統計圖等化法於雜訊語音辨識之進一步研究

林士翔[1]　葉耀明[2]　陳柏琳[2]

[1]國立台灣師範大學資訊教育研究所

[2]國立台灣師範大學資訊工程研究所

69308027@cc.ntnu.edu.tw, ymyeh@ice.ntnu.edu.tw, berlin@csie.ntnu.edu.tw

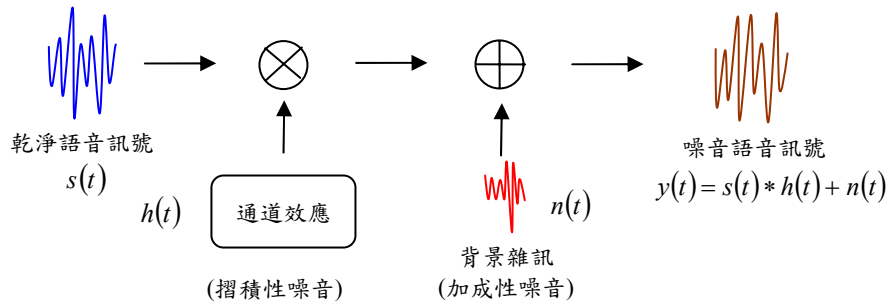
摘要

自動語音辨識系統通常會因語音訊號是否受各種環境雜訊干擾而產生某種程度上的影響。正因如此，語音強健(Speech Robustness)技術的發展長久以來一直被視為一個非常重要的研究領域，過去已有許多方法成功地被提出，可以在抗雜訊上有不錯的效果。其中，統計圖等化法(Histogram Equalization)能有效地補償語音訊號受噪音干擾而所產生的失真情形，因而被公認為非常有效果的方法之一。但前人所提出的統計圖等化法，往往需要大量的記憶體使用空間或是處理器運算時間，本論文探討利用數據擬合(Data Fitting)方法創造一逆函數(Inverse Function)，有效且快速地將測試語句累積密度函數近似至參考分佈的累積密度函數，以正規化雜訊語音特徵，藉由逆函數的使用，能夠節省統計圖等化時所需要的記憶體使用空間以及處理器運算時間。同時，本論文亦探討數據擬合統計圖等化法與時間軸上特徵值移動平均(Moving Average)之結合，來減輕非穩性噪音(Non-stationary Noise)所造成的異常尖峰或波谷的影響。此外，本論文更進一步將所提出的數據擬合統計圖等化法與其他特徵擷取或補償方法進行整合，初步實驗結果證實本論文所提出之方法，能有效補償語音受雜訊干擾所造成的失真情形，進而有效提昇辨識效能。實驗語料庫為由歐洲電信標準協會所發行的AURORA-2 語料，實驗結果初步地顯示數據擬合統計圖等法化確實為一有效的語音強健技術。

1. 序論

現今自動語音辨識 (Automatic Speech Recognition, ASR)系統在語音訊號不受噪音干擾的理想實驗室環境下，可獲得良好的辨識效果，但若要應用至實際日常生活環境中，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環境不匹配(Environment Mismatch)的差異，使得系統辨識效能大幅度降低，環境中複雜因素包括背景噪音(Background Noise)、錄音設備本身產生的噪音或是通道效應(Channel Effect)等。正因如此，語音強健(Speech Robustness)技術長久以來一直被視為重要的研究課題，主要是希望藉由對訊號本身、語音特徵參數或是模型參數做適當的處理與調整，以減緩雜訊干擾的影響、降低訓練環境與測試環境不匹配的情形或提升語音訊號或語音特徵參數本身的強健性，進而提高系統辨識效能。

環境中干擾語音訊號的雜訊可概略分為二種類型：(1)加成性噪音(Additive Noise)和(2)摺積性噪音(Convolutional Noise)。加成性噪音為錄製語音時，原始語音與背景噪音以線性加成(Linearly Additive)的關係同時被收錄進去，例如周遭人聊天的聲音或是機器設備所發出的噪音等；摺積性噪音通常是指語音訊號在經由不同傳輸通道時所產生的通道效應，例如電話線路通道



