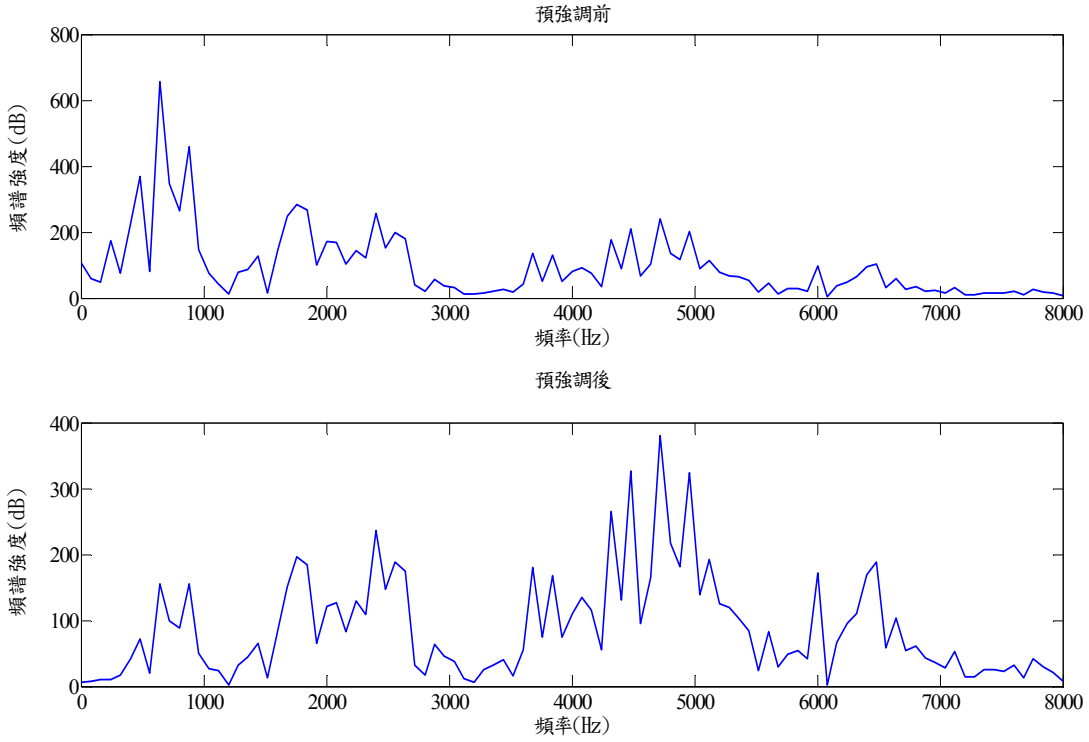


圖 2-2 預強調前與預強調後之振幅比較

數(MFCC)的擷取流程。

預強調(Pre-emphasis)

預強調的動作是讓語音訊號通過一個高通濾波器(High-Pass Filter)，主要作用是要加強聲波高頻的部份，其主要原因可以從幾個不同觀點進行解釋：第一種觀點是聲波在空氣中傳送時，頻率較高的部份會隨著時間增加而衰減，因此預強調可以補償衰減掉的高頻損失；另一種解釋是在發聲的過程中，聲門(Glottal)會抑制住高頻的部份，所以藉由預強調的動作補償高頻的衰減。常見用於預強調的高通濾波器之Z-轉換(Z-Transform)設計為：

$$H(z)=1-\alpha\cdot z^{-1} \qquad \text{(式 2-1)}$$

其中α為預強調的參數，α值的設定通常設為 0.95 左右，若將此高通濾波器之Z-轉換轉成時域(Time Domain)，那麼式 2-1 可改寫成：

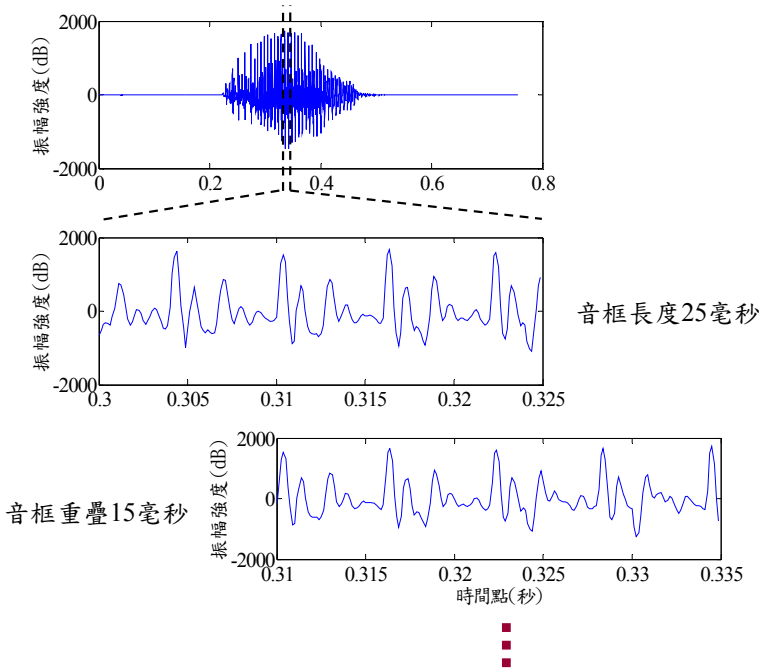


圖 2-3 音框化示意圖

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

其中 $L_y[n]$ 為語音訊號在時域上的第 n 個採樣點。經過預強調處理未經預強調處理處理的振幅比較圖如圖 2-2 所示，由圖中可清楚的看見，經過預強調處理後，低頻部分的能量確實被壓抑住，而高頻的能量相對被加強。

音框化(Windowing)

語音訊號是屬於短時域穩定(Short-Term Stationary)或稱為半穩定(Quasi-Stationary)的訊號，就長時間而言，雖然沒有固定的週期性規則，但在短時間內(約莫 20 毫秒至 30 毫秒)，語音訊號是屬於變化緩慢(Slowly Time Varying)的訊號，因此為了方便後續分析處理，傳統的作法是每隔一小段固定時間就對語音訊號取一個音框，且為了避免音框間的變化過大，二個相鄰的音框會採取部份的重疊(Overlap)，示意圖如圖 2-3 所示，範例的音框長度為 25 毫秒，音框重疊 15 毫秒。再者，因為在時域上對語音訊號做取窗的動作，會使得頻域(Frequency

Domain)上產生摺積的效果，使得訊號產生失真，所以通常會利用視窗函數(Window Function)做進一步處理，常見視窗函數包括矩形視窗(Rectangular Window)、高斯視窗(Gaussian Widow)、漢明視窗(Hamming Window)、漢尼視窗(Hanning Window)等。本論文是採用漢明視窗進行處理，因為漢明視窗的特性是主辦葉(Main Lobe)較寬，邊葉(Side Lobes)較窄，因此可以藉由漢明視窗的使用，可以減少音框化後語音訊號被破壞的情形。漢明視窗的計算方式如下：

$$\widetilde{L}_y[n] = L_y[n] \cdot w[n]$$
$$w[n] = \begin{cases} (1-\beta) - \beta \cdot \cos\left(\dfrac{2\pi \cdot n}{N}\right) & 0 \le n < N-1 \\ 0 & otherwise \end{cases} \qquad \text{(式 2-3)}$$

其中 N 為音框內樣本個數，$\widetilde{L}_y[n]$ 為經過漢明視窗處理的語音訊號，β 為漢明窗調整參數，通常設為 0.46。

離散傅立葉轉換(Discrete Fourier Transform, DFT)

在經過預強調和音框化的動作後，因為人類發音的特徵表現從時域上較難擷取住，所以通常會將語音訊號從時域轉換到頻域(Frequency Domain)上，藉由頻域上的表現，觀察語音訊號的特性，例如觀察語音訊號的共振峰(Formant)等資訊。一般採用離散傅立葉轉換將語音訊號從時域轉換到頻域上。假設 $\widetilde{L}_y[n]$ 是一個週期性為 N_t 的訊號，即

$$\widetilde{L}_y[n] = \widetilde{L}_y[n + N_t] \qquad \text{(式 2-4)}$$

那麼即可利用離散傅立葉轉換將語音訊號由時域轉成頻域

$$Y(k) = \sum_{n=0}^{N_t-1} \widetilde{L}_y[n] \cdot e^{-2j\pi nk/N_t} \quad , \quad 0 \le k < N_t \qquad \text{(式 2-5)}$$

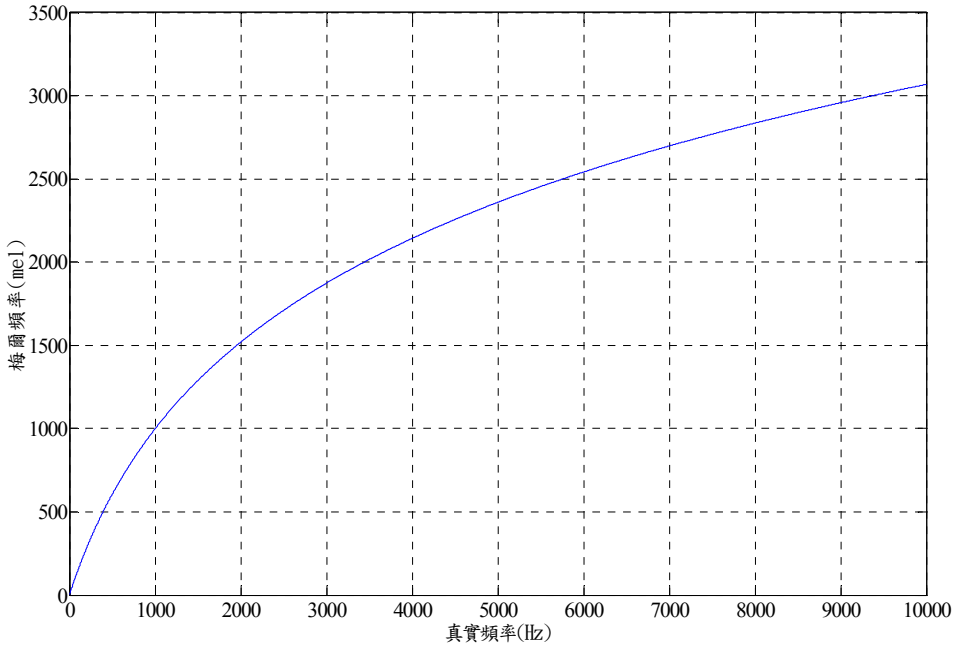


圖 2-4 真實頻率與梅爾頻率對應近似圖

實作上常常會使用快速傅立葉轉換(Fast Fourier Transform, FFT)取代離散傅立葉轉換以加快運算速度[Huang et al. 2001]。

梅爾三角濾波器組(Mel-Scaled Triangular Filterbank)

根據研究指出，人耳對於頻率的感知能力與實際的頻率並非呈線性的關係，在實際頻率為 1K 赫茲(Hz)以下，人類感知的頻率與實際頻率大約呈線性的對應關係，而當實際頻率大於 1K 赫茲以上，人類感知的頻率與實際頻率是呈對數關係的，換言之，人耳感受對低頻的頻率變化比較敏銳，而對高頻部份較為不敏銳。因此為了模擬人類的聽覺敏感度，我們可利用下列數學式模擬，通常下式所計算出的值又稱為梅爾頻率(Mel Frequency)：

$$Mel(f)=1125\cdot\ln\left(1+\frac{f}{700}\right) \qquad \text{(式 2-6)}$$

其中 f 為實際頻率，$Mel(f)$ 為近似的梅爾頻率，真實頻率與梅爾頻率的近似對應情形如圖 2-4 所示。此外為了模擬此人耳聽覺特性，還會進一步使用 M 個三角帶通濾波器(Triangular Band-Pass Filter)進行模擬，三角帶通濾波器的公式如下：

$$H_m[k]=\begin{cases}0 & , \quad k<f[m-1] \\ \dfrac{k-f[m-1]}{f[m]-f[m-1]} & , \quad f[m-1]\le k<f[m] \\ \dfrac{f[m+1]-k}{f[m+1]-f[m]} & , \quad f[m]\le k\le f[m+1] \\ 0 & , \quad k>f[m+1]\end{cases} \qquad \text{(式 2-7)}$$

其中 $f[m]$ 為第 m 個三角帶通濾波器的中心點，$H_m[k]$ 為頻率 k 在第 m 個三角帶通濾波器的權重。假設 f_l 和 f_h 為別代表濾波器組中最低與最高的頻率，F_s 為取樣頻率，N 為快速傅立葉轉換取樣點數，那麼 $f[m]$ 即可進一步表示成

$$f[m]=\left(\frac{N}{F_s}\right)Mel^{-1}\left(Mel(f_l)+m\cdot\frac{Mel(f_h)-Mel(f_l)}{M+1}\right) \qquad \text{(式 2-8)}$$

此外三角帶通濾波器組還有二個重要的特性：第一是降低資料量維度，第二是對頻譜進行平滑化並消除諧波(Harmonic)的作用，以凸顯原語音訊號的共振峰的作用。因此對於第 j 個梅爾三角濾波器輸出值 Mel_j 計算方式如下

$$Mel[j]=\sum_{k=0}^{N}|Y(k)|^2H_j[k] \qquad \text{(式 2-9)}$$

離散餘弦轉換(Discrete Cosine Transform, DCT)

經由對數運算轉換過後的梅爾三角濾波器輸出值，最後會再經由離散餘弦轉換達到降低語音特徵向量維度個數與維度間彼此關係的部份解相關(Partially decorrelation)的目的，離散餘弦轉換數學式表示如下：

$$c[n]=\sqrt{\frac{2}{M}}\sum_{j=1}^{M}\log(Mel[j])\cos\left(\frac{n\cdot\pi}{M}(j-0.5)\right) \quad , \quad n=0,1,\cdots,L<M \qquad \text{(式 2-10)}$$

其中 $c[n]$ 表示語音特徵向量中第 n 維的梅爾倒頻譜係數(特徵值)，L 為語音特徵向量的總維度個數，M 是三角帶通濾波器的個數，$Mel[j]$ 表示第 j 個梅爾三角帶通

濾波器的輸出值。此外，為了模擬人耳聽覺的特性，一般會對梅爾三角濾波器輸出的值作對數運算，同時對數處理也有著動態壓縮的意謂，使得具有較大振幅的濾波器輸出值與較低振幅的濾波器輸出值間差異不會太大。

對數能量(Logarithm Energy)與時間差量計算(Time Derivatives)

除了梅爾倒頻譜係數外，對數能量通常也是一個非常重要的聲學特徵，傳統作法會將對數能量與梅爾倒頻譜係數結合在一起，形成靜態(Static)語音特徵向量。對數能量的計算方式是在經過取窗動作後，將語音訊號值取平方加總起來，數學式表示如下：

$$Log\_E = \log \sum_{n=0}^{N-1} L_y[n]^2 \qquad \text{(式 2-11)}$$

此外，為了擷取住語音訊號在時間軸上的變化，除了先前求得的 L 維的梅爾倒頻譜係數加上對數能量外，還會額外計算其對應應一階差量 $\Delta c_t[n]$ 與二階差量 $\Delta^2 c_t[n]$，以捉住語音特徵參數在時間軸上的改變特質，此又可稱為動態(Dynamic)語音特徵向量，計算方式分別如下：

$$\Delta c_t[n] = \frac{\sum_{p=1}^{P}\left(c_{t+p}[n] - c_{t-p}[n]\right)}{2\sum_{p=1}^{P} p^2} \qquad \text{(式 2-12)}$$

$$\Delta^2 c_t[n] = \frac{\sum_{p=1}^{P} p\left(\Delta c_{t+p}[n] - \Delta c_{t-p}[n]\right)}{2\sum_{p=1}^{P} p^2} \qquad \text{(式 2-13)}$$

$c_t[n]$ 為在時間點 t 中第 n 維的梅爾倒頻譜係數，P 為音框前後的考量個數。

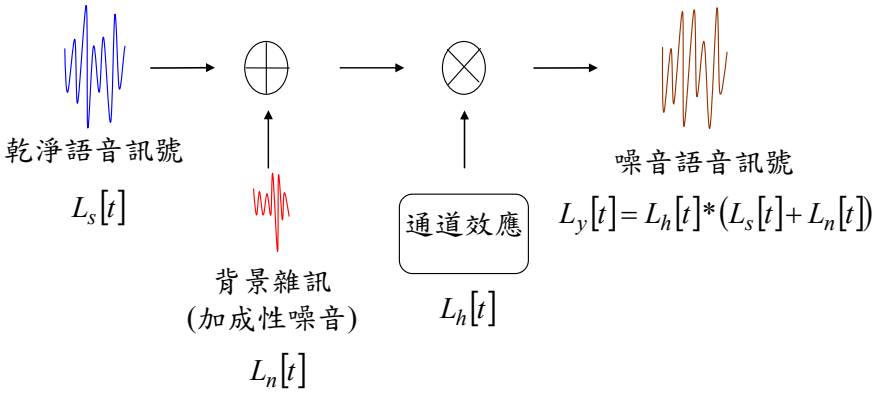


圖 2-5 雜訊干擾示意圖

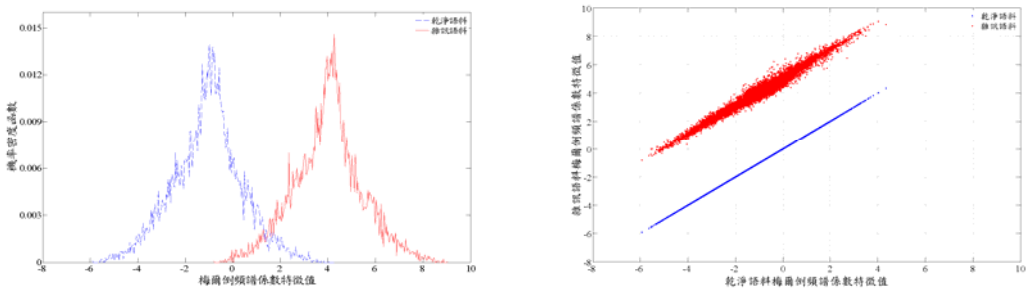
2.2 雜訊干擾影響情形

目前在強健性語音技術的研究上，大致上將干擾語音訊號的雜訊區分為二種類型，分別為加成性噪音(Additive Noise)和摺積性噪音(Convolution Noise)。加成性噪音指的是在收錄語音訊號時，原始語音訊號與其他背景雜訊訊號以線性加成(Linearly Additive)的關係同時被收錄至錄音設備裡，例如周遭人聊天的聲音或是機器設備所發出的噪音等噪音皆屬於加成性噪音；而摺積性噪音通常是指語音訊號經由不同傳輸通道被收錄至錄音設備時，所產生的通道效應，例如電話線路通道效應、麥克風通道效應等。此二類的噪音對於語音訊號的干擾過程，可利用圖 2-5 表示[Acero 1990]。若乾淨語音訊號同時受加成性噪音和摺積性噪音干擾，在時域上此干擾情形可表示成：

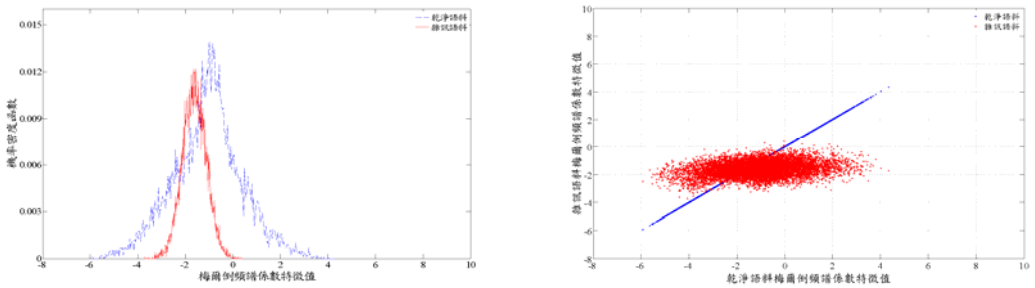
$$L_y[t]=L_h[t]*(L_s[t]+L_n[t]) \qquad \text{(式 2-14)}$$

其中 $L_y[t]$ 與 $L_s[t]$ 分別為雜訊語音訊號與乾淨語音訊號，$L_n[t]$ 與 $L_h[t]$ 分別為表示加成性噪音與摺積性噪音。若以對數頻域表示，對於某個頻率 k，式 2-14 則可重新表示成[Huang, Acero et al. 2001]：

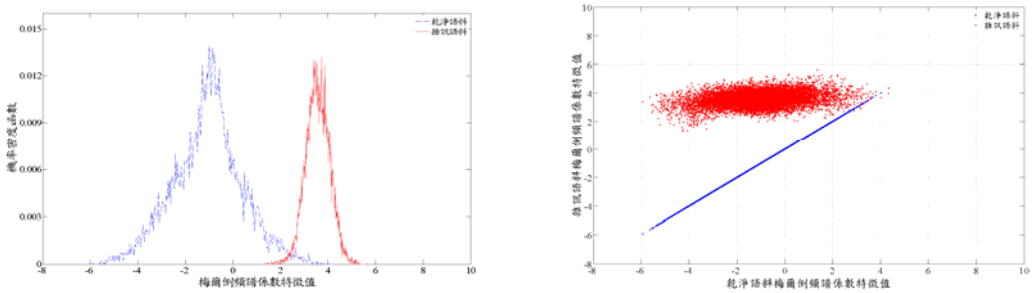
$$\begin{aligned}\log|Y_t(k)|^2 \approx{}& \log|X_t(k)|^2+\log|H_t(k)|^2 \\ &+\log\left(1+\exp\left(\log|N_t(k)|^2-\log|X_t(k)|^2-\log|H_t(k)|^2\right)\right)\end{aligned} \qquad \text{(式 2-15)}$$



(a)通道效應



(b)加成性噪音



(c)通道效應+加成性噪音

圖 2-6 通道效應與加成性噪音對乾淨語音的影響情形

若更進一步以語音辨識中最常用的倒頻譜特徵向量表示，那麼前式可改寫成

$$\begin{aligned} y_t &= x_t + h_t + \mathbf{C}\log\left(1+\exp\left(\mathbf{C}^{-1}\left(n_t - x_t - h_t\right)\right)\right) \\ &= x_t + h_t + g\left(x_t, h_t, n_t\right) \end{aligned} \qquad \text{(式 2-16)}$$

且

$$\begin{aligned} x_t &= \mathbf{C}\left(\log|X_t(k_1)|^2 \quad \log|X_t(k_2)|^2 \quad \cdots \quad \log|X_t(k_M)|^2\right) \\ h_t &= \mathbf{C}\left(\log|H_t(k_1)|^2 \quad \log|H_t(k_2)|^2 \quad \cdots \quad \log|H_t(k_M)|^2\right) \\ n_t &= \mathbf{C}\left(\log|N_t(k_1)|^2 \quad \log|N_t(k_2)|^2 \quad \cdots \quad \log|N_t(k_M)|^2\right) \\ y_t &= \mathbf{C}\left(\log|Y_t(k_1)|^2 \quad \log|Y_t(k_2)|^2 \quad \cdots \quad \log|Y_t(k_M)|^2\right) \end{aligned} \qquad \text{(式 2-17)}$$

其中 x_t 為乾淨語音倒頻譜特徵向量，n_t 為加成性噪音倒頻譜特徵向量，h_t 為摺積性噪音倒頻譜特徵向量，**C** 為離散餘弦轉換。因此從式 2-16 可以清楚地發現，摺積性噪音會造成乾淨語音訊號產生特徵參數值的值域偏移(Shift)情形發生，而加成性噪音則是因為經過對數轉換的關係，使得其對語音特徵參數產生非線性失真的影響，如 $g(x_t,h_t,n_t)$，因此此二種現象即為造成乾淨語音訊號和雜訊語音訊號二者間統計特性不匹配的主要原因。