圖一、雜訊干擾示意圖

效應、麥克風通道效應等。加成性噪音與摺積性噪音對於語音訊號的干擾過程示意圖如圖一所示。

語音強健技術的主要目的就是為了消除不同環境下的差異性以及減輕雜訊對語音訊號的影響，過去已有許多方法成功地被提出，依據方法的本質可概分為以下三種方向[1]：

(1) 語音強化技術(Speech Enhancement)

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)重建出乾淨語音(Clean Speech)訊號。常見的技術有頻譜消去法(Spectral Subtraction, SS)[2]、維爾濾波器(Wiener Filter, WF)[3]等。

(2) 強健性語音特徵(Robust Speech Feature)

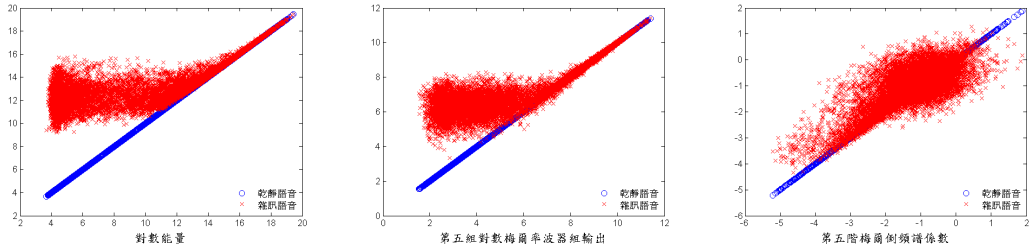
從語音訊號中擷取出較不易受到環境變化干擾而失真的強健性語音特徵參數。常見的技術有倒頻譜平均消去法(Cepstrum Mean Subtraction, CMS)[4]、倒頻譜正規化法(Cepstrum Mean and Variance Normalization, CMVN)[5]等。

(3) 聲學模型調適技術(Acoustic Model Adaptation)

藉由少量的調適語料(Adaptation Data)調整由乾淨語音所訓練而成的聲學模型中的機率分佈參數，如平均值向量(Mean Vector)或共變異矩陣(Covariance Matrix)，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)[6]、最大相似度線性回歸法(Maximum Likelihood Linear Regression, MLLR)[7]等。

本論文中將探討的方法是屬於上述第二類強健性語音特徵。目前，倒頻譜平均消去法(CMS)和倒頻譜正規化法(CMVN)已被廣泛的應用且也被成功地證實能有效的提升辨識效果，其分別是針對語音特徵參數第一階動差(Moment)或是第一階動差與第二階動差進行正規化。但因方法本身線性關係的限制，造成只能補償因受噪音干擾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。因此許多學者嘗試提出許多不同的補償方法，試圖解決因噪音干擾影響對語音特徵參數所產生的失真情形。例如[8]針對語音特徵參數的第三階動差進行正規化或[9]對語音特徵參數更高階動差進行正規化。此外，近年來亦有學者嘗試將在影像處理中已行之有年的統計圖等化法(Histogram Equalization)應用於語音辨識[10]。

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，更企圖使訓練語料和測試語料能夠具有相同的統計分佈特性，其作法是藉由將測試語料(Test Speech)的累積密度函數(Cumulative Density Function, CDF) 對應至由訓練語料(Training Speech)所統計出來的參考分佈(Reference Distribution)的累積密度函數，藉由此匹配轉換過程，降低測試