雜訊干擾對於乾淨語音訊號所造成的失真情形如圖 2-6 所示，圖中是由利用梅爾倒頻譜係數特徵向量第一維特徵值所畫而成，圖片左半部為乾淨語音訊號與雜訊語音訊號的統計分布圖(Histogram)，藍色線段是乾淨語音的統計分布圖，而紅色線段是受雜訊語音的統計分布圖，橫軸為梅爾倒頻譜係數特徵值，縱軸為機率密度函數(Probability Density Function, PDF)值；圖片右半部為乾淨語音訊號與雜訊語音訊號的特徵值的值域散布圖(Scattergram)，其中藍色散布點的描繪是利用乾淨語音的特徵值當做橫軸參考座標值與縱軸參考座標值；紅色散布點數的描繪是以乾淨語音的特徵值當做橫軸參考座標值，以及所對應雜訊語音的特徵值為縱軸參考座標值。圖片由上至下分別代表乾淨語音訊號受到(1)摺積性噪音干擾、(2)加成性噪音干擾－利用 0dB 地下鐵噪音與(3)摺積性噪音與加成性噪音同時干擾的影響狀況。對於摺積性噪音的干擾，很明顯可看出其對整個資料分布的值域產生偏移的情形，而加成性噪音則是改變了語音的統計分布的機率密度函數，若同時考慮二種噪音的影響，那麼不僅會產生值域偏移的現象，同時也會改變了原本的統計分布的機率密度函數。可想而知，如果語音辨識系統的聲學模型是利用乾淨語料訓練而成的，但測試環境卻是在吵雜的噪音環境下進行，那麼勢必會存在訓練環境與測試環境產生不匹配的情形，進而大幅降低系統的辨識效能。

2.3 強健性語音特徵技術

強健性語音特徵技術主要是從語音訊號中擷取出較不易受到環境變化干擾而失真的語音特徵參數，目前就常見的技術而言，可再細分為三個研究方向：(1)語音特徵參數轉換法(Feature Transformation)、(2)語音特徵參數補償法(Feature Compensation)和(3)語音特徵參數重建法(Feature Reconstruction)。下面三小節將分述近年來一些較被廣為討論的方法。

2.3.1 語音特徵參數轉換法(Feature Transformation)

如同 2.2 小節所述，語音特徵參數容易受到雜訊干擾而產生變化，因此便有研究學者嘗試不同的轉換方法，期望找出更具強健性的語音特徵，且不易受到雜訊或者通道效應影響。此研究方向又可再細分二種研究議題：第一種是資料相關線性語音特徵空間轉換(Data-Driven Linear Feature Transform)，主要將語音特徵參數轉換至另一種語音特徵向量空間(Feature Space)，使得轉換後的語音特徵向量能帶有或者保留較具有鑑別力的鑑別資訊(Discriminative Information)成份，第二種研究方向是語音特徵參數正規化(Normalization)，期望藉由對語音特徵向量的正規化過程中，進一步移除雜訊干擾的影響。

2.3.1.1 資料相關線性語音特徵空間轉換

資料相關線性語音特徵空間轉換的好處除了藉由統計訓練語料的統計資訊，自動地找出特徵空間中重要的基底向量，使得經轉換後的語音特徵參數能只保留具較大變異或者有鑑別力的特徵成分，並且能進一步去除多餘(Redundant)的維度。常見的方法有主成分分析(Principal Component Analysis, PCA)、線性鑑別分析(Linear Discriminant Analysis, LDA)[Duda and Hart 1973]、異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)[Kumar 1997; Gales 2002]、異質性鑑別分析(Heteroscedastic Discriminant Analysis, HDA)[Saon et al. 2000]。其中線性鑑別分析是假設所有類別特徵向量的分布變異是相同的；而異質性線性鑑別

分析與異質性鑑別分析則是打破這樣的假設，允許類別間的分布變異可以不同。同時，也有研究嘗試以核函數線性鑑別分析(Kernel Linear Discriminant Analysis, Kernel LDA) [Mika 1999]對語音特徵向量做進一步處理，利用核函數(Kernel Function)將語音特徵向量投影(Project)到高維度的特徵空間，再作線性鑑別分析，以解決在原本特徵空間中可能存在著非線性鑑別的問題。

另一方面，由於在聲學模型(例如隱藏式馬可夫模型狀態觀測機率分布)中為了計算方便與加快運算速度，常使用具對角化共變異矩陣(也就是假設特徵向量維度間彼此為無關的)的高斯分布，但是前述的語音特徵向量或是鑑別分析(如線性鑑別分析或異質性線性鑑別分析)並不保證一定能擁有此一特性，因而有學者提出最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)[Gales 1998]，嘗試讓轉換過後的共變異矩陣的值集中在對角線上，在對聲學模型相似度影響最小的條件下，儘量滿足對角化共變異矩陣的要求。因此，目前在語音辨識的語音特徵擷取上常見到以結合線性鑑別分析與最大相似度線性轉換(LDA-MLLT)或是異質性線性鑑別分析與最大相似度線性轉換(HLDA-MLLT)[Beyerlein et al. 2002; Hain et al. 2005]等作法。

2.3.1.2 語音特徵參數正規化

語音特徵參數正規化法通常只需很少量的運算時間，即可明顯地提昇辨識效能。目前最常見的方法之一是倒頻譜平均消去法(Cepstral Mean Substraction, CMS)[Furui 1981]，倒頻譜平均消去法主要是針對語音特徵參數第一階動差(First Moment)進行正規化，數學式表示如下：

$$\overline{Y}^i = \frac{1}{T}\sum_{t=1}^{T} y_t^i \qquad \text{(式 2-18)}$$

$$\tilde{y}_t^i = y_t^i - \overline{Y}^i \qquad \text{(式 2-19)}$$

其中 y_t^i 表示第 t 個音框的第 i 維語音特徵參數，T 表示總音框個數，$\overline{Y}^i$ 代表語音特徵參數中第 i 維的平均數，$\widetilde{Y}_t^i$ 為經過倒頻譜消去法所得到的新語音特徵參數。由式 2-16 可以明顯看出通道效應對語音特徵參數的影響是常數固定不變的，從發音開始至結束都會一直存在著，因此對於二種具有不同通道效應的語音特徵參數而言，若同時使用倒頻譜平均消去法，那麼這二種通道效應對語音特徵參數的影響已不復存在[[Huang et al. 2001]]。

因為倒頻譜平均消去法在調變頻譜(Modulation Frequency)上的表現類似一帶通濾波器(Band-Pass Filter)，因此後來有學者提出相類似的作法，例如相對頻譜法(RelAtive SpecTrAl, RASTA)[Hermansky and Morgan. 1994]，作者觀察到人類的發音特性，在調變頻譜上變化低於 1 赫茲或高於 12 赫茲的訊號源均屬於非人類發音的訊號，因此可視其為雜訊訊號，所以使用一個帶通濾波器，進而移除變化速度相對較快或較慢的雜訊訊號。基本做法是針對數個音框的語音特徵向量進行平滑動作(Smoothing)(大約是 150-170 毫秒)，所以 Z-轉換表示如下

$$R(z) = 0.1z^4 * \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \qquad \text{(式 2-20)}$$

此外，倒頻譜平均消去法亦有許多的延伸，例如倒頻譜正規化法(Cepstral Mean and Variance Normalization, CMVN)[Vikki and Laurila 1998]，其針對語音特徵參數第一和第二階動差(Second Moment)進行正規化，數學式如下所示：

$$\overline{Y}^i = \frac{1}{T}\sum_{t=1}^{T} y_t^i \qquad \text{(式 2-21)}$$

$$\sigma^i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(y_t^i - \overline{Y}^i\right)^2} \qquad \text{(式 2-22)}$$

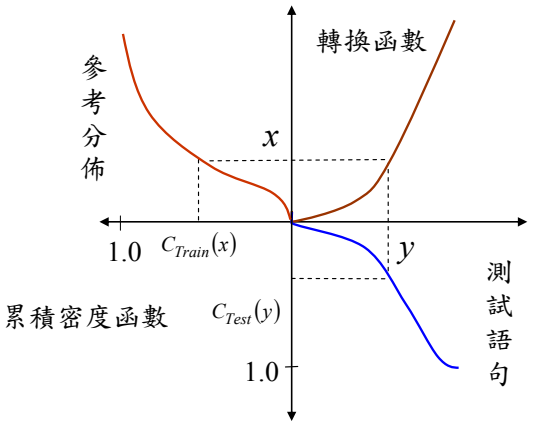


圖 2-7 統計圖等化法示意圖

$$\tilde{y}_t^i = \frac{y_t^i - \bar{Y}^i}{\sigma^i} \qquad \text{(式 2-23)}$$

其中σ^i代表語音特徵參數中第i維的標準差(Standard Deviation)，經由式 2-23 的處理可以使語音特徵參數的平均數為 0 且標準差 1(同特意謂著變異數亦為 1)，所以倒頻譜正規化法除了能移除通道效應所造成的影響，同時也降低了每個維度間語音特徵機率分布的差異程度，進一步的降低環境不匹配對特徵參數所造成的不良影響。此外，為了滿足即時回應的系統需求，亦有學者發展出只需短暫延遲(Delay)的演算法[Pujol et al. 2006]。

但是由於倒頻譜平均消去法與倒頻譜正規化法本身線性關係的限制，造成只能補償因受雜訊干擾影響所產生的線性失真部份，對於非線性失真部份的補償效果有限，因此便有許多學者嘗試提出許多不同的補償方法，試圖解決因雜訊干擾影響對語音特徵參數所產生的失真情形，例如針對語音特徵參數的第三階動差進行正規化[Suk et al. 1999]或對語音特徵參數更高階動差進行正規化[Hsu and Lee 2004; 2006]。此外，亦有學者嘗試將已經在影像處理中行之有年的統計圖等化法(Histogram Equalization)應用於語音辨識之特徵參數正規化[Dharanipragada and Padmanabhan 2000; Molau 2003; Torre et al. 2005; Hilger and Ney 2006]。

統計圖等化法除了試圖去匹配訓練語料與測試語料之語音特徵參數的平均數和變異數之外，更企圖讓訓練語料和測試語料的統計分布特性能夠相同(Identical)，其作法是藉由將測試語料的累積密度函數(Cumulative Density Function, CDF)值應對至由訓練語料所統計出來的參考分布的累積密度函數值，藉由此匹配轉換過程，降低測試語料與訓練語料由於環境因素影響所造成統計特性不匹配的現象，實驗結果證實統計圖等法化對提升辨識效果有很明顯的幫助[Molau 2003; Torre et al. 2005]。另外更有學者嘗試將統計圖等化法概念推廣至向量量化編碼(Vector Quantization，VQ)，且更進一步應用於分散式語音辨識(Distributed Speech Recognition, DSR)上[Wan and Lee 2005, 2006; Wan et al. 2007]，利用統計圖資訊做為向量之量化準則，有效解決傳統以距離為量化準則容易受環境雜訊影響或是容易形成量化失真(Quantization Distortion)的問題。

統計圖等化法主要假設測試語句之語音特徵參數的統計分布會與訓練語料特徵參數的統計分布(參考分布)是一致的，其最基本精神可以視為是要求取一個轉換函數(Transformation Function)，使得此函數能將測試語句的語音特徵參數中每一維特徵向量的統計分布轉換至先前已從訓練語句中定義好的參考分布。一般而言，若語音特徵參數擷取是利用梅爾倒頻譜係數，那麼統計圖等化法可以作用在梅爾濾波器組輸出[Molau et al. 2001; Molau 2003; Molau et al. 2003]或是梅爾倒頻譜係數[Dharanipragada and Padmanabhan 2000; Segura et al. 2004; Torre et al. 2005]。統計圖等化法數學式關係式可表示如下[Torre et al. 2002, 2005]：假設 y 為測試語音中某一維的語音特徵參數，且具有機率密度函數 $p_{Test}(y)$，那麼經由統計圖等化法的假設，即測試語音的機率密度函數 $p_{Test}(y)$ 與參考分布的機率密度函數 $p_{Train}(x)$ 的假設下，因此可以利用一轉換函數 $F(y)$，將測試語音的語音特徵參數 y 轉換至參考語音特徵參數 x：

$$p_{Train}(x) = p_{Test}(y)\frac{dy}{dx} = p_{Test}\left(F^{-1}(x)\right)\frac{d\left(F^{-1}(x)\right)}{dx} \qquad \text{(式 2-24)}$$

其中 $F^{-1}(x)$ 為 $F(x)$ 的逆函數(Inverse Function)，若將上述關係式以累積密度函數的觀點表達，即可表示成

$$\begin{aligned} C_{Test}(y) &= \int_{-\infty}^{y} p_{Test}(y')dy' \\ &= \int_{-\infty}^{F(y)} p_{Test}\left(F^{-1}(x')\right)\frac{dF^{-1}(x')}{dx'}dx' \\ &= \int_{-\infty}^{x} p_{Train}(x')dx'|_{x=F(y)} \\ &= C_{Train}(x) \end{aligned} \qquad \text{(式 2-25)}$$

其中 $C_{Test}(y)$ 和 $C_{Train}(x)$ 分別為測試語句和訓練語料的累積密度函數，x 為經由轉換函數 $F(y)$ 求得的結果，所以轉換函數 $F(y)$ 會具有下列特性

$$x = F(y) = C_{Train}^{-1}(C_{Test}(y)) \qquad \text{(式 2-26)}$$

其中 C_{Train}^{-1} 為 C_{Train} 的逆函數，轉換過程如圖 2-7 所示。

因此在實作上，由於訓練語料與測試語料的語音特徵參數通常皆為一有限集合，所以並無法精準估算其實際的累積密度函數值，較常見的作法是使用累積統計圖(Cumulative Histogram)來近似累積密度函數值。對於所有訓練語料而言，語音特徵參數中的每一維特徵向量會統計出一個累積統計圖，再依需求將累積統計圖設定為 i 個分位差(Quantile)，每個分位差區間皆以區間內所有特徵值的平均數做為該分位差的代表特徵值(Representative Value)，且此資訊亦被用來當做轉換的參考分布。對測試語句的每一維度特徵向量同樣統計出累積統計圖，也取 i 個分位差，接著對測試語句的每個分位差區間內的特徵值用先前以訓練語料建立好的特徵參數參考分布逐一進行轉換取代。一般實作可利用表格查詢(Table-lookup)的方式進行，首先先以表格方式紀錄參考分布的累積統計圖資訊，例如記錄成{分位差區間索引值，特徵值}；接著在進行等化(Equalization)過程時，將所有表格載入記憶體中以方便進行查表轉換。往往要得到良好的辨識效果，使用的分位差

區間數數不可太少，亦代表需耗費大量的記憶體空間，並且在進行查表轉換時，也需花費不少的搜尋時間。

上述介紹的統計圖等化法的轉換動作都是直接根據測試語句的累積統計圖進行，並無需使用任何額外的參數，但 Hilger 等研究學者後來提出一種參數型態(Parametric)的分位差統計圖等化法[Hilger and Ney 2001, 2006]，對於語音特徵向量中每一維的特徵值 y，改用一轉換函數 $H(y)$ 進行等化動作，數學關係式表示如下：

$$\tilde{y} = H(y) = Q_K\left(\alpha\left(\frac{y}{Q_K}\right)^{\gamma} + (1-\alpha)\left(\frac{y}{Q_K}\right)\right) \qquad \text{(式 2-27)}$$

$\tilde{y}$ 為轉換後新的語音特徵值，Q_K 為最後一個分位差值，亦即整句語句中最大的特徵值；α 和 γ 為轉換函數 $H(y)$ 所需的參數可利用式 2-28 求得。值得注意的是在對於每一句語句在進行等化過程前，需先對整句語句與參考分布進行分位差校正(Quantile Correction)，以求得最佳的參數，此校正動作是以最小平方差(Least Squares Error, LSE)進行，利用格式搜尋法(Grid Search)，將 α 和 γ 個別限制在一段值域區間內，以等距的數值代入式 2-28 進行搜尋，進而找出使得誤差最小的 α 和 γ 值，數學式定義如下：

$$\{\alpha,\gamma\} = \arg\min_{\{\alpha,\gamma\}}\left(\sum_{k=1}^{K-1}\left(H(Q_k) - Q_k^{train}\right)^2\right) \qquad \text{(式 2-28)}$$

其中 K 為分位差的個數；Q_k 為待轉換語句中第 k 個分位差的特徵值；Q_k^{train} 為訓練語料所統計出的參考分布中的第 k 個分位差值。分位差統計圖等化法的處理流程是先經由式 2-28 計算以求得最佳參數 α 和 γ，接著再利用式 2-27 中一組非線性函數和一組線性函數進行加權合併，期望轉換後的語音特徵參數的統計分布能夠和參考分布愈相似愈好，此外，此法對於雜訊干擾而形成的非線性失真部份，

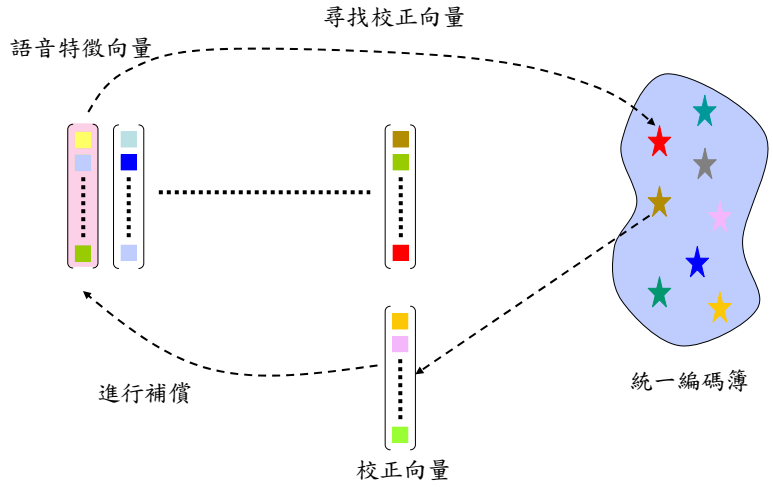


圖 2-8 編碼為基礎之倒頻譜向量補償法示意圖

可藉由γ項的設定進行補償。但由於針對每一測試語句都必經由式 2-28 求得最佳的參數α和γ而進行語音特徵參數正規化，因此在實際應用上，必須需耗費不少的處理器運算時間利用格式搜尋法做完整的搜尋。雖然統計圖等化法或分位差統計圖等化法能提昇語音辨識系統的辨識效能，但從上面的敘述可清楚地發現，此二種方法在執行等化過程，可能需耗費大量的記憶體空間或者是處理器運算時間，因此可想而知，若要將其運用在運算能力有限的行動裝置上，往往是不太可行的。

2.3.2 語音特徵參數補償法(Feature Compensation)

語音特徵參數轉換法是期望找出更具強健性的語音特徵，且不易受到雜訊影響，然而語音特徵參數補償法是希望能夠將受雜訊干擾語音特徵參數補償至未受雜訊干擾的語音特徵參數。在語音特徵參數補償法中，本論文在此將著重於介紹近年來被廣為研究與討論的議題，吾人稱之為以編碼為基礎之倒頻譜向量補償法(Codebook-based Cepstral Compensation)，在此類方法中，主要的精神是利用向量量化編碼技術對語音特徵向量建立一組統一編碼簿(Universal Codebook)，且統一編碼簿裡的每一個向量量化編碼區域(Vector Quantization Region)都有一組對應的補償向量，用來當作是將雜訊語音特徵向量還原至乾淨語音特徵向量的補償參

考值。在測試時，對於每個音框而言，都需至統一編碼簿中找出一個最相似的向量量化編碼區域，再利用該編碼區域所對應的補償向量進行補償，整體示意圖如圖 2-8 所示。目前常見的方法有編碼詞相關之倒頻譜正規化法(Codeword Dependent Cepstral Normalization, CDCN)[Acero 1990]、機率最佳化過濾法(Probabilistic Optimum Filtering, POF)[Neumeyer and Weintraub 1994]與雙聲源為基礎分段線性補償(Stereo-based Piecewise Linear Compensation, SPLICE)[Deng et al. 2000]等。

(1)編碼詞相關倒頻譜正規化法(CDCN)

編碼詞相關倒頻譜正規化法主要假設雜訊語音特徵向量 x 與乾淨語音特徵向量 y 間的關係如下：

$$\hat{x} = y - h - r \qquad \text{(式 2-29)}$$

其中 h 為摺積性噪音，r 為補償向量。且又假設乾淨語料的資料分布情形可以用一具有 K 個高斯分布的高斯混合模型(Gaussian Mixture Model, GMM)表示，換句話說，每個高斯分布可以被視為是某一些特定音素相關(Phoneme-Dependent)的資料群集分布。有了此二個假設後，在測試時，給定一連串雜訊語音特徵向量 Y 和乾淨語料訓練而成的高斯混合模型，即可透過最大相似度(Maximum Likelihood, ML)準則及使用期望值最大化演算法(Expectation Maximum, EM)線上即時估測摺積性噪音 $\hat{h}$ 與補償向量 $\hat{r}$。最大相似度的目標函數(Objective Function)為：

$$\begin{aligned}(\hat{h},\hat{n}) &= \arg\max_{(h,n)} \log p(Y \mid h,n) \\ &= \arg\max_{(h,n)} \sum_{t=0}^{T-1} \log p(y_t \mid h,n)\end{aligned} \qquad \text{(式 2-30)}$$

且

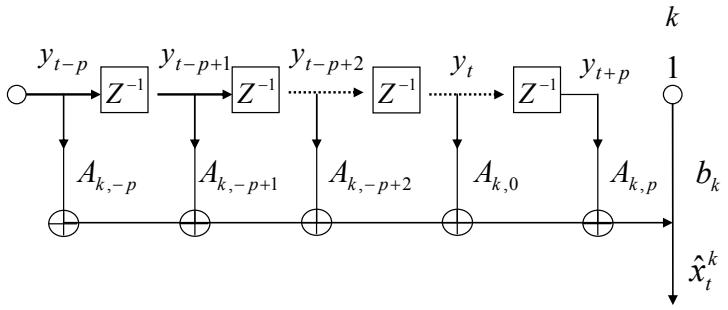


圖 2-9 多維度橫向濾波器示意圖

$$p(y_t \mid h,n) = \sum_{k=0}^{K-1} c_k \cdot \frac{1}{\sqrt{2\pi}\left|\Sigma_k\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\hat{x}_t^k - \mu_k\right)\Sigma_k^{-1}\left(\hat{x}_t^k - \mu_k\right)\right) \qquad \text{(式 2-31)}$$

其中 T 為總音框個數，因為如式 2-16 所示，加成性噪音 n 對在倒頻譜特徵上的表現為一非線性失真，因此本方法假設此一失真可經由一補償向量 $\hat{r}_k$ 進行補償，所以乾淨語音特徵向量 $\hat{x}_t^k$ 可經由 $y_t - \hat{h} - \hat{r}_k$ 求得，因此只需最小化式 2-30 即可估測出摺積性噪音 $\hat{h}$ 為何，且每個高斯分布亦可估測出相對應的補償向量 $\hat{r}_k$，最後補償後的語音特徵向量由下式求得：

$$\tilde{y}_t = y_t - \hat{h} - \sum_{k=1}^{K} P(k \mid y_t) \times \hat{r}_k \qquad \text{(式 2-32)}$$

$P(k \mid y_t)$ 為給定雜訊語音特徵向量 y_t 時，發生在某個高斯分布 k 的事後機率(Posteriori Probability)。但此法最大的缺點是每次測試都需線上利用期望最大值演算法進行數次迭代(Iterations)以求得摺積性噪音 $\hat{h}$ 與補償向量 $\hat{r}_k$，通常此迭代過程需耗費不少處理器運算時間，因此實作上速度較慢。

(2)機率最佳化過濾法(POF)

機率最佳化過濾法與編碼詞相關倒頻譜正規化法有相類似的概念，且可以視為是稍後將介紹的雙聲源為基礎分段線性補償的一般化通式(Generalization)，其主要

是利用雙聲源語料與分段最小平方差(Piecewise Minimum Square Errors)準則，設計出濾波器組以對雜訊語音特徵參進進行補償。雙聲源語料是指再收錄語音訊號時，同時收錄二份語音檔案，一份語音檔案表示未受雜訊干擾的語音訊號，通常可以利用麥克風較接近語者的嘴巴進行收錄，另一份語音檔案表示受雜訊干擾的語音訊號，收錄方式可將麥克風放至離語者較遠的地方，除了收錄語者的講話內容外，同時也將環境噪音收錄進去。機率最佳化過濾法與編碼詞相關倒頻譜正規化法相同之處在於其亦是利用向量量化編碼技術將乾淨訓練語料分成 K 個向量量化編碼區域，每個編碼區域 g_k 會對應到一組多維度橫向濾波器(Multi-dimensional Transversal Filter)，其作用示意圖如圖 2-9 所示，因此經過濾波器組補償後的語音特徵向量可經由下式求得：

$$\widetilde{y}_t = \sum_{k=0}^{K-1} P(g_k \mid z_t)\mathrm{W}_k^{\mathrm{T}}\mathrm{Y}_t = \left[\sum_{k=0}^{K-1} P(g_k \mid z_t)\mathrm{W}_k^{\mathrm{T}}\right]\mathrm{Y}_t \qquad \text{(式 2-33)}$$

其中 z_t 是雜訊語音中第 t 個語音特徵向量有關的指示條件向量(Conditional Vector)，主要是用來判斷第 t 個語音特徵向量落在某個向量編碼區域的事後機率，通常可利用訊噪比、訊號能量強度或倒頻譜特徵向量等資訊；$P(g_k \mid z_t)$ 為給定指示條件向量 z_t 時，發生在編碼區域 g_k 的事後機率，$\mathrm{W}_k^{\mathrm{T}}$ 為在編碼區域 g_k 所對應的濾波器的參數矩陣(Coefficient Matrix) ，Y_t 為第 t 個音框串接前後各 p 個音框的雜訊語音特徵向量所組成的超級向量(Super-vector)，其分別定義如下：

$$\mathrm{W}_k^{\mathrm{T}} = \left[A_{k,-p}\cdots A_{k,-1}A_{k,0}A_{k,1}\cdots A_{k,p}b_k\right] \qquad \text{(式 2-34)}$$

$$\mathrm{Y}_t^{\mathrm{T}} = \left[y_{t-p}^{\mathrm{T}}\cdots y_{t-1}^{\mathrm{T}}y_t^{\mathrm{T}}y_{t+1}^{\mathrm{T}}\cdots y_{t+p}^{\mathrm{T}}1\right] \qquad \text{(式 2-35)}$$

對於每個多維度橫向濾波器的參數矩陣可利用最小均方誤差法求得。首先先定義第 k 個多維度橫向濾波器的誤差方式計算如下：

$$e_{tk} = x_t - \hat{y}_t = x_t - \mathrm{W}_k^{\mathrm{T}}\mathrm{Y}_t \qquad \text{(式 2-36)}$$

x_t 為第 t 個音框的乾淨語音特徵向量，$\hat{y}_t$ 為經過多維度橫向濾波器補償後得到的語音特徵向量，那麼對於全部訓練語料而言，第 k 個向量量化編碼區域總條件誤差(Conditional Error)則為

$$E_k = \sum_{t=p}^{T-1-p} \left\| e_{tk} \right\|^2 P(g_k \mid z_t) \qquad \text{(式 2-37)}$$

為了使式 2-37 誤差最小，我們可以用每個多維度橫向濾波器的參數矩陣 $\mathrm{W}_i^{\mathrm{T}}$ 對式 2-37 做偏微分，令微分後的結果於零，最後每個多維度橫向濾波器可用下式求得

$$\mathrm{W}_k = R_k^{-1} r_k \qquad \text{(式 2-38)}$$

其中

$$R_k = \sum_{t=p}^{T-1-p} \mathrm{Y}_t \mathrm{Y}_t^{\mathrm{T}} P(g_k \mid z_t) \qquad \text{(式 2-39)}$$

且

$$r_k = \sum_{t=p}^{T-1-p} \mathrm{Y}_t x_t^{\mathrm{T}} P(g_k \mid z_t) \qquad \text{(式 2-40)}$$

(3) 雙聲源為基礎分段線性補償(SPLICE)

雙聲源為基礎分段線性補償為近來非常熱門之議題[Deng et al. 2000; Droppo et al. 2001, 2002, 2005]，其概念是從編碼詞相關倒頻譜正規化法與機率最佳化過濾法延伸而得，主要利用高斯混合模型來表示受雜訊干擾的語音特徵參數分布情形，在高斯混合模型中的每個高斯分布可被視為語音特徵參數在某一種特定噪音環境下的分布情形，高斯混合模型表式如下：

$$p(y_t)=\sum_{k=1}^{K}P(k)p(y_t\mid k)=\sum_{k=1}^{K}c_k N(y_t;\mu_k,\Sigma_k) \quad \text{(式 2-41)}$$

其中 y_t 代表雜訊語音特徵參數空間中第 t 個語音特徵向量，K 為所有高斯分布的個數，$p(y_t\mid k)$ 及 $N(y_t;\mu_k,\Sigma_k)$ 表示雜訊語音特徵向量 y_t 落在第 k 個高斯分布的相似度，$P(k)$ 及 c_k 表示第 k 個高斯分布在高斯混合模型內的的權重(Mixture Weight)。此外每個高斯分布都會有一組對應的補償向量表示受雜訊干擾的語音特徵參數和未受雜訊干擾的語音特徵參數之間的差異，最後，補償後的語音特徵向量 $\tilde{y}_t$ 可利用最小均方誤差法求得，求算方式如下。

$$\tilde{y}_t=\hat{x}_t=\mathbf{E}[x_t\mid y_t]=\sum_{k}p(k\mid y_t)\mathbf{E}_x[x_t\mid y_t,k] \quad \text{(式 2-42)}$$

$\mathbf{E}[\cdot]$ 表示取期望值的意思，$\tilde{y}_t$ 為補償後的語音特徵向量，同時還進一步假設就第 k 個高斯分布而言，雜訊語音特徵向量 y_t 和乾淨語音特徵向量 x_t 間的差異可以經由用一線性補償向量 r_k 進行補償，因此 $\mathbf{E}_x[x_t\mid y_t,k]$ 可表示成如下式：

$$\mathbf{E}_x[x_t\mid y_t,k]\approx y_t+r_k \quad \text{(式 2-43)}$$

最後，補償後的語音特徵向量 $\tilde{y}_t$ 可改寫成下式：

$$\tilde{y}_t=y_t+\sum_{k}P(k\mid y_t)\cdot r_k \quad \text{(式 2-44)}$$

假設有雙聲源語料乾淨語音 x_t 與雜訊語音 y_t，那麼 r_k 的求得可經由下式估測而得：

$$r_k=\frac{\sum_{t=0}^{T-1}P(k\mid y_t)(x_t-y_t)}{\sum_{t=0}^{T-1}P(k\mid y_t)} \quad \text{(式 2-45)}$$

其中 y_t 為時間點 t 的雜訊語音特徵向量，x_t 為相對應的乾淨語音特徵向量，k 表

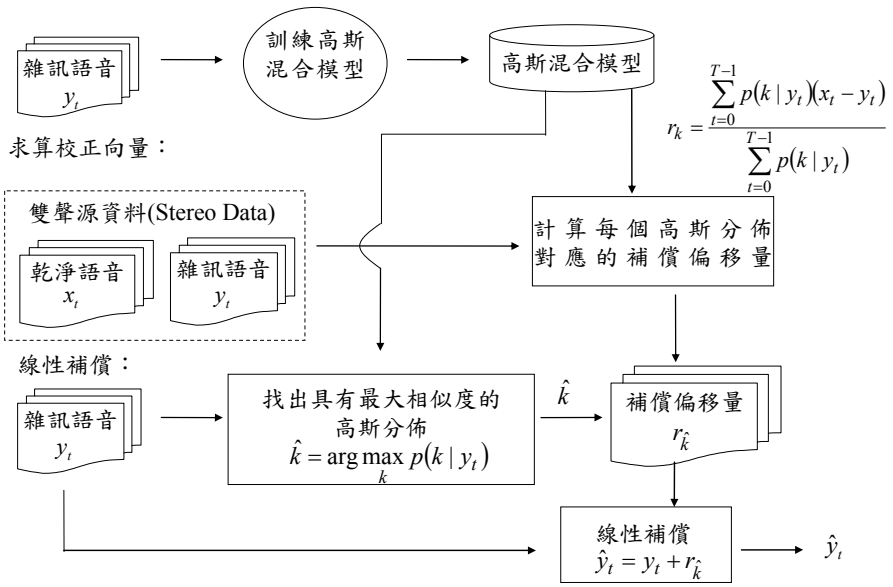


圖 2-10 雙聲源為基礎之分段線性補償流程圖

示高斯混合模型中第 k 個高斯分布。因為 k 與高斯混合模型的高斯分布個數有關，所以為了快速計算而言，可使用最大事後機率法則(Maximum a Posteriori, MAP)簡化式 2-44 的運算複雜度。

$$\hat{P}(k \mid y_t) \approx \begin{cases} 1 & k = \arg\max_k P(k \mid y_t) \\ 0 & otherwise \end{cases} \quad \text{(式 2-46)}$$

其中 $p(k \mid y_t)$ 為給定雜訊語音特徵向量 y_t，發生在第 k 個高斯分布的事後機率，計算方式如下：

$$P(k \mid y_t) = \frac{p(y_t \mid k)P(k)}{\sum_{k'=0}^{K-1} p(y_t \mid k')P(k')} \quad \text{(式 2-47)}$$

因此雙聲源為基礎之分段線性補償可分二個步驟完成，第一個步驟是找出高斯混合模型中和 y 具有最大事後機率的高斯分布 k，接著再利用和該高斯分布 k 所對應的補償向量進行補償，整體流程如圖 2-10 所示，數學關係式如下：

$$\hat{k} = \arg\max_k P(k \mid y_t) \quad \text{(式 2-48)}$$

$$\tilde{y}_t = y_t + r_{\hat{k}} \qquad \text{(式 2-49)}$$

根據過去研究成果指出，不論使用式 2-44 或是使用式 2-49，所得到的補償效果幾乎是相同的[Droppo et al. 2001]，所以在實作上，可以使用式 2-49 較節省運算複雜度。雖然雙聲源為基礎之分段線性補償雖然能有效的補償受雜訊干擾的語音特徵參數，但最大的缺點是對於實際上的語音辨識系統而言，雙聲源的訓練語料往往是不容易取得。因此，最近有學者嘗試將雙聲源為基礎之分段線性補償延伸改良成僅使用雜訊語料而已，該種方法稱為隨機特徵向量對映法(Stochastic Vector Mapping, SVM)[Droppo and Acero 2005; Wu et al. 2005; Huo and Zhu 2006; Wu and Huo 2006]。

(4) 隨機特徵向量對映法(SVM)

在隨機特徵向量對映法中，因為只有利用雜訊語料進行補償向量的估測，因此必須搭配聲學模型使用以求得補償向量的補償值。如果估測補償向量是以最大相似度為考量出發點的話，那麼可以解釋成雜訊語音特徵向量在加入補償向量後，落在聲學模型上相似度會愈大[Wu et al. 2005]；若是以鑑別力為出發點，那麼可以解釋成語音特徵向量加入了補償向量後會帶較多的鑑別性資訊[Droppo and Acero 2005]。在此回顧中，假設雜訊語音特徵向量是利用線性的方法進行補償，因式數學關係式可表示成式 2-50 或式 2-52。

$$\tilde{y}_t = F_1(y_t, \mathbf{B}) = y_t + \sum_{k=1}^{K} P(k \mid y_t) b_k \qquad \text{(式 2-50)}$$

$$P(k \mid y_t) = \frac{p(y_t \mid k) P(k)}{\sum_{k=1}^{K} p(y_t \mid k) P(k)} \qquad \text{(式 2-51)}$$

其中 **B** 為補償向量的集合。或

$$\tilde{y}_t = F_2(y_t, \mathbf{B}) = y_t + b_{k'} \qquad \text{(式 2-52)}$$

其中

$$k' = \arg\max_{k'} P(k' \mid y_t) \tag{式 2-53}$$

如果補償後的效果是以最大相似度為考量出發點[Wu et al. 2005]，那麼假定在給定聲學模型 Λ 下，目標函數則可定義為：

$$L(\mathbf{B}, \Lambda) = \prod_{t=1}^{T} p(F(y_t, \mathbf{B}) \mid \Lambda) \tag{式 2-54}$$

我們希望求出一組補償向量集合 $\mathbf{B}$，可以使得補償後的語音特徵向量 $\tilde{y}_t$ 落在聲學模型 Λ 中有最大的相似度，那麼可利用期望最大值演算法求得適當的 $\mathbf{B}$，補助函數(Auxiliary Function)定義成

$$\begin{aligned} Q &= \sum_t \sum_s \sum_m \zeta_t(s,m) \log N\left\{ y_t + \sum_{k=1}^{K} P(k \mid y_t) b_k ; \mu_{sm}, \Sigma_{sm} \right\} \\ &= \sum_t \sum_s \sum_m \zeta_t(s,m) \Sigma_{sm}^{-1} \left(y_t + \sum_{k=1}^{K} P(k \mid y_t) b_k - \mu_{sm} \right)^2 + Const \end{aligned} \tag{式 2-55}$$

其中 s 代表聲學模型 Λ 中的第 s 個狀態，m 代表聲學模型 Λ 中某個狀態下的第 m 個高斯分布，μ_{sm} 與 Σ_{sm} 分表代別第 s 個狀態裡的第 m 個高斯分布的平均值與變異數，$\zeta_t(s,m)$ 為時間點 t 落在第 s 個狀態裡的第 m 個高斯分布的出現機率(Occupation Probability)。為了求得適當的補償向量，則可利用 b_k 對式 2-55 進行偏微分令其為零求得。如果補償方法是利用式 2-50，那麼微分後可得到下列聯立方程式

$$\begin{aligned} &\sum_t \sum_s \sum_m \sum_{k'} \zeta_t(s,m) \Sigma_{sm}^{-1} P(k \mid y_t) P(k' \mid y_t) b_{k'} \\ &= \sum_t \sum_s \sum_m \zeta_t(s,m) \Sigma_{sm}^{-1} P(k \mid y_t) (\mu_{sm} - y_t) \end{aligned} \tag{式 2-56}$$

因為上式的 $b_{k'}$ 是對補償向量集合 $\mathbf{B}$ 中所有的 b_k，因此最後補償向量可經由下式

求得。

$$\mathrm{B}_d = \mathrm{A}_d^{-1}\mathrm{C}_d \qquad \text{(式 2-57)}$$

其中 $\mathrm{B}_d = [b_{1d},\cdots,b_{Kd}]^T$，$d$ 為補償向量中的第 d 維，A_d 與 C_d 分別代表 $K\times K$ 與 $K\times 1$ 的矩陣，而 A_d 與 C_d 內的每個元素(Element)分別表示如下：

$$a_d^{(e)}(k,k') = \sum_t \left[\sum_s \sum_m \frac{\zeta_t(s,m)}{\sigma_{smd}^2} P(k \mid y_t) P(k' \mid y_t) \right] \qquad \text{(式 2-58)}$$

$$c_d^{(e)}(k) = \sum_t \left[\sum_s \sum_m \frac{\zeta_t(s,m)(\mu_{smd} - y_{td})}{\sigma_{smd}^2} P(k \mid y_t, e) \right] \qquad \text{(式 2-59)}$$

不過此方法最大的缺點就是需要計算 A_d 的反矩陣(Inverse Matrix)，通常需耗費龐大的計算運算量。因此如果補償方法是利用式 2-52 的話，那麼 b_k 則可直接利用下式估測求得：

$$b_{kd}^{(e)} = \frac{\sum_t \sum_s \sum_m \frac{1[k = \arg\max_{k'} p(k' \mid y_t, e)]\zeta_t(s,m)(\mu_{smd} - y_{td})}{\sigma_{smd}^2}}{\sum_t \sum_s \sum_m \frac{1[k = \arg\max_{k'} p(k' \mid y_t, e)]\zeta_t(s,m)}{\sigma_{smd}^2}} \qquad \text{(式 2-60)}$$

另一方面，如果隨機特徵向量對映法的補償向量的估測是以鑑別力為考量出發點[Droppo and Acero 2005]，並且以最大交互訊息(Maximum Mutual Information, MMI)為鑑別準則，那麼目標函數可定義成：

$$F_{objective} = \sum_r \frac{p(F_r(\hat{y}, \mathbf{B}), w_r)}{\sum_{w'} p(F_r(\hat{y}, \mathbf{B}), w')} \qquad \text{(式 2-61)}$$

B 為補償向量集合，r 為訓練語料中第 r 句訓練語料，w_r 為第 r 句訓練語料對應的正確轉譯文字，w' 為對於第 r 句訓練語料的所有可能辨識結果，因為上式沒有辦法找到一個完全解(Closed Form Solution)，所以必須使用梯度下降法(Gradient

Descent)求得，因此用 Θ 對式 2-61 微分求斜率(gradient)以求得補償向量的更新值。

$$\frac{\partial F_{objective}}{\partial \Theta} = \sum_{r,t,s_t^r} \frac{\partial F_{objective}}{\partial \ln p\left(\hat{y}_t^r \mid s_t^r\right)} \frac{\partial \ln p\left(\hat{y}_t^r \mid s_t^r\right)}{\partial \hat{y}_t^r} \frac{\partial \hat{y}_t^r}{\partial \Theta} \qquad \text{(式 2-62)}$$

$\ln p\left(y_t^r \mid s_t^r\right)$ 為第 r 句訓練語料中第 t 個音框的語音特徵向量 $\hat{y}_t^r$ 落在聲學模型中 s 狀態的對數相似度，經過推導整理後，斜率可經由下式求得

$$\frac{\partial F_{objective}}{\partial b_k} = \sum_{r,t,s_t^r,m} P\left(k \mid y_t^r\right)\left(\gamma_{s_t(m)}^{num,r} - \gamma_{s_t(m)}^{den,r}\right)\Sigma_{s_t(m)}^{-1}\left(\mu_{s_t(m)} - \hat{y}_t^r\right) \qquad \text{(式 2-63)}$$

$P\left(k \mid y_t^r\right)$ 表示在給定第 r 句訓練語句中第 t 個音框的語音特徵向量 $\hat{y}_t^r$ 時，發生在高斯混合模型中第 k 個高斯分布的事後機率，$\gamma_{s_t(m)}^{num,r}$ 與 $\gamma_{s_t(m)}^{den,r}$ 分別代表在時間點 t 狀態 s 中的第 m 個高斯分布中正確轉譯文字與所有辨識結果的事後機率，$\Sigma_{s_t^r(m)}^{-1}$ 與 $\mu_{s_t^r(m)}$ 分別代表狀態 s 中的第 m 個高斯分布的共變異矩陣與平均值向量。在求得斜率之後，即可使用共軛梯度法(Conjugate Gradient Method)或 BFGS 更新(Broydon-Fletcher-Goldfarb-Shanno Update)等方法進行補償向量的更新[Droppo et al. 2005]。

2.3.3 語音特徵參數重建法(Feature Reconstruction)

對於語音特徵參數重建法，吾人在此主要是介紹遺失特徵理論(Missing Feature Theory, MST)。近年來分別有研究學者分別探討遺失特徵理論作用在前端語音特徵擷取[Raj et al. 2004; Raj and Stern 2005]或是後端語音解碼[Cooke et al. 2001; Hamme 2004]上。不論是做在前端特徵擷取或後端語言解碼，基於遺失特徵理論的語音特徵參數重建法基本上都包含二個步驟：第一步是決定語音特徵向量中哪些特徵參數是可靠(Reliable)，哪些是不可靠(Unreliable)或遺失(Missing)的，常見

的作法包括估測每個音框中頻譜的訊噪比[Vizinho et al. 1999]、利用語音訊號能量強度當為準則[EL-Maliki and Drygajlo 1999]、利用貝氏分類器(Bayesian Classifier)進行判別[Raj et al. 2004]或是利用一些聽覺感知(Perceptual)特性或音韻(Prosodic)資訊進行判別[Barker et al. 2001; Palomaki et al. 2004]；第二步則是針對不可靠的語音特徵參數進行參數重建，下列將分述如何將基於遺失特徵理論的語音特徵參數重建法應用於在語音辨識系統的前端特徵擷取或後端語言解碼模組中。

2.3.3.1 遺失特徵重建法作用在前端語音特徵擷取上

此法又可視為特徵向量設算(Feature Vector Imputation)，其最主要的目的是利用可靠的語音特徵參數(區域)重建不可靠語音特徵參數(區域)的特徵值，重建的方法主要有二種，分別為(1)以群集為基礎重建方法(Cluster-based Reconstruction)及(2)以共變異為基礎重建方法(Covariance-based Reconstruction)[Raj 2000; Raj et al. 2004]，且此二種法方皆是基於最大事後機率估測法則求算。

(1) 以群集為基礎重建方法

在以群集為基礎重建方法中，假設聲譜圖(Spectrogram)中的每個頻譜語音特徵向量(Spectral Vector) X 彼此是獨立(Independent)的，且整個乾淨語料的頻譜語音特徵向量的分布可以用一個高斯混合模型表示。

$$P(X)=\sum_{v} c_v \times \frac{1}{(2\pi)^{d/2}\left|\Sigma_v\right|^{1/2}} \times \exp\left(-\frac{1}{2}(X-\mu_v)^T \Sigma_v^{-1}(X-\mu_v)\right) \qquad \text{(式 2-64)}$$

其中 d 為語音特徵向量的維度個數，c_v 、Σ_v 和 μ_v 分別為高斯混合模型中第 v 個高斯分布所對應的混合權重、平均值向量與共變異矩陣。假設 Y 為受雜訊干擾的頻譜語音特徵向量，且有部份語音特徵參數是不可靠的，所以可將 Y 分解成二個子向量 Y_r (未受干擾的頻譜語音特徵向量)與 Y_u (受干擾的頻譜語音特徵向量)，分

別代表可靠與不可靠的語音特徵向量，且更進一步假設 X 為將 Y 進行重建後所得到的語音特徵頻譜向量，相同地 X 亦可被分解成 X_r 與 X_u 二部份，理論上，因為雜訊干擾語音訊號在頻譜上的表現為一線性加成的關係，因此可假設 Y_r 與 X_r 是相同的，X_u 是未知的但是必定小於等於 Y_u，因此對於不可靠的向量的估測可利用下式求得：

$$\hat{X}_u = \sum_v P(v \mid X_r, X_u \le Y_u) BMAP(X_u \mid X_r, X_u \le Y_u; \mu_v, \Sigma_v) \quad \text{(式 2-65)}$$

$P(v \mid X_r, X_u \le Y_u)$ 為給定受雜訊干擾的頻譜語音特徵向量 Y，其發生在第 v 個高斯分布的事後機率，計算方法如下：

$$P(v \mid X_r, X_u \le Y_u) = \frac{c_v p(X_r, X_u \le Y_u \mid v)}{\sum_j c_j p(X_r, X_u \le Y_u \mid j)} \quad \text{(式 2-66)}$$

但 $p(X_r, X_u \le Y_u \mid v) = p(X_r, -\infty \le X_u \le Y_u; \mu_v, \Sigma_v)$，因此無法以傳統求算機率密度函數的方法進行求算，所以必須配合邊際機率密度函數計算(Marginal Probability Density)進行計算，並且在此假設語音特徵向量中每一維度彼此是獨立的，故 $P(X_r, -\infty \le X_u \le Y_u; \mu_v, \Sigma_v)$ 的求算方式如下：

$$p(X_r, -\infty \le X_u \le Y_u; \mu_v, \Sigma_v) = \prod_{j \in X_r} \frac{1}{\sqrt{2\pi\sigma_v(j)}} e^{\left(\frac{-(X_r(j)-\mu_v(j))^2}{2\sigma_v(j)} \right)} \times \prod_{l \in Y_u} \int_{-\infty}^{Y_u(l)} \frac{1}{\sqrt{2\pi\sigma_v(l)}} e^{\left(\frac{-(x-\mu_v(l))^2}{2\sigma_v(l)} \right)} dx$$

(式 2-67)

而 $BMAP(X_u \mid X_r, X_u \le Y_u; \mu_v, \Sigma_v)$ 的計算可利用下面迭代步驟求得[Raj 2000]：

步驟 1：初始化 $\overline{X}(k) = Y(k) \forall k$，$k$ 為語音特徵向量中每一維的特徵索引值

步驟 2：對於 X_u 中的每個元素 $X_u(i)$ 進行計算，找出 $X_u(i)$ 的設算值

$$\tilde{X}(k) = \arg\max_{X(k)} P\left(X(k) \mid X(i) = \ddot{X}(i), \sigma_v(i), \mu_v(\mathrm{i}), i \neq k\right)$$
$$= \mu_v(i) + \frac{1}{\sigma_v(i)} \Theta_{X(k),\ddot{X}}\left(\ddot{X} - \bar{\mu}\right) \quad \text{(式 2-68)}$$

$$\ddot{X}(k) = \min\left(\tilde{X}(k), Y(k)\right)$$

$\mu_v(i)$與$\sigma_v(i)$為第v個高斯分布第i維的平均數與變異數，$\bar{\mu}$為所對應的平均向量，$\ddot{X}$為不包含i的所有$X(k)$組合，$\Theta_{X(i),\ddot{X}}$為一列矩陣(Raw Matrix)表示$X(i)$與$\ddot{X}$的交叉共變異數(Cross Covariance)。