圖二、加成性噪音對語音特徵參數的影響

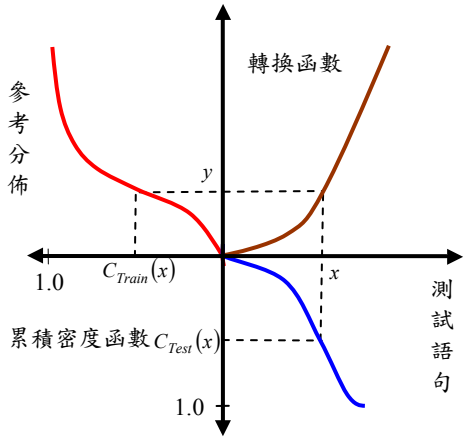
語料與訓練語料由於環境因素影響所造成統計特性不匹配的現象，實驗結果亦證實統計圖等法化對提升辨識效果有很明顯的幫助。另外在[11][12]中，更嘗試將統計圖等化法概念推廣至向量量化編碼(Vector Quantization)，進而應用於分散式語音辨識 (Distributed Speech Recognition, DSR)上，主要是利用統計圖資訊做為向量量化準則，有效解決傳統以距離為量化準則容易受環境噪音影響或是容易形成量化失真(Quantization Distortion)的問題。

雖然統計圖等化法近年來已被廣泛的應用於討論，但仍然有許多可以改善的地方，例如查表式統計圖等化法(Table Look-Up based Histogram Equalization, THEQ)[10]需要將龐大的表格資訊載入記憶體中，方能進行轉換匹配動作，且若要有良好的補償效果，表格所紀錄的點數不能太少，但當表格紀錄點數增加時，同時意謂著需耗費更大量的記憶體使用空間與進行查表轉換的處理器運算時間；又如分位差統計圖等化法(Quantile-based Histogram Equalization, QHEQ)[13,14]，雖然轉換過程不需透過查表動作，只需使用少量的參數即可進行等化動作，但是對每一句待轉換的語句在進行轉換動作前，必須利用格式搜尋(Grid Search)以線上即時運算求取參數，因此所需的處理器運算時間也是相當可觀的。

基於上述原因，本論文提出利用數據擬合(Data Fitting)的概念求算累積密度函數的逆函數，藉由少量的多項式係數與多項式函數的運用，達到具有和統計圖等化法相同的補償效果，同時也探討時間序列上特徵值移動平均法的使用，解決因等化過程中某些特徵值被異常的放大或縮小。本論文後續章節安排如下：第二章將介紹傳統查表式統計圖等化法與分位差統計圖等化法；第三章則介紹數據擬合統計圖等化法及時間序列上特徵值移動平均法的使用；第四章為實驗與討論；第五章結論。

2. 文獻回顧

雜訊干擾會使得語音訊號產生非線性失真，例如加成性噪音對訊號的對數能量影響，在高能量的區域，只有輕微的影響，相反地，在能量強度較低的區域則會有嚴重的失真情形，此一情形，即為造成乾淨語音訊號和雜訊語音訊號二者間統計特性差異的主要原因之一。雜訊干擾對於乾淨語音所造成的非線性失真情形如圖二所示，其中藍色散佈點數的描繪是利用乾淨語音的特徵值當做X軸參考座標值與Y軸參考座標值；紅色散佈點數的描繪是以乾淨語音的特徵值當做X軸參考座標值，雜訊語音的特徵值為Y軸參考座標值，圖片從左至右分別是作用在對數能量(Log Energy)、對數梅爾濾波器組(Mel Filter-Bank)輸出以及梅爾倒頻譜係數(Mel-Frequency Cepstral Coefficient)。由圖中可觀察發現傳統的補償方法諸如倒頻譜平均消去法或倒頻譜正規化法因本身線性特性所限制，可能會使得對於雜訊干擾所造成的非線性失真部份補償效果非常有限。因此，過去幾年有許多學者提出各種方法來補償此非線性失真的情形，其中統計圖等化法為有非常顯著



圖三、統計圖等化法示意圖

補償效果的方法之一[10][15]，下面章節將分述傳統查表式統計圖等化法(THEQ)與分位差統計圖等化法(QHEQ)的概念並分析各方法的優點和缺點。

2.1 統計圖等化法(Histogram Equalization, HEQ)