步驟 3：重覆執行步驟 2 直到收斂為止

(2) 以共變異為基礎重建方法

而在以以共變異為基礎重建方法中，主要是假設對數頻譜向量(Log-Spectral Vectors)是由穩態高斯隨機程序(Stationary Gaussian Random Process)所產生的，頻譜上第m個頻譜語音特徵向量中的第k個語音特徵向量元素表示成$Y(m,k)$，因為穩態高斯隨機程序，對任何時間而言，所有頻譜圖中的第k個向量元素的平均值均會為$\mu(k)$，且$Y(m,k_1)$與$Y(m+\xi,k_2)$的共變異數表示成$c(\xi,k_1,k_2)$，相關係數為$r(\xi,k_1,k_2)$，各別變數計算方法定義如下[Raj 2000]：

$$\mu_k = \mathbf{E}[X(m,k)] \quad \text{(式 2-69)}$$

$$c(\xi,k_1,k_2) = \mathbf{E}\left[\left(Y(m,k_1) - \mu_{k_1}\right)\left(Y(m+\xi,k_2) - \mu_{k_2}\right)\right] \quad \text{(式 2-70)}$$

$$r(\xi,k_1,k_2) = \frac{c(\xi,k_1,k_2)}{\sqrt{c(\xi,k_1,k_1) \times c(\xi,k_2,k_2)}} \quad \text{(式 2-71)}$$

其中$\mathbf{E}[\cdot]$表示取期望值的意思。對於第m個頻譜向量$Y(m)$而言，所有不可靠元素可以表示成$Y_u(m)$，並且收集頻譜圖中所有的可靠頻譜向量元素與$Y_u(m)$中任一個頻譜向量元素的相關係數大於 0.5 的元素形成一組相鄰向量(Neighborhood Vector)$Y_n(m)$，假設$X(m)$為$Y(m)$重建後所得到的特徵值，那麼相同地$X(m)$亦可

被拆解成$X_r(m)$與$X_u(m)$二部份，最後不可靠的向量$X_u(m)$就可利用和前面敘述相似的方法$BMAP(X_u(m)|Y_n(m),Y_u(m)\le X_u(m);\mu,\Sigma)$求得。

2.3.3.2 遺失特徵重建法作用在後端語音解碼上

遺失特徵重建法除了可作用在前端語音特徵擷取上，亦可作用於後端語音解碼上，常見的方法亦有二種：資料設算法(Data imputation)與邊際化法(Marginalization)。

在資料設算法中，假設隱藏式馬可夫模型中的每個狀態都是用一組高斯混合模型表示，因此在給定任意頻譜向量$X(t)$下，$X(t)$可分解成二個子向量$X_r(t)$與$X_u(t)$，所以$X(t)$落在隱藏式馬可夫模型中某個狀態s的相似度計算方法如下：

$$P(X(t)|s)=P(X_r(t),X_u(t)|s)=\sum_v c_{s,v}G(X_r(t),X_u(t);\mu_{s,v},\Sigma_{s,v}) \qquad \text{(式 2-72)}$$

因為$X_u(t)$是屬於不可靠的向量，所以並不能拿來直接計算落在某個高斯分布的相似度，因此與上述方法一樣，必須重建向量$X_u(t)$，可行的做法包括前面章節所描述的群集為基礎的重建方法進行重建，或是利用最小平方差錯誤法進行重建[Josifovski et al. 1999]，計算方法如下：

$$\hat{X}_u^s(t)=\sum_v \gamma_{s,v}(X(t))U(t)\mu_{s,v} \qquad \text{(式 2-73)}$$

其中$U(t)$表示排列矩陣(Permutation Matrix)，$\gamma_{s,v}(X(t))$的計算方法如下

$$\gamma_{s,v}(X(t))=\frac{c_{s,v}\int_{-\infty}^{X_u(t)}G(X_r(t),Y_u(t);\mu_{s,v},\Sigma_{s,v})dY_u(t)}{\sum_k c_{s,k}\int_{-\infty}^{X_u(t)}G(X_r(t),Y_u(t);\mu_{s,k},\Sigma_{s,k})dY_u(t)} \qquad \text{(式 2-74)}$$

在邊際化法中，並不用重建任何向量，而是直接利用可靠與不可靠的頻譜向量進行相似度估測，計算方式如下：

$$P(X(t)|s)=\sum_{v} c_{s,v} P\left(X_r(t), -\infty \le X_u(t) \le Y_u(t); \mu_{s,v}, \Sigma_{s,v}\right) \quad \text{(式 2-75)}$$

$P\left(X_r(t), -\infty \le X_u(t) \le Y_u(t); \mu_{s,v}, \Sigma_{s,v}\right)$的計算方法與式 2-67 相同。但就目前大多數將遺失特徵重建法作用在後端語音解碼上，都存在一個限制，只能作用在頻譜特徵上，因為就目前的方法而言，語音特徵向量中特徵參數的可靠與否的判定，大多只能作用在頻譜上，主要原因是因為雜訊干擾語音訊號的表現，可能只在某些頻段上，所以可用一些方法進行可靠與否的判定，例如訊噪比等；但是若轉換至倒頻譜上，倒頻譜中的每一維特徵參數會含有數個頻段的頻譜資訊，所以判定並非那麼容易。但是就一般而言，但倒頻譜語音特徵向量的辨識效能通常卻又較頻譜語音特徵向量好。因此若將遺失特徵重建法能作用在前端語音特徵擷取上，那麼在重建完所有不可靠的頻譜特徵後，可再將頻譜的資訊轉換至倒頻譜上，勢必辨識效能會較直接將遺失特徵重建法作用在後端語音解碼上來得好。

第三章 實驗語料庫與相關基礎實驗結果

本章節主要是介紹本論文中實驗語料庫與相關實驗設定。第一小節將介紹本論文所使用的實驗語料庫；第二小節將說明本論文所使用的相關實驗設定；第三小節介紹辨識效能的評估方式，最後呈現相關基礎實驗結果與觀察。

3.1 實驗語料庫

為了測試與驗證本論文所提出的方法是否對提升語音強健性有幫助，本論文所使用的實驗的語料庫為 Aurora-2 語料庫，Aurora-2 是由歐洲電信標準協會(European Telecommunications Standards Institute, ESTI)所發行的語料庫[Hirsch and Pearce 2000]，其本身為一套含有雜訊的連續英文數字語料庫，參與錄音計畫的語者，皆是美國成年人。其中雜訊包含八種來源不同的加成性噪音和二種不同特性的通道。加成性噪音包括機場(Airport)、人聲(Babble)、汽車(Car)、展覽會館(Exhibition)、餐廳(Restaurant)、地下鐵(Subway)、街道(Street)及火車站(Train Station)，且依不同訊噪比(Signal-to-Noise Ratio, SNR)各自加入乾淨語音裡，訊噪比包括 20dB、15dB、10dB、5dB、0dB 和-5dB；通道效應包含由國際電信聯合會(International Telecommunication Union, ITU)所訂立的二個標準-G.712 和 MIRS。

根據測試語料中加入之通道效應(Channel Effect)以及加成性噪音(Additive Noise)之類型不同，Aurora-2 共分為三組測試群組 Set A、Set B 和 Set C，並且提供二種不同的訓練模式：乾淨語料訓練(Clean-Condition Training)模式與複合情境訓練(Multi-Condition Training)模式，詳細內內容如表 3-1 所示。

3.2 實驗設定

在前端處理方面，本論文的基礎實驗是採用梅爾倒頻譜係數作為語音特徵參數，

表 3-1 Aurora 2.0 語料庫詳細說明

<table>
<tr><td>取樣頻率</td><td colspan="3">8KHz</td></tr>
<tr><td>語音內容</td><td colspan="3">包含英文數字單詞：One、Two、Three、Four、Five、Six、Seven、Eight、Nine、Zero、Oh</td></tr>
<tr><td rowspan="2">訓練模式</td><td>乾淨語料訓練</td><td colspan="2">複合情境訓練</td></tr>
<tr><td>語句個數：8440 句
加成性噪音：無
訊噪比範圍：完全乾淨
通道效應：G.712</td><td colspan="2">語句個數：8440 句
加成性噪音：
地下鐵、人聲、汽車、展覽會館
訊噪比範圍：完全乾淨及 20dB 至 5dB
通道效應：G.712</td></tr>
<tr><td rowspan="2">測試組合</td><td>Set A</td><td>Set B</td><td>Set C</td></tr>
<tr><td>語句個數：28028 句
加成性噪音：地下鐵、人聲、汽車、展覽會館
訊噪比範圍：完全乾淨及 20dB 至-5dB
通道效應：G.712</td><td>語句個數：28028 句
加成性噪音：餐廳、街道、機場、火車
訊噪比範圍：完全乾淨及 20dB 至-5dB
通道效應：G.712</td><td>語句個數：14014 句
加成性噪音：地下鐵、街道
訊噪比範圍：完全乾淨及 20dB 至-5dB
通道效應：MIRS</td></tr>
</table>

預強調參數α設為 0.975，漢明窗參數β設為 0.46，取樣音框長度(Frame Length)為 25 毫秒，音框間距(Frame Shift)為 10 毫秒，每個音框的資訊是以 39 維語音特徵向量表示， 其中包含 12 維的梅爾倒頻譜係數以及一維的對數能量(Log Energy)，同時並對 13 維語音特徵參數取其相對的一階差量係數(Delta Coefficient)和二階差量係數(Acceleration Coefficient)。

在聲學模型的設定，每個數字模型(1~9 及 Zero 和 Oh)皆由一個由左到右(Left-to-Right)形式的連續密度隱藏式馬可夫模型(CDHMM)表示，其中包含 16 個狀態(State)，並且每個狀態是利用 3 個高斯分布的高斯混合模型表示。另外靜音模型的部份有二種模型，一個為靜音(Silence)模型包含 3 個狀態，每個狀態用

6 個高斯分布的高斯混合模型，用來表示語句開始跟結束時的靜音；另一個為間歇(Pause)模型包含 1 個狀態，以 6 個高斯分布的高斯混合模型建模，表示語句內數字與數字之間的短暫停止，上述所有聲學模型的訓練與本論文所有的實驗都是使用 HTK 工具套件完成[Young et al. 2006]。

3.3 辨識效能評估方式

辨識效能評估的方式是採用美國標準與科技組織(The National Institute of Standards and Technology，NIST)所訂立的評估標準，進行正確轉譯文句字串與辨識字串的比較。評估單位是以字正確率(Word Accuracy)為單位，計算正確轉譯文句字串與辨識字串間的字取代個數(Substitutions)、字插入個數(Insertions)和字刪除個數(Deletions)；計算的方式有二種，字正確率(Word Accuracy Rate)與字錯誤率(Word Error Rate)，分別如下所示：

$$\text{字正確率}(\%) = \frac{\text{字正確辨識個數} - \text{字插入個數}}{\text{輸入字總數}} \times 100\%$$

$$\text{字錯誤率}(\%) = \frac{\text{字取代個數} + \text{字插入個數} + \text{字刪除個數}}{\text{輸入字總數}} \times 100\%$$

因為 Aurora-2 的語料庫裡，每一種噪音對於同一個測試集都會用七種不同程度的訊噪比添加，依照國際學者對數據呈現的習慣，對於每一種噪音的平均字正確率或平均字錯誤率的計算方式是加總 20dB 至 0dB 的辨識結果取平均，排除掉乾淨和-5dB 二種極端的訊噪比，所以後續本論文的所有實驗結果亦是遵循此種呈現方式。

3.4 基礎實驗結果

首先吾人先以梅爾倒頻譜係數(MFCC)(設定如 3.2 節所描述)當作語音特徵參數，求算其在各種不同雜訊與通道效應下的辨識結果，當作本論文的基礎實驗結

表 3-2 使用梅爾倒頻譜係數(MFCC)於乾淨語料訓練模式與複合情境訓練模式下的辨識結果

乾淨語料訓練模式										
平均字錯誤率(%	測試集A				測試集B				測試集C	
訊噪比	地下鐵	人聲	汽車	展覽會館	餐廳	街道	機場	火車站	地下鐵	街道
Clean	1.01	1.00	1.13	0.89	1.01	1.00	1.13	0.89	0.80	0.91
20dB	4.70	9.37	4.18	4.81	7.37	4.96	6.83	4.35	12.25	8.28
15dB	12.65	25.85	13.93	10.46	20.39	14.33	17.86	13.27	21.43	15.21
10dB	31.35	48.55	35.79	26.87	40.80	35.55	39.64	34.53	37.83	31.29
5dB	60.24	71.25	66.00	54.89	65.92	62.21	65.02	65.13	61.77	53.99
0dB	85.57	85.97	86.46	82.35	85.63	79.53	81.99	85.04	82.41	75.82
-5dB	92.05	92.20	92.07	90.96	92.60	90.11	90.96	91.92	89.87	86.85
平均	38.90	48.20	41.27	35.88	44.02	39.32	42.27	40.46	43.14	36.92

複合情境訓練模式										
平均字錯誤率(%	測試集A				測試集B				測試集C	
訊噪比	地下鐵	人聲	汽車	展覽會館	餐廳	街道	機場	火車站	地下鐵	街道
Clean	1.78	1.66	1.52	1.64	1.78	1.66	1.52	1.64	1.63	1.72
20dB	3.72	3.42	2.45	2.87	4.36	2.96	2.45	3.80	3.41	3.81
15dB	5.50	5.47	2.98	4.29	8.07	4.72	3.94	5.92	5.22	5.32
10dB	9.24	9.22	5.10	6.97	12.93	7.44	6.74	8.02	8.17	8.59
5dB	17.47	19.29	13.12	13.64	23.24	19.47	14.46	17.86	22.57	22.55
0dB	41.39	43.11	45.72	40.70	46.24	45.89	35.88	45.88	59.10	54.56
-5dB	77.16	76.57	81.99	79.05	78.14	78.93	71.01	81.86	85.39	82.41
平均	15.46	16.10	13.87	13.69	18.97	16.10	12.69	16.30	19.69	18.97

果(Baseline)。表 3-2 階分別呈現於乾淨語料訓練模式與複合情境訓練模式的辨識結果，在乾淨語料訓練模式下的平均字錯誤約莫是 41.03%，在複合情境訓練模式下的總平均字正確率約莫是 16.18%，從二種不同訓練模式的數據呈現，吾人總結下列數點觀察到的現象：

(1) 在不受任何雜訊干擾時，字正確率可高達 98% 甚至是 99%，然而隨著雜訊干擾程度愈來愈大，辨識效能會下降非常的快速，尤其當訊噪比低於 5dB 時，下降的程度更為明顯。

(2) 複合情境訓練模式的辨識效能較乾淨語料訓練模式好，是因為複合情境訓練模式是收集許多受不同訊噪比干擾的語料，加以訓練聲學模型，因此使得測試語料與聲學模型間的不匹配問題降低，所以才能獲得較好的辨識效能。

(3) 在複合情境訓練模式下，測試集 A 的辨識效果比測試集 B 好，主要是因為測試集 B 的噪音型態是沒出現在複合情境訓練模式的訓練語料中，所

表 3-3 倒頻譜平均消去法(CMS)作用在梅爾倒頻譜係數上的辨識結果

乾淨語料訓練模式										
平均字錯誤率(%	測試集A				測試集B				測試集C	
訊噪比	地下鐵	人聲	汽車	展覽會館	餐廳	街道	機場	火車站	地下鐵	街道
Clean	0.92	0.85	0.98	0.62	0.92	0.85	0.98	0.62	0.86	0.73
20dB	3.65	2.21	2.80	3.55	2.09	2.81	2.21	1.76	4.08	2.96
15dB	8.93	5.62	6.77	9.16	4.67	6.59	4.62	5.31	10.04	8.37
10dB	25.18	16.90	22.04	24.04	14.58	20.95	11.51	16.08	27.30	23.79
5dB	53.09	44.29	55.53	55.75	37.43	48.31	35.28	42.92	56.83	53.05
0dB	76.79	73.85	78.14	79.76	69.76	75.42	67.04	73.84	78.14	76.90
-5dB	86.28	86.61	86.31	88.98	85.39	87.15	82.76	85.90	87.14	87.39
平均	33.53	28.57	33.06	34.45	25.71	30.82	24.13	27.98	35.28	33.01

表 3-4 頻譜正規化法(CMVN)作用在梅爾倒頻譜係數上的辨識結果

乾淨語料訓練模式										
平均字錯誤率(%	測試集A				測試集B				測試集C	
訊噪比	地下鐵	人聲	汽車	展覽會館	餐廳	街道	機場	火車站	地下鐵	街道
Clean	0.77	0.85	0.92	0.74	0.77	0.85	0.92	0.74	3.01	3.02
20dB	3.53	2.21	2.18	3.30	2.15	2.57	1.91	2.31	6.75	6.26
15dB	6.85	4.38	4.35	7.07	4.02	5.11	3.70	4.26	12.77	11.12
10dB	14.09	10.16	10.65	15.40	9.79	12.09	8.20	10.37	25.27	23.88
5dB	32.70	27.18	28.15	33.32	24.29	28.48	22.31	26.66	48.11	43.80
0dB	66.38	58.07	59.50	65.04	53.36	59.76	52.07	58.59	70.89	68.17
-5dB	88.06	84.16	84.79	86.12	81.64	84.76	81.33	83.49	86.37	84.64
平均	24.71	20.40	20.97	24.83	18.72	21.60	17.64	20.44	32.76	30.65

以可想而知辨識效能必較差。

(4) 測試集 C 的辨識結果一般而言，較測試集 A 與測試集 B 差，主要原因是因為測試集 C 的通道效應是和訓練語料不相同的。

由於複合情境訓練模式已經能大幅度降低測試語料與聲學模型間不匹配的問題，但就乾淨語料訓練模式而言，仍尚有許多可努力的空間，因此吾人在本論文後續章節將只探討乾淨語料訓練模式下的語音辨識。首先，吾人先討探倒頻譜平均消去法(CMS)與倒頻譜正規化法(CMVN)的辨識效能，實驗結果如表 3-3 與表 3-4 所示。從表中可清楚的發現倒頻譜平均消去法對於移除通道效應的影響有非常顯著的效果，此外，頻譜平均消去法亦對減緩一些加成性噪音所帶來的失真情形有所幫助；而倒頻譜正規化法除了與倒頻譜平均消去法一樣能移除通道效應的影響外，同時會對每一維語音特徵參數的分布變異做正規化，進而降低語音特徵參數各個維度間彼此分布的差異程度，因此可預期的，倒頻譜正規化的辨識效果會較倒頻譜平均消去法好。

表 3-5 使用不同統計圖組距數與不同表格記錄點數之查表式統計圖等化法(THEQ)於乾淨語料訓練模式的辨識結果

乾淨語料訓練模式									
平均字錯誤率(%)		表格記錄點數							
		10	50	100	500	1000	5000	10000	50000
統計圖組距數	100	41.32	45.65	46.39	44.55	44.59	44.65	44.67	44.65
	500	33.21	28.60	25.44	22.42	22.42	22.41	22.45	22.41
	1000	29.63	24.19	22.12	19.04	19.19	19.46	19.88	19.87
	5000	28.13	23.72	20.68	18.22	18.02	18.18	18.19	18.10
	10000	27.64	23.50	20.50	18.35	18.33	18.13	18.30	18.32
	50000	27.46	23.30	20.29	18.41	18.58	18.46	18.47	18.45

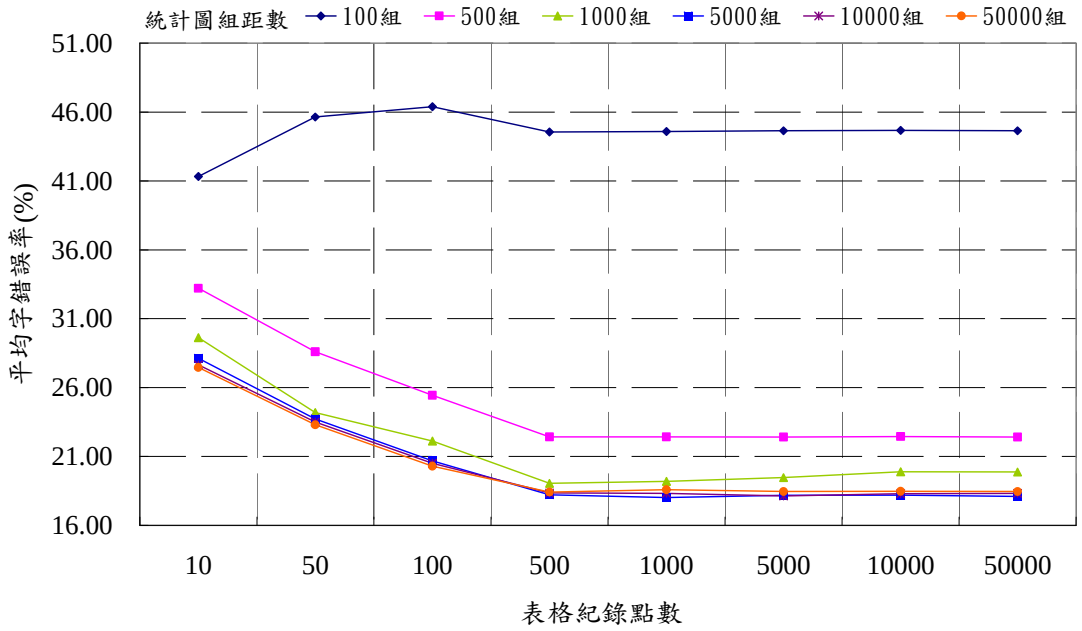


圖 3-1 使用不同統計圖組距數與不同表格記錄點數之查表式統計圖等化法於乾淨語料訓練模式的辨識結果比較圖

再者，吾人實作傳統查表式統計圖等化法(THEQ)，在查表式統計圖等化法中，辨識效能乃取決於二種參數的設定，分別是查表表格的記錄點數與統計圖中組距 (Histogram Bin)個數(參見 2.3.1 節)，因此吾人利用不同的表格紀錄點數與不同的統計圖組距個數進行實驗，檢視不同的設定環境下，辨識效能為何，整體辨識結果如表 3-5 所示。如同表格所呈現的數據，查表表格的記錄點數與統計圖中的組距個數影響辨識效能甚鉅，平均字錯誤率隨著表格記錄點數與統計圖組距個數增加而降低，相較於梅爾倒頻譜係數基礎實驗而言，在乾淨語料訓練模式下，可達平均字錯誤率 55%左右的相對減少(Relative Improvement)。從圖 3-1

表 3-6 使用不同分位差點數於分位差統計圖等化法(QHEQ)於乾淨語料訓練模式與複合情境訓練模式下的辨識結果

平均字錯誤率(%)	分位差個數						
訓練模式	2	3	4	5	8	16	32
乾淨語料訓練模式	24.02	23.67	22.86	23.00	24.93	24.83	24.95

表 3-7 使用不同分群數於雙聲源為基礎分段線性補償(SPLICE)在乾淨語料訓練模式的辨識結果

平均字錯誤率(%)	分群個數					
訓練模式	32	64	128	256	512	1024
乾淨語料訓練模式	27.64	24.31	21.89	21.03	20.52	19.04

中亦可清楚發現，若要得到良好的辨識效能，那麼查表表格的記錄點數與統計圖組距個數不可太少，此現象同時意謂著在執行等化的過程，需較多的記憶體空間供表格的存放與查表需花費額外較多的處理器運算時間。

接下來，吾人討論分位差統計圖等化法(QHEQ)的分位差個數對於辨識效能的影響情形，辨識結果如表 3-6 所示，隨著分位差個數不同，辨識效能也有所差異，當分位差點數使用太少，代表分位差統計圖等化法的轉換函數 $H(x)$ 會較粗糙，因此辨識效能較差；相反地，若分位差點數使用太多，使得轉換函數 $H(x)$ 太精細，反而會降低辨識效能，此現象與作者提出的論證相同[Hilger and Ney 2001]，因此分位差統計圖等化法中分位差點數的選擇需特別注意。以在 Aurora-2 語料庫上，吾人實驗結果是以 4 點分位差個數能獲得較好的辨識效能，相較於梅爾倒頻譜係數基礎實驗，在乾淨語料訓練模式下可達平均字錯誤率 44%左右的相對減少。

再者，吾人實作雙聲源為基礎分段線性補償，在 Aurora-2 語料庫中，雙聲源語料可從乾淨語料訓練模式的語音及其對應的複合情境訓練模式的語音而得，吾人嘗試使用 32、64、128、256、512、1024 個不同分群數的高斯混合模型，欲探討群集數的個數對於辨識效能影響程度為何，實驗結果如表 3-7 所示。從實

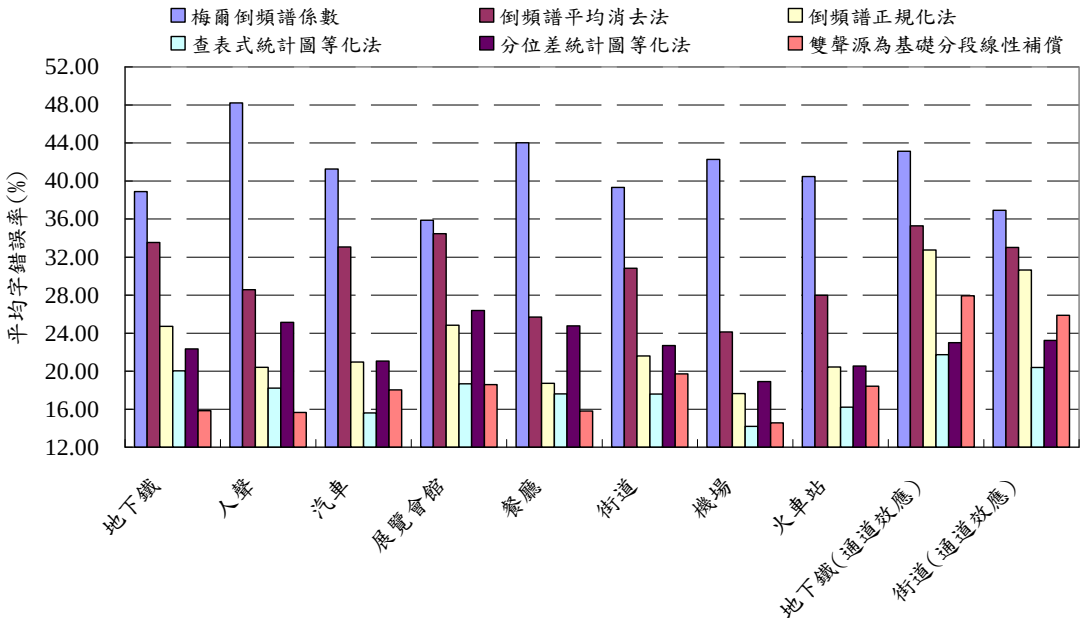


圖 3-2 不同強健性語音技術作用在 Aurora-2 語料庫的比較圖

驗表格可清楚發現，隨著分群數增加，辨識效能亦隨之提高，此舉意謂著較多的群集數對於雜訊干擾乾淨語音的情形能有更精細的描述。在使用 1024 個高斯分布的高斯混合模型下，辨識結果相較於梅爾倒頻譜係數基礎實驗結果，可達平均字錯誤率 53%左右的相對減少。

最後，吾人以圖 3-2 總結上述所有實驗，從圖中可發現即使是簡單的倒頻譜消去法(CMS)或倒頻譜正規化法(CMVN)皆達到不錯的辨識效能。而查表式統計圖等化法(THEQ)確實較倒頻譜消去與或倒頻譜正規化法好，因為統計圖等化法可視為是對語音特徵參數統計分布的每一動差進行正規化。分位差統計圖等化法(QHEQ)雖然對提升辨識效能有幫助，但效果相較於倒頻譜正規化法或查表式統計圖等化法較不明顯，主要原因是因為查表式統計圖等化法是利用語句的全部累積密度函數進行等法動作，而分位差統計圖等化法是只以數點分位差對累積密度函數做分位差校正(Quantile-Corrective)；此外，雙聲源為基礎分段線性補償(SPLICE)在測試集 A 與測試集 B 有較顯著的幫助，而在測試集 C 中，因為語音訊號是含有與訓練語料不同的通道效應，因此會使得在計算式 2-46 的過程中產生誤差，因此補償效能會較差，解決的方法之一是在計算式 2-46 前，可先利用

倒頻譜消去法將通道效應移除掉[Droppo et al. 2002]。

第四章 特徵參數補償法之相關改進

4.1 群集式為基礎之多項式擬合統計圖等化法

縱觀前面章節回顧，若依照語音特徵參數的處理層面與出發點不同，大致上目前的主要研究方向，可概分為二類：第一種是直接從雜訊語音特徵參數的特徵值域(Feature Domain)進行特徵參數補償或特徵參數轉換，此類研究方向通常假設有關於雜訊干擾語音訊號的先備知識(Prior Knowledge)或是假設乾淨語音特徵參數與雜訊語音特徵參數間存在著某種固定的關係，所以在訓練階段可以事先求算此一關係，而在測試階段利用此一關係進行補償，因此此種研究方向通常能有較佳的辨識效能。然而此種研究方向卻存在個潛在的隱藏問題，因為雜訊干擾語音訊號的影響並非絕對是一對一的線性關係，所以可能因某些非預期的因素影響，造成特徵參數補償或特徵參數轉換的效果不佳，或更進一步使得辨識效能驟降；另一種研究方向是利用一些較不容易受雜訊干擾而有所影響的語音特徵參數特徵值的統計分布特性(Distribution Characteristics)，當作是特徵參數補償或是特徵參數轉換的依據，通常此研究方向相較於前者只需額外的短暫運算時間即可獲得良好的辨識效能。但有些方法往往會事先對語音特徵參數的統計分布做一些限制或假設，例如假設語音特徵參數的分布是高斯分布等，然而此類假設並非完全正確，因此可能使得方法的補償效果有所侷限，此外，雜訊干擾語音訊號除了會產生統計特性不匹配的問題外，因其本身的隨機特性，同時也會對語音訊號加入了不確定性(Uncertainties)，而此研究方向只能有效處理統計特性不匹配的問題，卻無法解決由雜訊干擾所產生的不確定性問題。在此，吾人針對目前一些較廣為討論的語音強建技術做一分類圖，分類圖如圖 4-1 所示。有鑑於此，吾人嘗試結合二種研究方向的優點，結合語音特徵參數的統計分布特性與雙聲源語料進行語音特徵參數補償，此外為了能符合實際語音特徵參數的統計分布，吾人搭配多項式

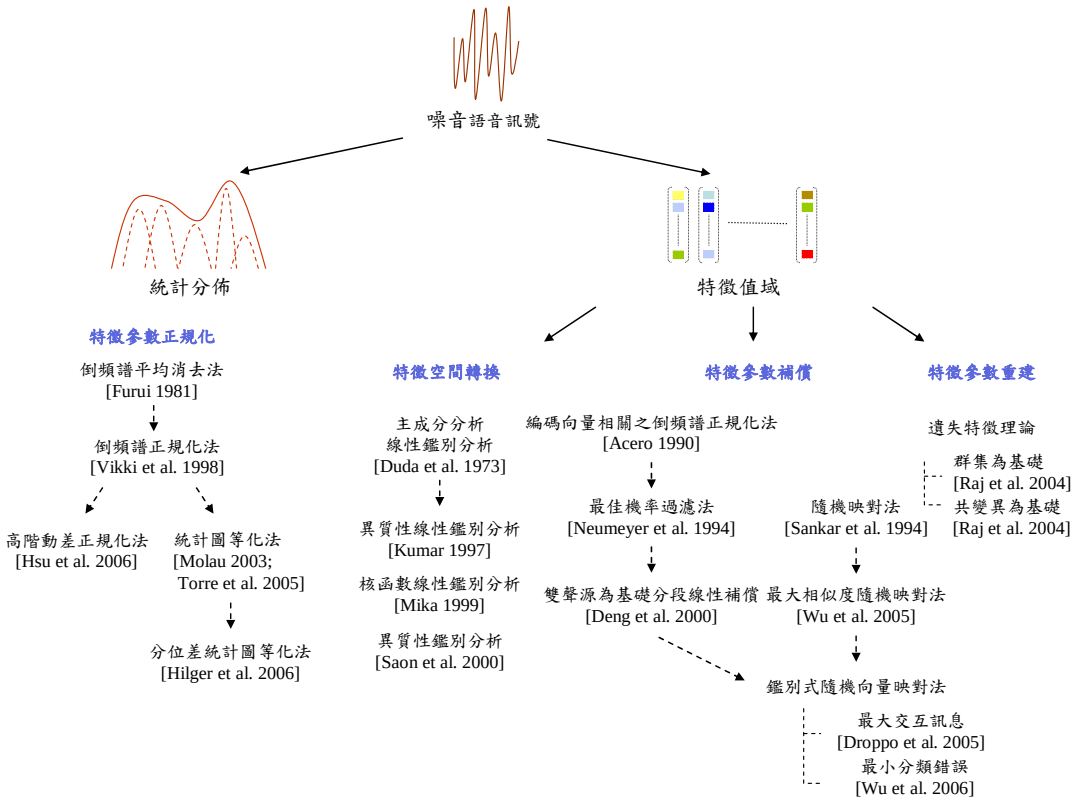


圖 4-1 語音特徵參數補償或轉換的研究方向分類圖

數據擬合(Polynomial Data Fitting)方法的使用，以資料導向(Data Driven)的方式近似實際的統計分布，因此吾人稱其為群集式為基礎之多項式擬合統計圖等化法(Cluster-based Polynomial-Fit Histogram Equalization, CPHEQ)[Lin et al. 2007a, 2007b]。

多項式數據擬合的精神是當給定一些資料點數(u_i, v_i)，若要以一個函數來描述反應變數(Response Variable)v_i與解釋變數(Explanatory Variable)u_i關係，通常可使用迴歸模型(Regression Model) $G(u_i)$來表示，換句話說迴歸模型可用來解釋在給定u_i的情況下，預測v_i的可能值為何。通常迴歸模型$G(u_i)$可依係數(Coefficients)組合不同而表示成線性或非線性型式，並且$G(u_i)$係數的選擇影響預測值$\tilde{v}_i$的準確性甚鉅，一般可利用最小誤差平方和 (Minimum Sum of Squares Error)求得，換言之，若將所有u_i分別代入迴歸模型所求得的預測值$\tilde{v}_i$和實際觀測值v_i的誤差值平方和為最小，此意謂著經由迴歸模型所預測出的值會跟實際的

值較相似，此法又可稱最小平方迴歸法(Least-Squares Regression)[Montgomery et al. 2006]。假設$G(u_i)$為一M階的線性多項式函數，那麼在給定u_i的情況下，v_i的預測值$\tilde{v}_i$可經由下列迴歸模型進行預測：

$$\tilde{v}_i = G(u_i) = a_0 + a_1 u_i + a_2 u_i^2 + \cdots + a_M u_i^M = \sum_{m=0}^{M} a_m u_i^m \qquad \text{(式 4-1)}$$

因此吾人利用與統計圖等化法相同的假設，假設測試語句之語音特徵參數的統計分布會和訓練語料特徵參數的統計分布一致，並且新的語音特徵參數可經由其對應的累積密度函數透過一轉換函數$G(\bullet)$求得。此外，為了解決傳統統計圖等化法利用查表進行轉換動作所需的記憶體空間與運算器處理時間的問題，吾人採用多項式轉換函數描述轉換函數$G(\bullet)$進行補償動作，且在此吾人假設語音特徵向量中，每一維語音特徵參數間彼此為獨立的，因此每一維語音特徵參數皆可獨立分開進行補償。此外為了方便表示，吾人重新定義符號Y_t為語音特徵向量及y_t代表某一維的語音特徵參數(省略維度的索引值)，因此對於每一維語音特徵參數的補償可經由用下式求得：

$$\hat{x}_t = G(CDF(x_t)) = \sum_{m=0}^{M} a_m (CDF(x_t))^m \qquad \text{(式 4-2)}$$

其中M為多項式轉換函數階數(Order)，a_m為多項式轉換函數的係數，$CDF(x_t)$為第t個音框的語音特徵參數所對應的累積密度函數，此外為了擷取住不同雜訊對語音訊號的干擾情形，吾人利用向量編碼技術對雙聲源語料中的雜訊語料訓練出一組高斯混合模型，先以K-分群演算法(K-Mean)[Alpaydin 2004]計算每個高斯分布參數的初始種子(Initial Seeds)，接著再以期望值最大化演算法，迭代數次以更新高斯混合模型內每個高斯分布的參數，最後，每個高斯分布可視為一個群集，且每個高斯分布可被視為是某些雜訊干擾某些音素的影響情形，高斯混合模型表示如下：

$$p(Y_t)=\sum_{k=1}^{K}P(k)p(Y_t\mid k)=\sum_{k=1}^{K}c_k N(Y_t;\mu_k,\Sigma_k) \qquad \text{(式 4-3)}$$

其中 K 為高斯混合模型中所有高斯分布的個數，$p(Y_t\mid k)$ 為雜訊語音特徵向量 Y_t 落在第 k 個高斯分布的相似度，$P(k)$ 為第 k 個高斯分布的事前機率。因此，吾人使用和雙聲源為基礎分段線性補償相同的概念(如式 2-43)，假設對於每一群集 k 裡的雜訊語音特徵參數 y_t 和乾淨語音特徵參數 x_t 間的關係式可經由式 4-2 進行補償，因此利用最小均方誤差的概念結合式 4-2，補償後的雜訊語音特徵參數 $\tilde{y}_t$ 可得下列關係式求得：

$$\begin{aligned}\tilde{y}_t=\hat{x}_t&=\mathbf{E}[x_t\mid Y_t]=\mathbf{E}[\mathbf{E}[x_t\mid Y_t,k]]=\sum_{k=1}^{K}P(k\mid Y_t)\mathbf{E}[x_t\mid Y_t,k]\\&=\sum_{k=1}^{K}\left(P(k\mid Y_t)\times\left(\sum_{m=0}^{M}a_{km}(CDF(y_t))^m\right)\right)\end{aligned} \qquad \text{(式 4-4)}$$

其中 $P(k\mid Y_t)$ 為給定雜訊語音特徵向量 Y_t 下，發生在第 k 個高斯分布的事後機率，且每一群集皆有一組對應的多項式轉換函數 $G_k(\bullet)$。對於每一群集 k 所對應的多項式轉換函數 $G_k(\bullet)$ 的係數 a_{km}，則可經由最小化下列均方誤差(Squares Error)而得：

$$E_k^2=\sum_{t=0}^{T-1}\left(\left(x_t-\sum_{m=0}^{M}a_{km}(CDF(y_t))^m\right)\times p(k\mid Y_t)\right)^2 \qquad \text{(式 4-5)}$$

其中 T 為所有訓練語料的音框個數，x_t 為雙聲源語料中乾淨語料所擷取出的語音特徵參數，y_t 為雙聲源語料中雜訊語料所擷取出的語音特徵參數，而 $CDF(y_t)$ 則是 y_t 所對應的累積密度函數值。因此只需利用 a_{km} 對式 4-4 做偏微分令其為零，即可透過解聯立方程式求得每個多項式轉換函數的係數。

但由式 4-4 可看出，新的語音特徵參數的求得必須將 $CDF(y_i)$ 代入所有每一群集 k 的多項式轉換函數 $G_k(\bullet)$，然後再將多項式轉換函數輸出的特徵值乘上給

定Y_t，會發生在群集k的事後機率，因此，在實作上，當分群數變多，所需的處理器運算時間會隨之增加，因此吾人利用最大事後機率的概念，重新定義式 4-4 成如下：

$$\begin{aligned}\tilde{y}_t = \hat{x}_t &= \sum_{k=1}^{K} \delta(k \mid Y_t) G_k\left(CDF(y_t)\right) \\ &= \sum_{k=1}^{K} \left(\left(\sum_{m=0}^{M} a_{km} \left(CDF(y_t)\right)^m \right) \times \delta(k \mid Y_t) \right)\end{aligned} \qquad \text{(式 4-6)}$$

且

$$\delta(k \mid Y_t) = \begin{cases} 1 & \text{if } k = \arg\max_{k'} p(k' \mid Y_t) \\ 0 & \text{otherwise} \end{cases} \qquad \text{(式 4-7)}$$

其中$\delta(k \mid Y_t)$為克羅內克函數(Kronecker Delta Function)，判斷的依據是計算在給定給定雜訊語音特徵向量Y_t下，發生在不同高斯分佈的事後機率，唯具有最大事後機率的高斯分布k設為 1，其於的皆設為 0。因此實作上對於每個雜訊語音特徵向量Y_t而言，需先計算在給定Y_t情況下，找出其發生在某個高斯分布下的事後機為最大的高斯分佈，再將該音框裡每一維特徵參數所對應的累積密度函數帶至該高斯分布所對應的多項式轉換函數$G_k(\bullet)$即可，而非像式 4-4 需將累積密度函數代入至每個高斯分布所對應的多項式轉換函數。因此式 4-6 多項式轉換函數的係數估測方式如下：

$$E_k^2 = \sum_{t=0}^{T-1} \left(\left(x_t - \sum_{m=0}^{M} a_{km} CDF(y_t) \right) \times \delta(k \mid Y_t) \right)^2 \qquad \text{(式 4-8)}$$

然而式 4-5 與式 4-8 最大的差異只是在於群集指派(Cluster Assignment)的方式不同，式 4-5 屬於軟性指派(Soft Assignment)，每個雙聲源訓練語料樣本對於每個群集的均方誤差皆有貢獻，貢獻程度取決於該雜訊語音特徵落在對應群集的事後機率，而式 4-8 屬於硬性指派(Hard Assignment)，因為每一個訓練樣本只會單單

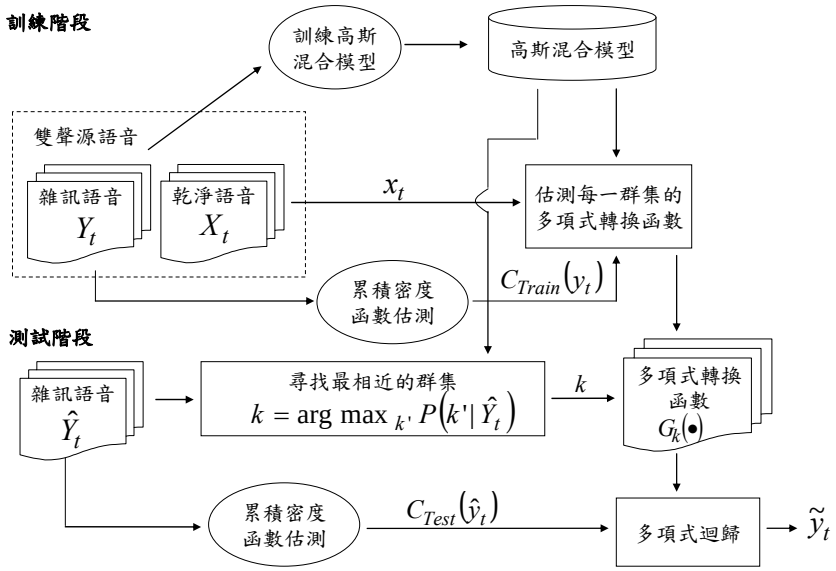


圖 4-2 群集式為基礎之多項式擬合統計圖等化法的流程圖

地落在某一群集內，對於其他群集並不會有任何影響，最後群集式為基礎之多項式擬合統計圖等化法的整體實作流程如圖 4-2 所示。

本論文所提出的方法與以雙聲源為基礎之分段線性補償(SPLICE)的不同之處可由幾點不同層面進行探討：

(1) 在以雙聲源為基礎之分段線性補償中，假設每一群集內所有的乾淨語音特徵與雜訊語音特徵間的差異是固定的，並且只利用一線性補償向量(Linear Compensation Vector)進行補償，然而實際上，縱使二個不同的音框被分到同一群集內，其受影響的程度亦會是不相同的，乾淨語音特徵與雜訊語音特徵並非是線性關係的，因此吾人採用非線性的多項式轉換函數來描述此非線性補償的關係。

(2) 雙聲源為基礎之分段線性補償直接對語音特徵向量進行補償，而本論文所提出的方法除了利用語音特徵向量做補償的依據外，更結合了不易受雜訊干擾的統計分布特性－累積密度函數達到補償的作用，因此，本論文所提出的方法，勢必會比以雙聲源為基礎之分段線性補償更具強健性。

表 4-1 群集式為基礎之多項式擬合統計圖等化法中使用硬性指派與軟性指派的辨識結果

乾淨語料訓練模式						
平均字錯誤率(%)	分 群 個 數					
指派方式	32	64	128	256	512	1024
硬性指派	19.73	19.35	18.19	17.27	16.36	15.41
軟性指派	19.77	19.34	18.19	17.24	16.33	15.40

4.2 群集式為基礎之多項式擬合統計圖等化法相關實驗結果

首先，吾人先探討硬性指派與軟性指派的對於辨識效能的影響程度，吾人嘗試使用不同群集數的高斯混合模型，包括 32 個、64 個、128 個、256 個、512 個與 1024 個，而多項式轉換函數的階數初步設為 3 階，實驗結果如表 4-1 所示，由表格可清楚發現隨著高斯混合模型內的高斯分布個數增多，平均字錯誤率會隨之下降，在分群數達 1024 個群集時，平均字錯誤率達 15.41%左右，相較於梅爾倒頻譜係數(MFCC)基礎實驗結果，平均字錯誤率達 62%左右的相對減少。此外，從表中亦可看出硬性指派與軟性指派的辨識結果並無任何顯著差異，主要的原因可能是因為在式 4-5 中，需計算在給定雜訊語音特徵參數 y_t 下，其發生在第 k 個高斯分布的事後機率，而此事後機率有可能只會被某一個高斯分布所支配著(Dominate)，相對其他 $K-1$ 個高斯分布的事後機率都會變得很小，也就意謂著該訓練樣本對於其他 $K-1$ 個高斯分布的誤差貢獻幾近可忽略，所以式 4-5 與式 4-8 的效用大致相同，因此在實作上，只需利用硬性指派即可，以便降低處理器運算時間。

下一個實驗吾人採用硬性指派的方式，觀察使用不同階數的多項式轉換函數與不同群集個數的高斯分布混合模型對於辨識效能的影響情形，欲探討此二種因素與群集式為基礎之多項式擬合統計圖等化法間的關係，實驗結果如表 4-2 所示，且不同設定下的辨識結果比較折線圖如圖 4-3 所示。由圖 4-3 中可看到，隨著分群數增加，平均字錯誤率會隨著降低，在分群數較少時，高階的多項式轉換

表 4-2 群集式為基礎之多項式擬合統計圖等化法中使用不同分群數與搭配不同多項式階數的辨識結果

乾淨語料訓練模式							
平均字錯誤率(%)		多項式階數					
		1	2	3	4	5	6
分群個數	32	20.70	20.36	19.84	19.80	19.75	19.73
	64	20.29	20.00	19.49	19.44	19.37	19.35
	128	18.95	18.71	18.24	18.21	18.20	18.19
	256	17.82	17.60	17.33	17.32	17.27	17.43
	512	16.84	16.58	16.36	16.40	16.53	16.84
	1024	15.69	15.57	15.41	15.62	16.04	17.14

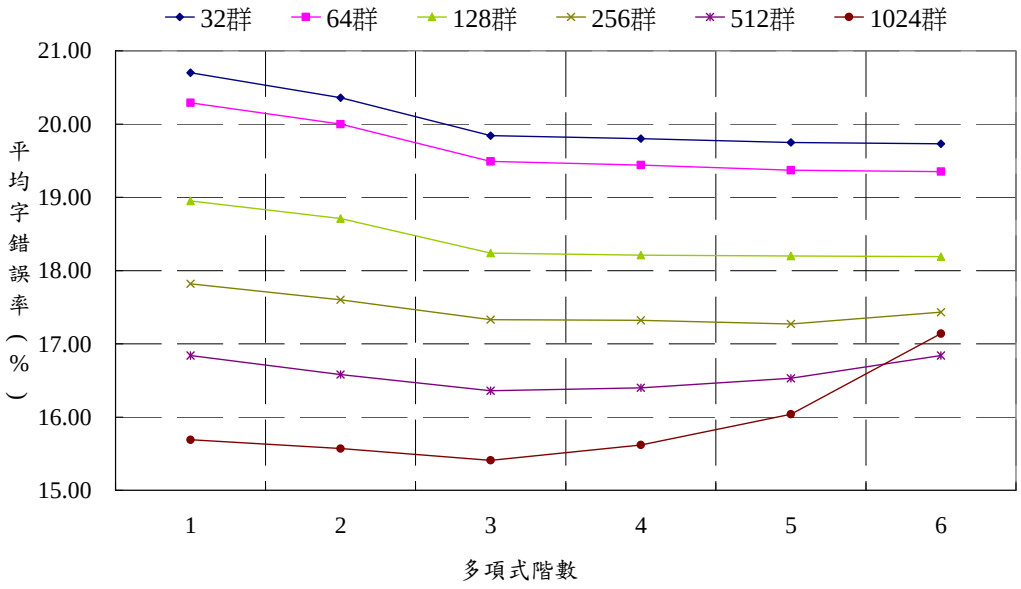


圖 4-3 群集式為基礎之多項式擬合統計圖等化法中使用不同分群數與搭配不同多項式階數的辨識結果比較圖

函數有較好的效果，相反地，在分群數較多時，低階的多項式轉換函數會有較好的效果，且愈往高階辨識效能愈差，主要原因是因為雙聲源訓練語料樣本有限，若分群數太多，每一群集內的訓練樣本數相對會減少，若再利用高階的多項式轉換函數，則此函數會過度擬合(Over-Fit)訓練樣本，造成多項式轉換函數的估測容易受到異常值(Outlier)的影響，而失去多項式迴歸的一般化(Generalization)能力，此情形亦可解釋為維度的詛咒(The Curse of Dimensionality)。因此在實作上，若分群數較少時，即每一群集內的資料樣本較多時，則可利用較高階的多項式轉換

表 4-3 群集式為基礎之多項式擬合統計圖等化法中以 1024 分群數搭配 3 階多項式轉換函數的辨識結果

乾淨語料訓練模式										
平均字錯誤率(%)	測試集A				測試集B				測試集C	
訊噪比	地下鐵	人聲	汽車	展覽會館	餐廳	街道	機場	火車站	地下鐵	街道
Clean	1.04	1.06	1.25	1.02	1.04	1.06	1.25	1.02	1.04	1.09
20dB	2.30	1.75	1.70	2.34	1.78	2.39	1.91	1.85	2.52	3.36
15dB	3.99	2.63	2.51	3.39	2.76	3.14	2.59	3.61	4.11	4.38
10dB	7.03	4.72	4.68	7.47	5.59	7.44	5.22	6.76	9.21	8.83
5dB	16.89	14.30	15.72	18.64	15.60	19.32	13.63	18.02	26.90	24.21
0dB	45.10	40.93	45.45	45.36	40.22	47.73	36.56	44.74	64.51	54.78
-5dB	79.34	73.94	80.85	80.96	73.87	78.30	71.37	77.88	85.94	79.99
平均	15.06	12.87	14.01	15.44	13.19	16.00	11.98	15.00	21.45	19.11

表 4-4 群集式為基礎之多項式擬合統計圖等化法中以 1024 分群數搭配 3 階多項式轉換函數結合倒頻譜平均消去法的辨識結果

乾淨語料訓練模式										
平均字錯誤率(%)	測試集A				測試集B				測試集C	
訊噪比	地下鐵	人聲	汽車	展覽會館	餐廳	街道	機場	火車站	地下鐵	街道
Clean	1.14	1.06	1.28	0.89	1.14	1.06	1.28	0.89	1.35	0.97
20dB	2.46	1.87	1.76	2.13	1.84	2.06	1.61	1.70	3.10	2.48
15dB	3.99	3.11	2.74	3.27	2.27	2.96	2.18	3.42	4.24	3.45
10dB	7.74	4.47	4.92	7.31	5.19	6.56	3.97	5.55	8.87	7.13
5dB	18.21	13.91	16.25	18.94	12.83	16.11	12.35	16.11	20.88	19.17
0dB	45.19	41.20	47.72	43.54	37.00	46.16	35.73	44.99	49.74	50.63
-5dB	80.35	75.42	82.85	77.35	71.11	80.53	70.47	80.96	82.10	80.56
平均	15.52	12.91	14.68	15.04	11.83	14.77	11.17	14.35	17.37	16.57

函數以求得較精細的轉換函數，若分群數較多時，則只需用低階的多項式轉換函數即可，在 Aurora-2 上的實驗，又以 1024 群集搭配 3 階的多項式轉換函數能有最好的補償效果，其於不同噪音與訊噪比下的辨識效能如表 4-3 所示。

從表中可發現在測試集 A 與測試 B 都有良好的辨識效果，但在測試集 C 的表現卻較測試集 A 與測試 B 來得差，主要原因是因為在測試集 C 中的測試語料含有與訓練語料不同的通道效應影響，所以在做分群的指派 (如式 4-7) 時，估測事後機率 $p(k \mid Y_t)$ 會產生誤差，因此效能不及測試集 A 或測試集 B。但此問題可利用倒頻譜消去法解決通道效應的影響，因此不論是在訓練高斯混合模型前或是分群指派前，所有語音特徵參數可先經過倒頻譜消去法處理，移除通道效應的影響，再接續做模型訓練或分群指派。但為了不破壞原本的統計分布，所以在求算累積密度函數時還是以未經過倒頻譜消去法處理的語音特徵參數為主。實驗結果對於測試集 C 而言，平均字錯誤率從 20.28% 下降至 16.96%，詳細辨識結果

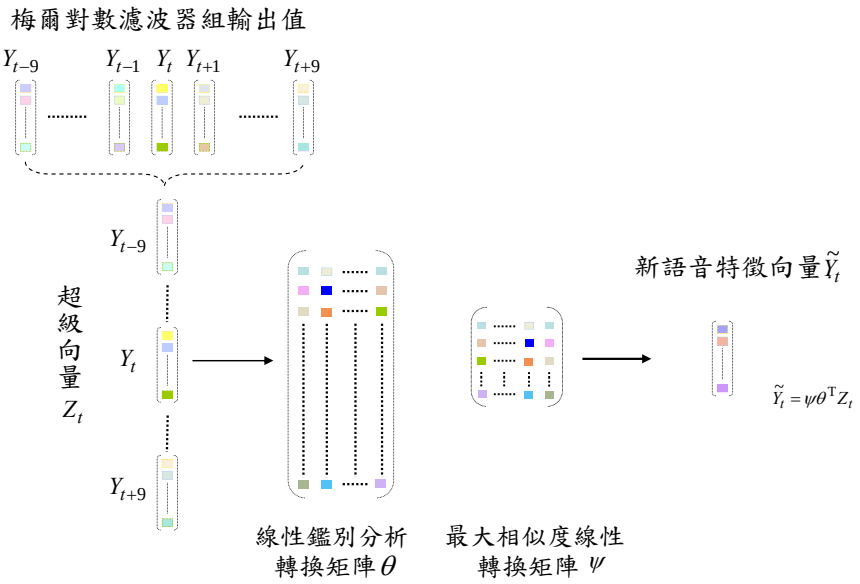


圖 4-4 鑑別性特徵擷取法示意圖

如表 4-4 所示。

4.3 群集式為基礎之多項式擬合統計圖等化法結合不同語音特徵參數相關實驗結果

下一個實驗吾人欲探討群集式為基礎之多項式擬合統計圖等化法結合不同特徵擷取方法，是否能與結合梅爾倒頻譜係數一樣，對於提升辨識效能有顯著的進步，因此吾人嘗試使用三種不同的語音特徵，第一種是原本第 13 維語音特徵參數是對數能量改成梅爾倒頻譜係數的第 0 維(C0)，第二與第三種語音特徵參數，是利用線性鑑別分析(LDA)或異質性線性鑑別分析(HLDA)加上最大相似度線性轉換(MLLT)作用在梅爾對數濾波器組輸出值之後，用來取代傳統梅爾倒頻譜係數擷取過程中需透過離散餘弦轉換達到各維度特徵向量部份解相關的效果，整體語音特徵擷取示意圖如圖 4-4 所示，對每個時間點 t 的特徵向量，是採用該時間點特徵向量加上前後各取九個時間點特徵向量形成超級特徵向量 Z_t (Feature Supervector)，此特徵向量 Z_t 經由線性鑑別分析或異質性線性鑑別分析的轉換矩陣 θ 與最大相似度線性轉換底矩陣 ψ 進行線性轉換後，以得新語音特徵向量 $\tilde{Y}_t$ ，

表 4-5 群集式為基礎之多項式擬合統計圖等化法結合不同語音特徵參數的辨識結果

乾淨語料訓練模式						
平均字錯誤率(%)	分 群 個 數					
語音特徵參數	32	64	128	256	512	1024
倒頻譜係數(Log_E)	19.73	19.35	18.19	17.27	16.36	15.41
倒頻譜係數(C0)	19.63	18.69	17.61	16.74	15.89	15.07
線性鑑別分析	18.40	17.55	16.60	15.40	14.70	13.68
異質性線性鑑別分析	17.78	17.13	16.72	15.57	14.76	13.96

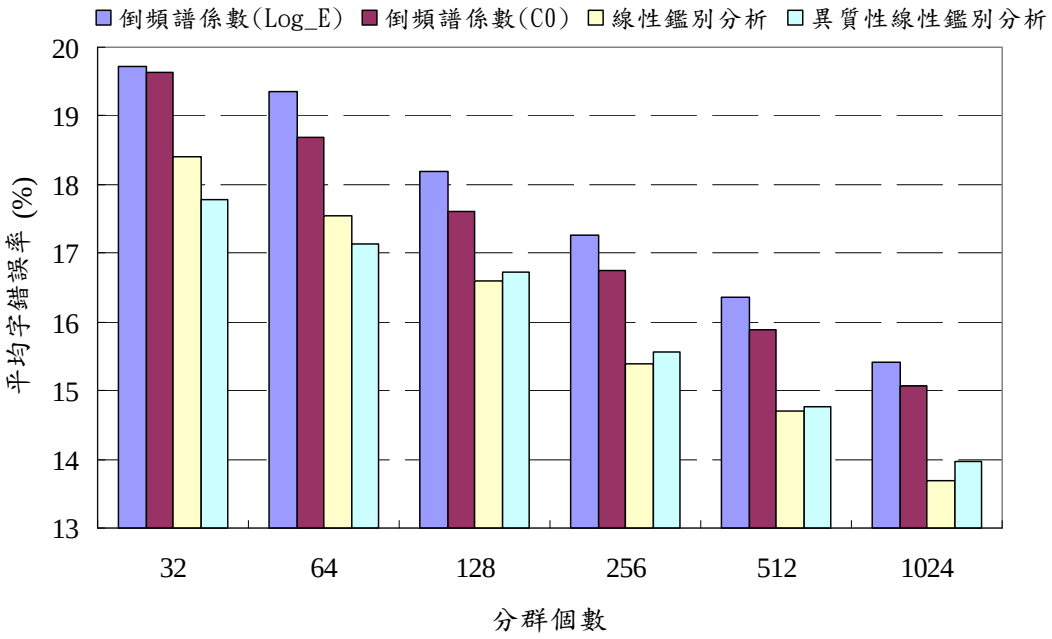


圖 4-5 群集式為基礎之多項式擬合統計圖等化法結合不同語音特徵參數的辨識結果比較圖

其數學式表示如下：

$$\tilde{Y}_t = \psi\theta^{\mathrm{T}} Z_t \qquad \text{(式 4-9)}$$

在擷取完語音特徵參數後，再使用群集式為基礎之多項式擬合統計圖等化法進行補償，其中多項式轉換函數的階數設定是參考圖 4-3 最佳設定，實驗結果如表 4-5 所示。表格中倒頻譜係數(Log_E)等同第三章梅爾倒頻譜係數擷取設定，即表 4-3 的結果，而倒頻譜係數(C0)表示將利用梅爾倒頻譜係數的第 0 維的特徵值取代原

本的對數能量維的特徵值。

表 4-6 群集式為基礎之多項式擬合統計圖等化法中以 1024 分群數搭配 3 階多項式轉換函數作用在經過線性鑑別分析處理的語音特徵參數的辨識結果

乾淨語料訓練模式										
平均字錯誤率(%	測試集A				測試集B				測試集C	
訊噪比	地下鐵	人聲	汽車	展覽會館	餐廳	街道	機場	火車站	地下鐵	街道
Clean	0.92	0.79	1.13	0.99	0.92	0.79	1.13	0.99	0.89	0.73
20dB	1.84	1.81	1.64	1.64	1.29	2.12	1.43	1.27	2.06	3.39
15dB	3.19	2.60	2.51	2.99	2.03	3.23	2.06	2.38	3.44	4.05
10dB	5.37	4.20	4.74	6.51	4.79	6.86	3.82	5.18	7.25	9.70
5dB	11.61	13.75	14.29	15.64	13.26	18.29	13.24	16.08	19.34	23.19
0dB	32.30	42.20	42.08	35.64	38.16	46.25	37.40	44.92	48.76	52.21
-5dB	68.74	77.45	78.62	66.68	75.68	77.57	73.64	78.06	77.43	80.71
平均	10.86	12.91	13.05	12.48	11.91	15.35	11.59	13.97	16.17	18.51

由圖 4-5 可看出不論是使用何種語音特徵參數結合群集式為基礎之多項式擬合統計圖等化法皆有良好的辨識效能，又以線性鑑別分析結合最大相似度線性轉換 (LDA_MLLT) 或是異質性線性鑑別分析結合最大相似度線性轉換 (HLDA_MLLT)表現更為顯著，主要原因可能是因為此二種方法本屬於鑑別性語音特徵，若利用其來訓練分群用的高斯混合模型或在計算落在某個高斯分布的事後機率，勢必會較精確。在 Aurora-2 上，最好的結果是利用線性鑑別分析結合最大相似度線性轉換搭配上 1024 群集與 3 階的多項式轉換函數，平均字錯誤率達 13.68%，相較於梅爾倒頻譜基礎實驗結果，平均字錯誤率達 67%的相對減少，詳細結果如表 4-6 所示。

第五章 群集式為基礎之多項式擬合統計圖等化法之延伸

前章節敘述的群集式為基礎之多項式擬合統計圖等化法屬於較一般化(General)的通式，若重新做一些假設後，可應用推導出不同作法。在本論文吾人延伸出二種不同作法，第一個作法是簡化假設並只使用乾淨訓練語料(Clean Training Speech Data)，應用至統計圖等化法中，吾人稱其為多項式擬合統計圖等化法(Polynomial-Fit Histogram Equalization, PHEQ)[Lin et al. 2006a, 2006b]；另一個作法是搭配遺失特徵理論，利用多項式迴歸函數本身具有預測(Prediction)的能力，達到遺失特徵的重建，吾人稱其為(Selective Cluster-based Polynomial-Fit Histogram Equalization, SCPHEQ)。下列章節將分別詳述此二種作法及其實驗結果。

5.1 多項式擬合統計圖等化法

在統計圖等化法最主要精神可以視為是利用一個轉換函數(Transformation Function)，此函數能將測試語句的語音特徵向量每一維的統計分布分別轉換至先前已從訓練語句中定義好的對應參考分布，數學式關係式表示如式 2-21 與式 2-22，因此，若將群集式為基礎之多項式擬合統計圖等化法中的假設，簡化成只使用單聲源語料(即乾淨語料)，且對於訓練語料語音特徵向量的每一維只有一個全域的(Global)多項式轉換函數表示，那麼即可將群集式為基礎之多項式擬合統計圖等化法(CPHEQ)簡化成多項式擬合統計圖等化法(PHEQ)，數學關係示表示如下

$$\tilde{y}_t = G\left(CDF\left(y_t\right)\right) = \sum_{m=0}^{M} a_m \left(C\left(y_t\right)\right)^m \qquad \text{(式 5-1)}$$

且其均方誤差定義為：

$$E^2 = \sum_{t=0}^{T-1}\left(y_t - \sum_{m=0}^{M} a_m \left(CDF_{Train}\left(y_t\right)\right)^m \right)^2 \qquad \text{(式 5-2)}$$

其中 T 為訓練語料中所有音框的個數，若要使均方誤差最小，則所有多項式係數 $a_0, a_1, \ldots, a_M$ 會滿足式 5-3 的關係，只需透過解聯立方程式，即可求得 $a_0, a_1, \ldots, a_M$ 係數。

$$\frac{\partial E'^2}{\partial a_m} = 0, \forall m = 1 \ldots M \qquad \text{(式 5-3)}$$

在辨識階段，只需要將測試語句語音特徵向量中的每一維特徵值 y_t 的對應累積密度函數值 $CDF\left(y_t\right)$ 代入先前已於訓練階段中求得的多項式函數(式 5-2)即可進行等化動作，此做法不僅能有效地解決傳統統計圖等化法或分位差統計圖等化法需耗費的大量記憶體資源與處理器運算時間的缺點，只需透過少量的多項式係數與多項式函數的運用，便能迅速的將測試語句語音特徵向量每一維的統計分布轉換至先前已從訓練語句中定義好的參考分布，並且能擁有和統計圖等化法相同的補償效果。

此外，雖然統計圖等化法對於補償因雜訊干擾所產生的非線性失真有顯著效果，但值得一提的是，由非穩性噪音(Non-stationary Noise)所造成的異常尖峰(Sharp Peak)或波谷(Valley)，可能會造成統計圖等化法在等化的過程中，某些語音特徵值被異常的放大或縮小，此異常情形可以由圖 5-1 中觀察到，最上方的圖為尚未做統計圖等化法前，原本的乾淨語料與雜訊語料的梅爾倒譜頻係數的第二維特徵值；中間的圖為做完統計圖等化法後的梅爾倒譜頻係數的第二維特徵值，可清楚看見有二個區域被過度強調。因此此問題可利用語音訊號本身是屬於變化緩慢的特性，利用移動平均法來達到音框間特徵值的平滑(Smoothing)，減緩音框間過度劇烈的快速變化。

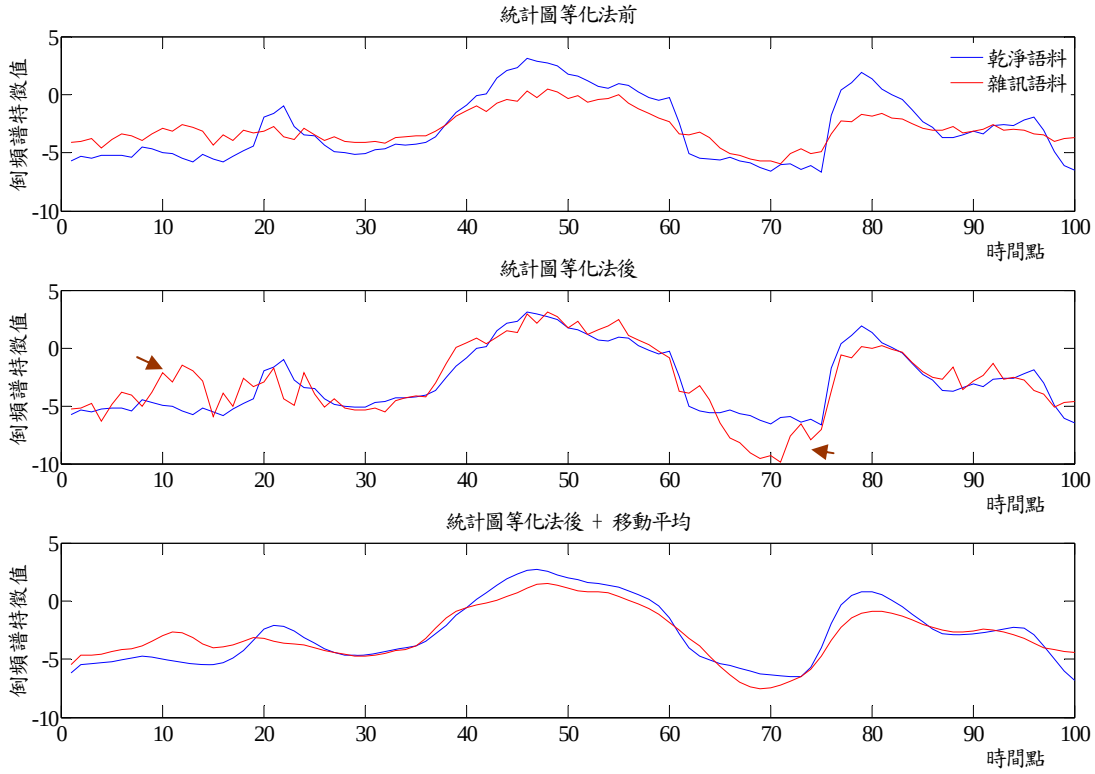


圖 5-1 非穩性噪音所造成的異常尖峰或波谷示意圖

移動平均的使用在語音辨識研究上，並非是一個全新的議題，例如[Chen et al. 2002]曾利用移動平均的概念，提出一種不同特徵向量正規化的方法，首先先對語音特徵向量進行平均消去法和變異數正規化，接著再利用自動迴歸移動平均(Auto-Regression Moving Average, ARMA)對特徵向量進行平滑的動作，實驗結果證實移動平均的使用對於提升整體辨識率有很大的幫助。然而依照移動平均所考慮語音特徵來源與時間軸點數不同，可以有下列數種選擇[Chen et al. 2002; Chen and Bilmes 2007]。

- 非因果關係移動平均(Non-Causal Moving Average)

$$\hat{y}_t = \begin{cases} \dfrac{\sum_{i=-L}^{L} \tilde{y}_{(t+i)}}{2L+1} & if\ L < t \le T-L, \\ \tilde{y}_t & otherwise \end{cases} \qquad \text{(式 5-4)}$$

- 因果關係自動迴歸移動平均(Causal Moving Average)

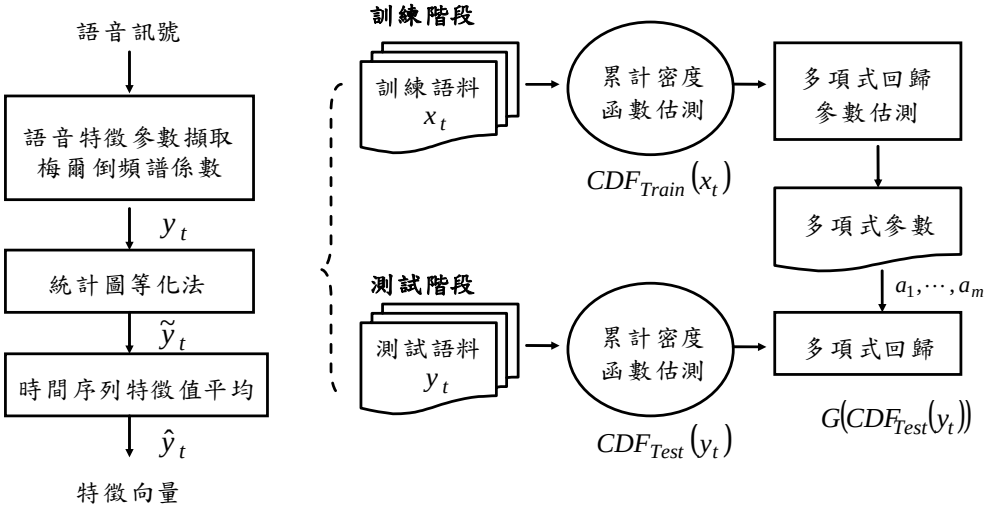


圖 5-2 多項式擬合統計圖等化法的流程圖

$$\hat{y}_t = \begin{cases} \dfrac{\sum_{i=0}^{L} \tilde{y}_{(t-i)}}{L+1} & if\ L < t \le T, \\ \tilde{y}_t & otherwise \end{cases} \qquad \text{(式 5-5)}$$

- 非因果關係自動迴歸移動平均(Non-Causal Auto Regression Moving Average)

$$\hat{y}_t = \begin{cases} \dfrac{\sum_{i=1}^{L} \hat{y}_{(t-i)} + \sum_{j=0}^{L} \tilde{y}_{(t+j)}}{2L+1} & if\ L < t \le T-L, \\ \tilde{y}_t & otherwise \end{cases} \qquad \text{(式 5-6)}$$

- 因果關係自動迴歸移動平均(Causal Auto Regression Moving Average)

$$\hat{y}_t = \begin{cases} \dfrac{\sum_{i=1}^{L} \hat{y}_{(t-i)} + \sum_{j=0}^{L} \tilde{y}_{(t-j)}}{2L+1} & if\ L < t \le T, \\ \tilde{y}_t & otherwise \end{cases} \qquad \text{(式 5-7)}$$

其中 $\tilde{y}_t$ 為輸入的語音特徵值，$\hat{y}_t$ 為經由移動平均法後所求得新的語音特徵值，L 表示移動平均項階數(Order of Moving Average)，圖 5-1 最下方的為做完統計圖等化法加移動平均後的倒譜頻特徵向量，從圖中可清楚看到做完移動平均後，被異常放大或縮小的語音特徵值已被平滑掉。

表 5-1 多項式擬合統計圖等化法的辨識結果

乾淨語料訓練模式					
平均字錯誤率(%)		多項式階數			
		3	5	7	9
訓練數量	所有語料	22.39	21.54	21.08	21.30
	1000組	21.80	21.46	21.13	21.16
	100組	22.68	21.31	20.75	20.55
	10組	23.42	22.20	22.54	23.42

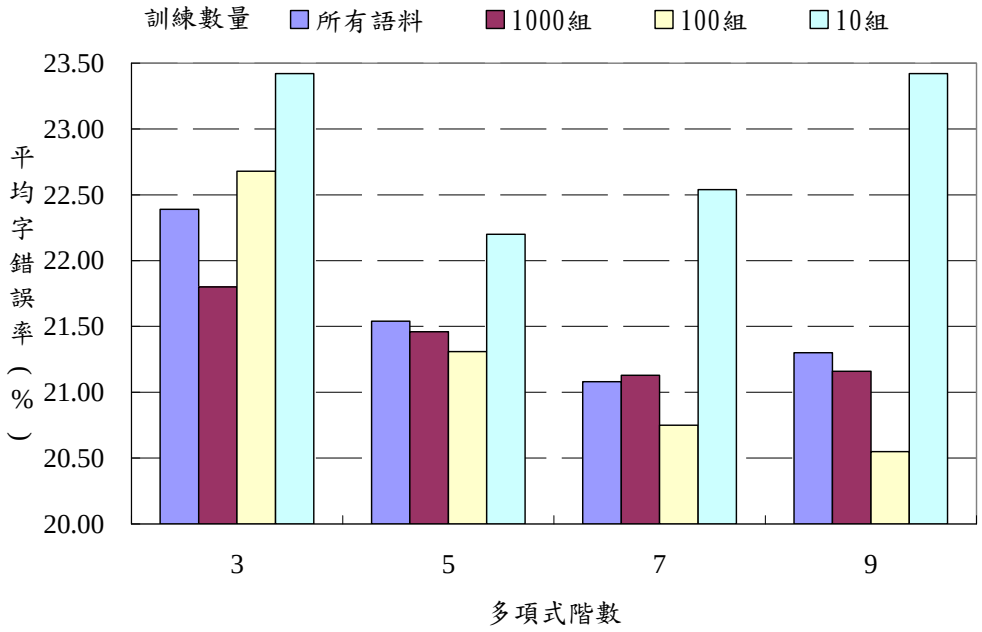


圖 5-3 多項式擬合統計圖等化於不同設定下之實驗結果比較圖

5.1.1 多項式擬合統計圖等化法(PHEQ)相關實驗結果

在多項式擬合統計圖等化法相關實驗中，第一個實驗是利用多項式迴歸模型描述參考分布的累積密度函數分布情形，探討使用所有訓練語料與否以及不同多項式階數對於整體辨識效能影響結果如何。其中參考分布的資訊是由乾淨訓練語料統計而成的，在累積密度函數的求取，除了使用所有的訓練語料外，亦嘗試將訓練語料分成 1000 組、100 組和 10 組，每一分組是以組內所有特徵值的平均數做為該組代表特徵值；同時也使用不同階數的多項式進行等化動作，辨識結果如表 5-1 所示。其中值得注意的是，由於多項式階數結束行為(End Behavior)的特性，會使得使用偶數階數的多項式可能無法滿足累積密度函數結束行為的特性，所以

表 5-2 多項式擬合統計圖等化法使用 7 階的多項式迴歸以及 100 分組組數實驗的辨識結果

乾淨語料訓練模式										
平均字錯誤率(%)	測試集A				測試集B				測試集C	
訊噪比	地下鐵	人聲	汽車	展覽會館	餐廳	街道	機場	火車站	地下鐵	街道
Clean	0.92	1.03	0.81	0.89	0.92	1.03	0.81	0.89	0.92	1.03
20dB	3.29	2.30	2.24	3.55	2.03	2.72	1.88	2.07	3.65	2.90
15dB	6.14	4.08	3.79	7.00	4.33	4.75	3.52	4.57	7.58	5.86
10dB	13.45	9.10	8.92	13.58	9.33	10.97	7.13	9.35	16.70	13.69
5dB	30.80	25.00	23.62	29.99	22.75	27.03	19.45	22.86	39.58	32.07
0dB	63.37	54.53	54.73	58.87	49.12	56.92	47.33	54.21	70.68	64.12
-5dB	86.77	82.83	82.91	84.29	79.31	82.26	78.32	80.84	88.06	84.58
平均	23.41	19.00	18.66	22.60	17.51	20.48	15.86	18.61	27.64	23.73

表 5-3 多項式擬合統計圖等化法結合不同移動平均法的辨識結果

乾淨語料訓練模式						
平均字錯誤率(%)	移 動 平 均 項					
平均方式	0	1	2	3	4	5
非因果移動平均	20.75	17.75	16.83	17.26	18.15	19.66
因果移動平均	20.75	19.23	18.28	17.44	17.12	17.28
非因果自動迴歸	20.75	17.83	16.90	16.38	16.99	17.34
因果自動迴歸	20.75	17.93	16.84	19.20	17.44	19.20

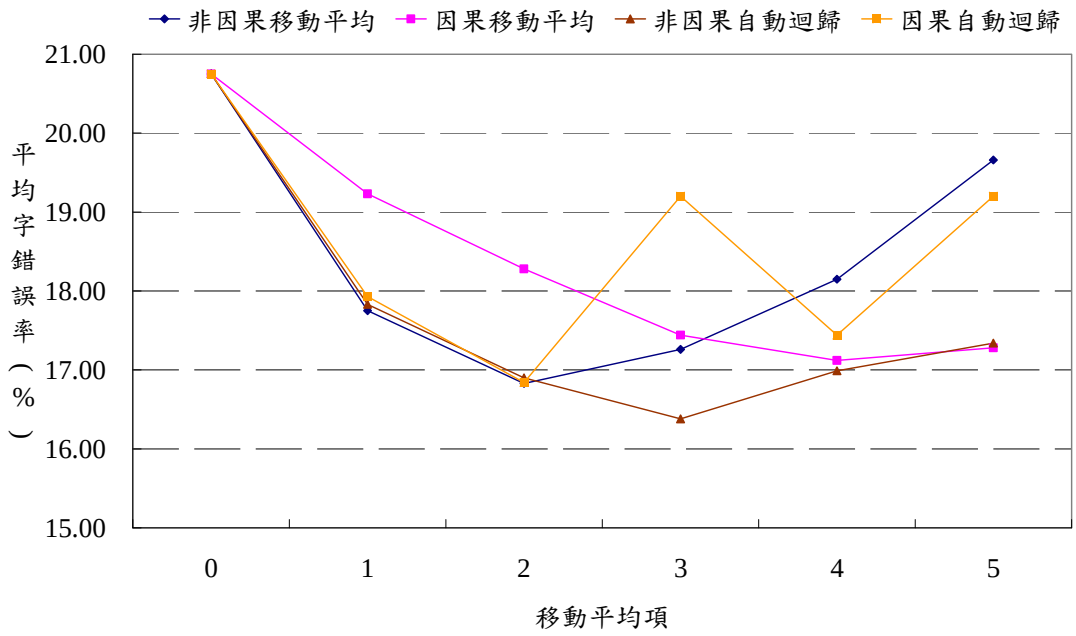


圖 5-4 多項式擬合統計圖等化法結合不同移動方均方法的辨識結果比較圖

在多項式擬合統計圖等化法並不考慮偶數階數的使用。由圖 5-3 可清楚看到，辨識效能隨著多項式階數增加有所進步，然而並非使用階數愈高愈好，因為資料的

表 5-4 多項式擬合統計圖等化法使用 7 階的多項式迴歸以及 100 分組組數搭配 3 階的非因果關係自動迴歸移動平均的辨識結果

乾淨語料訓練模式										
平均字錯誤率(%)	測試集A				測試集B				測試集C	
訊噪比	地下鐵	人聲	汽車	展覽會館	餐廳	街道	機場	火車站	地下鐵	街道
Clean	0.89	1.24	1.13	1.02	0.89	1.24	1.13	1.02	1.04	1.24
20dB	2.39	2.33	1.73	2.62	2.43	2.24	1.97	1.73	2.67	2.39
15dB	4.79	4.69	2.80	5.34	5.13	3.54	3.52	4.07	5.53	4.87
10dB	10.65	9.46	6.41	11.17	9.98	8.89	6.62	7.81	11.82	10.52
5dB	22.90	22.34	15.15	24.28	22.04	20.10	17.33	17.16	27.33	23.10
0dB	47.19	48.34	34.77	46.71	46.24	42.96	39.28	39.22	53.76	48.73
-5dB	75.25	78.60	65.55	72.79	76.02	74.43	69.28	68.22	78.26	75.88
平均	17.58	17.43	12.17	18.02	17.16	15.55	13.74	14.00	20.22	17.92

散布情形，可能使高階多項式為了要更符合資料分布情形，而造成過度擬合的情形；同樣地，若使用所有訓練語料來求算多項式係數亦會有過度擬合的情形或受異常值影響，Aurora-2 語料庫上最好的結果是利用 7 階的多項式迴歸以及 100 分組組數的累積統計圖有較佳的辨識結果。平均字錯誤率達 20.75%，相較於查表式統計圖等化法或是分位差統計圖等化法，辨識效能不相上下。下列所有實驗將使用 7 階多項式迴歸，並且參考分布是利用 100 組的累積統計圖中統計而得，其於不同雜訊與訊噪比的辨識效果如表 5-2 所示。

接下來，吾人探討移動平均的使用對於減輕由雜訊或等化過程中所造成的異常情形，而提高辨識能的效果如何，實驗結果如表 5-3 所示，表中清楚的呈現無論是使用哪種移動平均法，對於提升多項式擬合統計圖等化後語音特徵的辨識效果皆有明顯的幫助，其中當移動平均項為 0 時，表示不做任何平均動作，亦即單純使用多項式擬合統計圖等化法所得到的辨識結果。實驗結果和[Chen et al. 2002]呈現的相同，使用非因果關係自動迴歸移動平均有較佳的辨識結果，在 Aurora-2 語料庫上以搭配 3 階非因果關係自動迴歸移動平均達最好的效果，其於不同雜訊與訊噪比的辨識效果如表 5-4 所示，若相較於單純使用多項式擬合統計圖等化法而言，字錯率可相對地減少約 20%左右。此外由圖 5-4 可看出若移動平均項的階數若使用太高，可能會造成原本帶有鑑別資訊的特徵值，因此被平滑掉，使得辨識效能下降。

5.2 群集式為基礎之選擇性多項式擬合統計圖等化法

群集式為基礎之多項式擬合統計圖等化法的另一種延伸是搭配遺失特徵理論，利用多項式迴歸模型來預測遺失特徵的特徵值，吾人稱之為群集式為基礎之選擇性多項式擬合統計圖等化法(Selective Cluster-based Polynomial-Fit Histogram Equalization, SCPHEQ)。在遺失特徵理論中，特徵重建主要包含二個主要步驟，第一步是決定語音特徵向量中哪些特徵參數是可靠，哪些是不可靠(或遺失)的，因為此部份非本論文重點，所以在本論文是假設已經有一個非常好的方法可以用來判定每個音框中每一維度的特徵值是可靠或是不可靠的。而第二步是針對不可靠的特徵參數進行重建，此步驟乃為本小節的探討重點。

首先，吾人假設和群集式為基礎之多項式擬合統計圖等化法(CPHEQ)一樣，假設重建的語音特徵參數可利用多項式迴歸模型，因此重建的語音特徵值可經由如式 4-6 求得。但與群集式為基礎之多項式擬合統計圖不同的是Y_t並非倒頻譜特徵值，而是頻譜特徵值(Spectral Value)，主要原因是因為雜訊干擾語音訊號可能只在某些頻段(Frequency Band)上發生，對於其他頻段並不會有所干擾，且也因為倒頻譜係數會含蓋數個頻段的資訊，所以較難進行可靠語音特徵參數與不可靠語音特徵參數的判定。此外，因為在頻譜特徵值的值域變化差距非常大，若利用統計圖來求算對應的累積密度函數並不適合，因此，吾人利用高斯誤差函數(Gaussian Error Function)求得每個音框的累積密度函數。若隨機變數的統計分布是一高斯分布，即$Y_t \sim N\left(\mu,\sigma^2\right)$，那麼累積密度函數值式求得[Abramowitz and Stegun 1965]：

$$CDF\left(y_t\right)=\Phi\left(x_t\right)=\frac{1}{2}\left(1+erf\left(\frac{y_t-\mu}{\sigma\sqrt{2}}\right)\right)$$
$$erf\left(y_t\right)=\frac{2}{\sqrt{\pi}}\int_0^{y_t} e^{-t^2}dt \qquad \text{(式 5-8)}$$

其中μ為平均數，σ為標準差，但如同前面章節所敘述的，語音特徵參數的分布

並非是一常態分布，所以亦可將式 5-8 延伸至高斯混合模型。而在高斯混合模型中，累積密度函數的求得可利用下式近似：

$$CDF(y_t)=\sum_{j=1}^{J}c_j\times\Phi_j(y_t)=\frac{1}{2}\sum_{j=1}^{J}c_j\left(1+erf\left(\frac{y_t-\mu_j}{\sigma_j\sqrt{2}}\right)\right) \quad \text{(式 5-9)}$$

其中 J 為高斯混合模型中所有高斯分布的個數，c_j 為高斯混合模型中第 j 個高斯分布的權重，μ_j 和 σ_j 分別為第 j 個高斯分布所對應的平均數與標準差。因此在實作上，群集式為基礎之選擇性多項式擬合統計圖等化法需使用二組不同的高斯混合模型，一組與群集式為基礎之多項式擬合統計圖等化法相同，即利用雙聲源語料中的雜訊語料 Y_t 訓練出一組高斯混合模型：

$$p(Y_t)=\sum_{k=1}^{K}P(k)p(Y_t\mid k)=\sum_{k=1}^{K}c_k N(Y_t;\mu_k,\Sigma_k) \quad \text{(式 5-10)}$$

另一組高斯混合模型則是用來近似語音特徵參數的累積密度函數，對於每一群集 k 皆有一組對應的高斯混合模型，高斯混合模型是由收集所有落至該群集的語音特徵向量訓練而得：

$$p_k(Y_t)=\sum_{j=1}^{J}p_k(j)p_k(Y_t\mid j)=\sum_{j=1}^{J}c_{kj}N_k(Y_t;\mu_{kj},\Sigma_{kj}),\forall Y_t\in Cluster_k \quad \text{(式 5-11)}$$

而參數的重建是利用迴歸模型 $R(y_t)$ 求得，對於每一群集的多項式迴歸模型求得方式，與群集式為基礎之多項式擬合統計圖相同，迴歸模型的係數亦是經由最小化下列均方錯誤求得。

$$\tilde{y}_t=R(y_t)=\sum_{k=1}^{K}\left(\left(\sum_{m=0}^{M}a_{km}(CDF(y_t))^m\right)\times\delta(k\mid Y_t)\right) \quad \text{(式 5-12)}$$

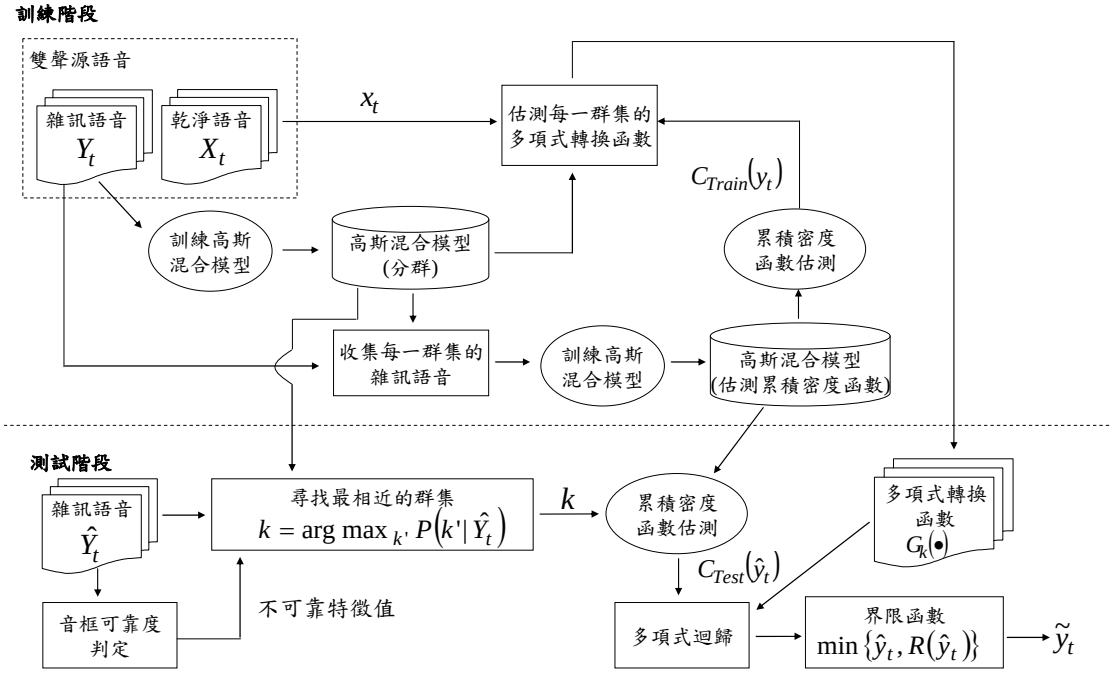


圖 5-5 群集式為基礎之選擇性多項式擬合統計圖等化法的流程圖

$$E_k^2 = \sum_{t=0}^{T-1}\left(\left(x_t - \sum_{m=0}^{M} a_{km} CDF(y_t)\right)\times \delta(k \mid Y_t)\right)^2 \qquad \text{(式 5-13)}$$

與群集式為基礎之多項式擬合統計圖等化法最大的差異是在測試階段時，頻譜特徵值的重建並非直接取用迴歸模型的輸出值，因為雜訊干擾在頻譜的表現上，受雜訊干擾的頻譜特徵值必會大於或等於原本的特徵值，所以吾人利用一界限函數(Bounded Function)避免重建後的頻譜特徵值大於原本的頻譜特徵值，因此頻譜語音特徵參數重建的判斷如下：

$$\tilde{y}_t = \begin{cases} \min\{y_t, R(y_t)\} & \text{，不可靠特徵值} \\ y_t & \text{，可靠特徵值} \end{cases} \qquad \text{(式 5-14)}$$

群集式為基礎之選擇性多項式擬合統計圖等化法的實作流程如圖 5-5 所示。

5.2.1 群集式為基礎之選擇性多項式擬合統計圖等化法相關實驗結果

在本小節的實驗，由於 Aurora-2 語料庫中，對於每一雜訊測試語料皆有其相對

表 5-5 群集式為基礎之選擇性多項式擬合統計圖等化法中使用不同分群數與搭配不同多項式階數的辨識結果

乾淨語料訓練模式							
平均字錯誤率(%)		多項式階數					
		1	2	3	4	5	6
分群個數	32	18.10	17.95	17.90	17.89	17.96	17.97
	64	16.42	16.03	15.86	15.72	15.74	15.82
	128	14.23	14.05	13.87	13.89	13.75	13.80
	256	13.26	12.98	12.87	12.85	13.03	12.96
	512	12.30	12.19	12.03	12.02	12.06	12.01
	1024	11.64	11.57	11.37	11.38	11.34	11.36

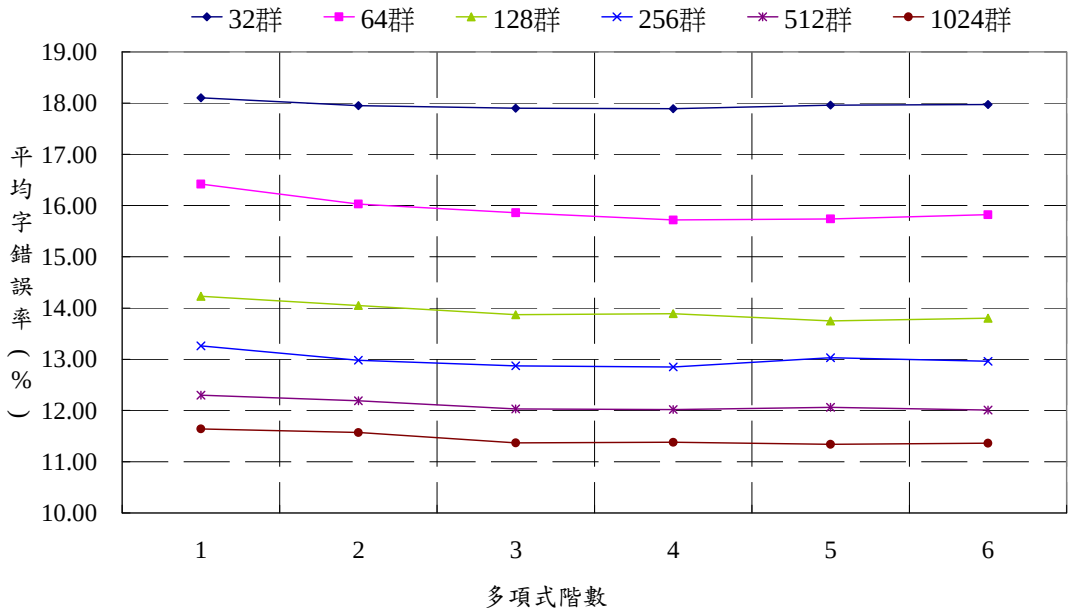


圖 5-6 群集式為基礎之選擇性多項式擬合統計圖等化法中使用不同分群數與搭配不同多項式階數的辨識結果比較圖

應的乾淨語料，因此吾人可以利用此相對應關係料來判定語音特徵向量中哪些語音特徵參數是可靠的，哪些是不可靠的。判定的方法主要是利用訊噪比的方式，若訊噪比小於一門檻值(Threshold)，則吾人判定其為不可靠的語音特徵值，則使用群集式為基礎之選擇性多項式擬合統計圖等化法進行語音特徵參數的重建，若為可靠的語音特徵值，即保留原本的語音特徵值，訊噪比求算方式為：

$$SNR_t(d) = 20\log_{10}\left(\frac{x_t(d)}{y_t(d) - x_t(d)}\right) \qquad \text{(式 5-15)}$$

表 5-6 群集式為基礎之選擇性多項式擬合統計圖等化法中以 1024 分群數搭配 3 階多項式轉換函數的辨識結果

乾淨語料訓練模式										
平均字錯誤率(%	測試集A				測試集B				測試集C	
訊噪比	地下鐵	人聲	汽車	展覽會館	餐廳	街道	機場	火車站	地下鐵	街道
Clean	0.86	0.94	1.10	0.86	0.86	0.94	1.10	0.86	0.64	0.79
20dB	1.41	1.30	1.22	1.14	1.26	1.51	1.61	1.20	1.20	1.54
15dB	2.06	1.93	1.40	2.04	1.90	2.93	2.68	2.72	1.93	2.42
10dB	3.87	5.86	3.19	3.70	5.59	5.86	7.01	7.65	3.56	4.11
5dB	8.01	21.89	11.66	8.86	17.38	18.44	21.74	21.63	6.23	8.74
0dB	20.48	50.18	31.58	22.40	44.06	42.17	46.38	48.01	12.43	20.77
-5dB	46.91	77.60	65.17	51.84	73.23	71.10	74.41	76.95	36.17	47.97
平均	7.17	16.23	9.81	7.63	14.04	14.18	15.88	16.24	5.07	7.52

其中 $y_t(d)$ 為雜訊語料中第 t 個音框的第 d 維的頻譜特徵值，$x_t(d)$ 為所對應乾淨語料中第 t 個音框的第 d 維頻譜特徵值，當計算出的訊噪比小於-5dB，則吾人標註其為不可靠的語音特徵值。實驗結果發現每一群集內用來求算累積密度函數的高斯混合模型只需使用 4 個高斯分布即足夠用來近似，表 5-5 為使用不同群集數與不同階數的迴歸模型所得到的辨識結果，辨識結果的趨勢和群集式為基礎之多項式擬合統計圖等化法(CPHEQ)非常相似，隨著分群數增加，平均字錯誤率會隨著降低，在分群數較少時，高階的多項式轉換函數有較好的效果。但不同的是，在分群數較多時，高階的多項式轉換函數並不一定讓使辨識效能變差，主要原因可能是因為界限函數的使用(如式 5-14)，界限函數確保重建後的頻譜特徵值一定要比目前受雜訊干擾的頻譜特徵值小，因此縱使由多項式轉換函數所求得的重建頻譜特徵值發生異常情形，透過此界限函數的使用，能消減掉異常值出現的可能。在以 1024 群集搭配 3 階的多項式轉換函數下的平均字錯誤率為 11.37%，相較於梅爾倒頻譜基礎實驗結果，高達 72%的相對字錯誤率減少，其於不同雜訊與訊噪比的辨識效果如表 5-6 所示。

第六章 結論與未來展望

6.1 結論

強健性語音辨識技術在語音辨識領域裡，一直是一個非常急迫的議題，唯有解決雜訊干擾語音所造成的失真問題，語音辨識技術方能慢慢地被大眾所接受。目前語音強健技術大致分為三個方向：語音強化技術、強建性語音特徵技術與聲學模型調適技術，本論文主要針對強建性語音特徵技術進行改良。

在本論文，吾人提出三種不同強健性語音特徵技術的改良方法，下列分別總結此三種方法的研究成果：

(1) 群集式為基礎之多項式擬合統計圖等化法(CPHEQ)

此方法主要是探討結合語音特徵參數本身及其所對應的統計分布特性，再加上雙聲源語料進行語音特徵參數補償動作，以最小均方誤差為概念出發，進而使用多項式數據擬合方法達到非線性的補償效果。在乾淨語料訓練模式下，若使用 1024 群集與 3 階的多項式轉換函數，平均字錯誤率為 15.41%，相較於梅爾倒頻譜基礎實驗結果，此方法達 62%左右的相對字錯誤率減少。若與雙聲源為基礎分段線性補償法比較，在同樣使用 1024 群集的情況下，亦有 19%的相對字錯誤率減少。如果再更進一步搭配鑑別性特徵，平均字錯誤率為 13.68%，平均字錯誤率更達 67%左右的相對減少[Lin et al. 2007a, 2007b]。

(2) 多項式擬合統計圖等化法(PHEQ)

再者，吾人對群集式為基礎之多項式擬合統計圖等化法的假設做進一步簡化，簡化為只利用單聲源語料(乾淨語料)求得多項式轉換函數，且語音特徵向量中每一維語音特徵只利用一個全域的多項式轉換函數表

示。在乾淨語料訓練模式下，若使用 7 階的多項式轉換函數，平均字錯誤率 20.75%，辨識效能與查表式統計圖等法化(THEQ)與分位差統計圖等化法(QHEQ)不相上下，但本法確能避免傳統作法中需耗費的大量記憶體資源與處理器運算時間的缺點。若更進一步結合 3 階非因果關係自動迴歸移動平均(Non-Causal ARMA)，減輕因由非穩性噪音所造成的異常尖峰或波谷及等化過程中造成部份特徵值被過度放大或縮小的異常情形，辨識結果達平均字錯誤率 16.38%，相較於梅爾倒頻譜基礎實驗結果能有 60%左右的平均字錯誤率相對減少。[Lin et al. 2006a, 2006b]

(3) 群集式為基礎之選擇性多項式擬合統計圖等化法(CPHEQ)

在此方法中，吾人結合遺失特徵理論和多項式迴歸模型的預測與一般化能力，重建被標記為不可靠的語音特徵參數；並且進一步考慮雜訊干擾語音訊號在頻譜上的表現特性，使用一界限函數避免異常的語音特徵值。實驗結果在乾淨語料訓練模式下，若使用 1024 群集與 3 階的多項式轉換函數，平均字錯誤率達 11.37%，相較於梅爾倒頻譜基礎實驗結果，此方法達 72%左右的相對字錯誤率減少。

吾人總結前面章節各種不同強健性語音辨識技術之辨識結果以圖表方式呈現，結果如圖 6-1 所示，從圖中可以清楚看見本論文所提出的三種方法，相較於傳統作法，皆有不錯的效果。

6.2 未來展望

最後，吾人列舉出數點未來可能的研究方向：

(1) 本論文提出的方法大多數需要使用雙聲源語料，然而雙聲源語料並非如此容易取得。因此吾人希望未來能只用單聲源語料進行多項式函數的估測，可能的作法有二種：

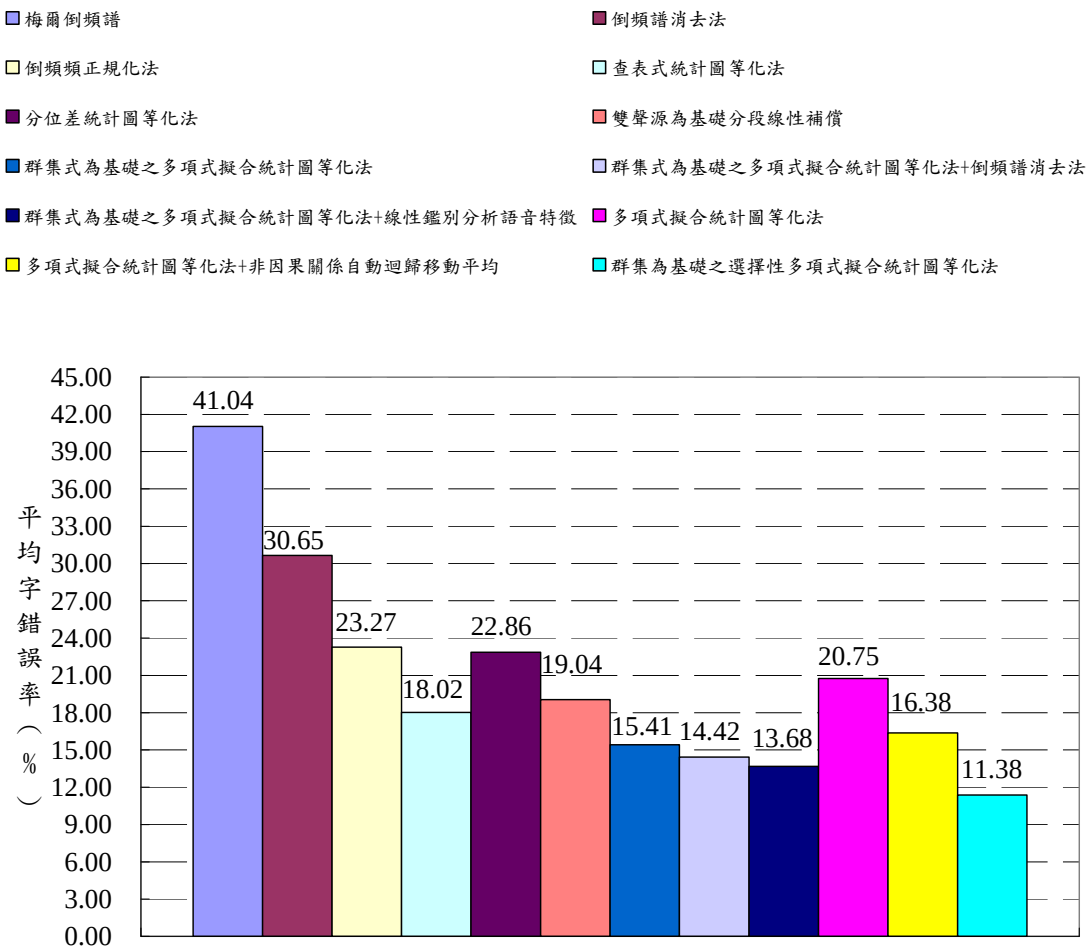


圖 6-1 各種不同強健性語音辨識技術之辨識結果比較圖

甲、 單純利用乾淨語料，再人工加入大量不同訊噪比的噪音或利用蒙地卡羅(Monte Carlo)法模擬雜訊干擾語音情形，再進行多項式轉換函數的求得。

乙、 利用雜訊語料配合上聲學模型的使用，利用最大相似度或其他鑑別式訓練法則來求得多項式轉換函數。

(2) 多項式數據擬合法容易受異常值影響，尤其當資料樣本較少時，影響情形更為嚴重，因此另一個可能的研究方向是異常值偵測(Outlier Detection)與排除或是利用強健迴歸(Robust Regression)來求得多項式函數[Montgomery et al. 2006]。

(3) 目前本論文假設語音特徵向量中每個維度間彼此為獨立的，未來的可能

研究方向之一是利用向量迴歸(Vector Regression)的方式求得迴歸模型，進而求算補償後的特徵值[Xiao et al. 2006]。

(4) 目前的迴歸模型的輸入只單單考慮單一個音框的資訊，並未考慮語音特徵參數的前後文資訊(Contextual Information)，然而語音本身是屬於變化緩慢的訊號，此特性意謂著相鄰的音框或許有高度的相關資訊，若能善用此特性，那麼勢必能有效地提昇辨識效能。

(5) 本論文的所有實驗皆是作用在 Aurora-2 語料庫上，未來吾人亦會嘗試將本論文所提出的方法，應用至不同語音辨識任務上，例如大詞彙連續語音辨識系統(LVCSR)或是其它辨識任務上，探討吾人所提出的方法在不同辨識任務上的表現如何。

參考文獻

Abramowitz, M., and I. A. Stegun (1972), "Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables," Dover.

Acero, A. (1990), "Acoustical and Environmental Robustness for Automatic Speech Recognition," Ph. D. Dissertation, Carnegie Mellon University.

Alpaydin, E. (2004), "Introduction to Machine Learning," The MIT Press.

Barker, J., M. P. Cooke, et al. (2001), "Robust ASR based on Clean Speech Models: An Evaluation of Missing Data Techniques for Connected Digit Recognition in Noise," *Interspeech'2001 - 7th European Conference on Speech Communication and Technology (Eurospeech)*, Alaborg, Denmark.

Beyerlein, P., X. Aubert, et al. (2002), "Large Vocabulary Continuous Speech Recognition of Broadcast News - The Philips/RWTH Approach," *Speech Communication*, vol. 37: pp. 109-131.

Boll, S. F. (1979), "Supperssion of Acoutstic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27(2): pp. 113-120.

Chen, C. P. and J. Bilmes (2007), "MVA Processing of Speech Features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15(1): pp. 257-270.

Chen, C. P., J. Bilmes, et al. (2002), "Low-Resource Noise-Robust Feature Post-Processing on Aurora 2.0," *Interspeech'2002 - 7th International Conference on Spoken Language Processing (ICSLP)*, Denver, Colorado.

Cooke, M., P. Green, et al. (2001), "Robust Automatic Speech Recognition with Missing and Uncertain Acoustic Data," *Speech Communication*, vol. 34: pp. 267-285.

Davis, S. B. and P. Mermelstein (1980). "Comparison of Parametric Representations

for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28(4): pp. 357-366.

Deng, L., A. Acero, et al. (2000), "Large Vocabulary Speech Recognition under Adverse Acoustic Environments," Interspeech'2000 - 6th International Conference on Spoken Language Processing (ICSLP), Beijing, China.

Dharanipragada, S. and M. Padmanabhan (2000), "A Nonlinear Unsupervised Adaptation Technique for Speech Recognition," Interspeech'2000 - 6th International Conference on Spoken Language Processing (ICSLP), Beijing, China.

Droppo, J. and A. Acero (2005), "Maximum Mutual Information SPLICE Transform for Seen and Unseen Conditions," Interspeech'2005 - 9th European Conference on Speech Communication and Technology (Eurospeech), Lisbon, Portugal.

Droppo, J., A. Acero, et al. (2001), "Evaluation of the SPLICE Algorithm on the Aurora2 Database," *Interspeech'2001 - 7th European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg, Denmark.

Droppo, J., L. Deng, et al. (2002), "Evaluation of SPLICE on the Aurora 2 and 3 Tasks," *Interspeech'2002 - 7th International Conference on Spoken Language Processing (ICSLP)*, Denver, Colorado.

Droppo, J., M. Mahajan, et al. (2005), "How to Train a Discriminative Front End with Stochastic Gradient and Maximum Mututal Information," *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'05)*, San Juan, Puerto Rico.

Duda, R. O. and P. E. Hart (1973), "Pattern Classification and Scene Analysis," New York, John Wiley and Sons.

EL-Maliki, M. and A. Drygajlo (1999), "Missing Features Detection and Handling for Robust Speaker Verification," *Interspeech'1999 - 6th European Conference on*

Speech Communication and Technology (Eurospeech), Budapest, Hungary.

Fiscus, J. (1997), "A Post-Processing System To Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'97)*, Santa Barbara, California.

Furui, S. (1981), "Cepstral Analysis Techniques for Automatic Speaker Verification," *IEEE Transaction on Acoustic, Speech and Signal Processing*, vol. 29(2): pp. 254-272.

Gales, M. J. F. (1998), "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition," *Computer Speech and Language*, vol. 12(2): pp. 75-98.

Gales, M. J. F. (2002), "Maximum Likelihood Multiple Subspace Projections for Hidden Markov Models," *IEEE Transactions on Speech and Audio Processing* , vol. 10(2): pp. 37-47.

Gales, M. J. F. and S. J. Young (1995), "Robust Speech Recognition in Additive and Convolutional Noise Using Parallel Model Combination.," *Computer Speech and Language*, vol. 9: pp. 289-307.

Gales, M. J. F. and S. J. Young (1996), "Robust Continuous Speech Recognition using Parallel Model Combination," *IEEE Transaction on Speech and Audio Processing*, vol. 4(5): pp. 352-359.

Gauvain, J.-L. and C.-H. Lee (1994), "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transaction on Speech and Audio Processing*, vol. 2(2): pp. 291-297.

Gong, Y. (1995), "Speech Recognition in Noisy Environments: A Survey," *Speech Communication*, vol. 16(3): pp. 261-291.

Hain, T., P. C. Woodland, et al. (2005), "Automatic Transcription of Conversational Telephone Speech," *IEEE Transactions on Speech and Audio Processing*, vol.

13(6): pp. 1173-1185.

Hamme, H. V. (2004), “Robust Speech Reocgnition Using Cepstral Domain Missing Data Techniques and Noisy Mask,” IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP '04), Quebec, Canada.

Hermansky, H. (1991), “Perceptual Linear Predictive (PLP) Analysis of Speech,” *Journal of the Acoustical Society of America*, vol. 87: pp. 1738-1752.

Hermansky, H. and N. Morgan. (1994), “RASTA processing of speech,” *IEEE Transactions on Speech and Audio Processing*, vol. 2(4): pp. 578-589.

Hilger, F. and H. Ney (2001), “Quantile Based Histogram Equalization for Noise Robust Speech Recognition,” *Interspeech'2001 - 7th European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg, Denmark.

Hilger, F. and H. Ney (2006), “Quantile Based Histogram Equalization for Noise Robust Large Vocabulary Speech Recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14(3): pp. 845-854.

Hirsch, H. G. and D. Pearce (2002), “The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions,” *in Proc. ISCA ITRW ASR2000*, Paris, France.

Hsu, C. W. and L. S. Lee (2004), “Higher Order Cepstral Moment Normalization (HOCMN) for Robust Speech Recognition,” *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP '04)*, Quebec, Canada.

Hsu, C. W. and L. S. Lee (2006), “Extension and Further Analysis of Higher Order Cepstral Moment Normalization (HOCMN) for Robust Features in Speech Recognition,” *Interspeech'2006 - 9th International Conference on Spoken Language Processing (ICSLP)*, Pittsburgh, Pennsylvania.

Huang, X., A. Acero, et al. (2001), “Spoken Language Processing: A Guide to Theory, Algorithm and System Development,” Upper Saddle River, NJ, USA, Prentice Hall PTR.

Hung, J. W., J. L. Shen, et al. (2002), “New Approaches for Domain Transformation and Parameter Combination for Improved Accuracy in Parallel Model Combination (PMC) Technologies,” *IEEE Transactions on Speech and Audio Processing,* vol. 9(8): pp. 842-855

Huo, Q., C. Chany, et al. (1995), “Bayesian Adaptive Learning of the Parameters of Hidden Markov Model for Speech Recognition,” *IEEE Transaction on Speech and Audio Processing,* vol. 3(4): pp. 334-345.

Huo, Q. and D. Zhu (2006), “A Maximum Likelihood Training Approach to Irrelevant Variability Compensation Based on Piecewise Linear Transformations,” *Interspeech'2006 - 9th International Conference on Spoken Language Processing (ICSLP),* Pittsburgh, Pennsylvania.

Josifovski, L., M. Cooke, et al. (1999), “State Based Imputation of Missing Data for Robust Speech Recognition and Speech Enhancement,” *Interspeech'1999 - 6th European Conference on Speech Communication and Technology (Eurospeech),* Budapest, Hungary.

Juang, B. H. and S. Frui (2000), “Automatic Recognition and Understanding of Spoken Language— A First Step Toward Natural Human-Machine Communication,” *Proceedings of the IEEE,* vol. 88(8): pp. 1142-1165.

Junqua, J. C., J. P. Haton, et al. (1996), “Robustness in Automatic Speech Recognition,” Norwell, MA:Kluwer.

Koo, J. D. Gibson, et al. (1989), “Filtering of Colored Noise for Speech Enhancement and Coding,” *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP '89),* Glasgow, Scotland.

Kumar, N. (1997), “Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition,” Ph. D. Dissertation, John Hopkins University.

Lee, L. S. and B. Chen (2005), “Spoken Document Understanding and Organization,” *IEEE Signal Processing Magazine (IEEE SPM),* vol. 22(5): pp. 42-60.

Leggeter, C. J. and P. C. Woodland (1995), “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models,” *Computer Speech and Language*, vol. 9: pp. 171-185.

Lin, S. H., S.-H. Liu, et al. (2007a), “Improved Histogram Equalization (HEQ) for Robust Speech Recognition,” *IEEE International Conference on Multimedia & Expo (ICME 2007)*, Beijing, China.

Lin, S. H., Y. M. Yeh, et al. (2006a), “Exploiting Polynomial-Fit Histogram Equalization and Temporal Average for Robust Speech Recognition,” *Interspeech'2006 - 9th International Conference on Spoken Language Processing (ICSLP)*, Pittsburgh, Pennsylvania.

Lin, S. H., Y. M. Yeh, et al. (2006b), “An Improved Histogram Equalization Approach for Robust Speech Recognition,” *ROCLING XVIII: Conference on Computational Linguistics and Speech Processing (ROCLING 2006)*, Hsinchu, Taiwan.

Lin, S. H., Y. M. Yeh, et al. (2007b), “Cluster-based Polynomial-Fit Histogram Equalization (CPHEQ) for Robust Speech Recognition,” *Interspeech'2007 - 10th European Conference on Speech Communication and Technology (Eurospeech)*, Antwerp, Belgium.

Mika, S. (1999), “Fisher Discriminant Analysis With Kernels,” *IEEE International Workshop on Neural Networks for Signal Processing (NNSP 1999)*, Madison, Wisconsin.

Molau, S. (2003), “Normalization in the Acoustic Feature Space for Improved Speech Recognition,” Ph. D. Dissertation, RWTH Aachen University.

Molau, S., F. Hilger, et al. (2003), “Feature Space Normalization in Adverse Acoustic Conditions,” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, Hong Kong.

Molau, S., M. Pitz, et al. (2001), “Histogram Based Normalization in the Acoustic Feature Space,” *IEEE Workshop on Automatic Speech Recognition and*

Understanding (ASRU '01), Trento, Italy.

Montgomery, D. C., E. A. Peck, et al. (2006), "Introduction to Linear Regression Analysis," Wiley-Interscience.

Neumeyer, L. and M. Weintraub (1994), "Probabilistic Optimum Filtering for Robust Speech Recognition," *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP '94)*, Albuquerque, New Mexico.

Palomaki, K. J., G. J. Brown, et al. (2004), "A Binaural Processor for Missing Data Speech Recognition in the Presence of Noise and Small-Room Reverberation," *Speech Communication*, vol. 43(4): pp. 361-378.

Pujol, P., D. Macho, et al. (2006), "On Real-Time Mean-and-Variance Normalization of Speech Recognition Features," *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP '06)*, Toulouse, France.

Raj, B. (2000), "Reconstruction of Incomplete Spectrograms for Robust Speech Recognition. ECE Department. Pittsburgh," Ph. D. Dissertation, Carnegie Mellon University.

Raj, B., M. L. Seltzer, et al. (2004), "Reconstruction of Missing Features for Robust Speech Recognition," *Speech Communication*, vol. 43(4): pp. 275-296.

Raj, B. and R. M. Stern (2005), "Missing-feature Approaches in Speech Recognition," *Signal Processing Magazine*, vol. 22(5): pp. 101-116.

Saon, G., M. Padmanabhan, et al. (2000), "Maximum Likelihood Discriminant Feature Spaces," *IEEE International Conference on Acoustics, Speech, Signal processing (ICASSP '00)*, Istanbul, Turkey.

Segura, J. C., C. Benitez, et al. (2004), "Cepstral Domain Segmental Nonlinear Feature Transformations for Robust Speech Recognition," *IEEE Signal Processing Letters*, vol. 11(5): pp. 517-520.

Suk, Y. H., S. H. Choi, et al. (1999), "Cepstrum Third-Order Normalisation Method for Noisy Speech Recognition," *Electronics Letters*, vol. 35(7): pp. 527-528.

Torre, A., A. M. Peinado, et al. (2005), “Histogram Equalization of Speech Representation for Robust Speech Recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 13(3): pp. 355-366.

Torre, A., J. C. Segura, et al. (2002), “Non-Linear Transformations of the Feature Space for Robust Speech Recognition,” *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP '02)*, Orlando, Florida.

Vikki, A. and K. Laurila (1998), “Segmental Feature Vector Normalization for Noise Robust Speech Recognition,” *Speech Communication*, vol. 25: pp. 133-147.

Vizinho, A., P. Green, et al. (1999), “Missing Data Theory, Spectral Subtraction and Signal-to-Noise estimation for Robust ASR,” *Interspeech'1999 - 6th European Conference on Speech Communication and Technology (Eurospeech)*, Budapest, Hungary.

Wan, C. Y., Y. Chen, et al. (2007), “Three-Stage Error Concealment for Distributed Speech Recognition (DSR) with Histogram-based Quantization (HQ) under Noisy Environment,” *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP '07)*, Honolulu, Hawai'i.

Wan, C. Y. and L. S. Lee (2005), “Histogram-based Quantization (HQ) for Robust and Scalable Distributed Speech Recognition,” *Interspeech'2005 - 9th European Conference on Speech Communication and Technology (Eurospeech)*, Lisbon, Portugal.

Wan, C. Y. and L. S. Lee (2006), “Joint Uncertainty Decoding (JUD) with Histogram-Based Quantization (HQ) for Robust and/or Distributed Speech Recognition,” *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP '06)*, Toulouse, France.

Wu, J. and Q. Huo (2006), “An Environment-Compensated Minimum Classification Error Training Approach Based on Stochastic Vector Mapping,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14(6): pp. 2147-2155.

Wu, J., Q. Huo, et al. (2005), “An Environment Compensated Maximum Likelihood Training Approach based on Stochastic Vector Mapping,” *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP '05)*, Philadelphia, Pennsylvania.

Xiao, X., H. Li, et al. (2006), “Vector Autoregressive Model for Missing Feature Reconstruction,” *The Fifth International Symposium on Chinese Spoken Language Processing (ISCSLP 2006)*, Singapore.

Young, S., G. Evermann, et al. (2006), “The HTK Book (for HTK Verson 3.4),” Cambridge University.