統計圖等化法假設測試語句之語音特徵參數的統計分佈會和訓練語料語音特徵參數的統計分佈(或稱作參考分佈)是一致的，若以目前較常用的語音特徵參數－梅爾倒頻譜係數(Mel- Frequency Cepstral Coefficients, MFCC)而言，統計圖等化法可以作用在對數梅爾濾波器組輸出[16][17][18]或是梅爾倒頻譜係數[10][19][20]上。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佈分別轉換至先前已從訓練語句中定義好的對應參考分佈，數學式關係式表示如下[19][20]：假設 x 為測試語句語音特徵向量的某一維特徵參數，且具有機率密度函數(Probability Density Function, PDF) $p_{Test}(x)$，那麼轉換函數 $F(x)$ 可依照下列的數學式將 x 轉換成在訓練語料所對應到的 y，並且讓 $p_{Train}(y)$ 與 $p_{Test}(x)$ 能有式(1)的關係：

$$p_{Train}(y)=p_{Test}(x)\frac{dx}{dy}=p_{Test}\left(F^{-1}(y)\right)\frac{d\left(F^{-1}(y)\right)}{dy} \tag{1}$$

其中 $F^{-1}(y)$ 為 $F(x)$ 的逆函數(Inverse Function)，若上述關係式以累積機率密度函數(Cumulative Probability Density Function, CDF)的觀點表達即為

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

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

$$F(x)=C_{Train}^{-1}(C_{Test}(x)) \tag{3}$$

其中 C_{Train}^{-1} 為 C_{Train} 的逆函數，轉換過程如圖三所意示。

在實作上，因為語音特徵參數為一有限集合，所以並無法非常精準估算實際的累積密度函

數，通常實作會使用累積統計圖(Cumulative Histogram)去近似累積密度函數。對於所有訓練語料而言，語音特徵向量的每一維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(Quantiles)，每個分位差區間皆以區間內所有特徵值的平均數(Mean)做為該分位差的代表特徵值，此資訊可被用來當作轉換的參考分佈。對測試語句語音特徵向量的每一維度同樣統計出累積統計圖，也取 i 個分位差，接著每個分位差區間內的特徵值用先前建立好的參考分佈逐一進行轉換取代。一般實作可利用表格查詢的方式進行：首先，以表格方式紀錄參考分佈的累積統計圖資訊，例如記錄成｛區間，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(Table-Lookup)轉換，故可稱為查表式統計圖等化法(Table Look-Up based Histogram Equalization, THEQ)。但是往往為了要得到良好的辨識效果，使用的分位差個數不可太少，即代表需耗費大量的記憶體空間紀錄表格資訊；並且在進行查表轉換時，也需花費不少的表格搜尋時間。

2.2 分位差統計圖等化法(Quantile-based Histogram equalization, QHEQ)

前一章節所介紹的查表式統計圖等法化為一種非參數(Nonparametric)型態的統計圖等化法，所有的等化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參數，在[19][20]中提出一種參數型態的分位差統計圖等化法，對於語音特徵向量的每一維可利用一轉換函數 $H(x)$ 進行等化動作，數學關係式表示如下

$$H(x)=Q_K\left(\alpha\left(\frac{x}{Q_K}\right)^{\gamma}+(1-\alpha)\left(\frac{x}{Q_K}\right)\right) \tag{4}$$

Q_K 為最後一個分位差值，亦即整句語句中在此一維特徵參數中最大的特徵值；α 和 γ 為轉換函數 $H(x)$ 所需的參數可利用式(5)求得，值得注意的是在對於每一句語句在進行等化過程前，需先將整句語句與參考分佈進行分位差校正(Quantile Correction)，以求得最佳的參數，此校正動作是以最小平方誤差(Minimum Mean Square Error, MMSE)法進行，可以利用格式搜尋法，將一段區間內的 α 和 γ 以等距的數值代入式(5)，找出最佳的 α 和 γ。

$$\{\alpha,\gamma\}=\arg\min_{\{\alpha,\gamma\}}\left(\sum_{k=1}^{K-1}\left(H(Q_k)-Q_k^{train}\right)^2\right) \tag{5}$$

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佈中的第 k 個分位差值。

分位差統計圖等化法是經由式(5)計算以求得參數 α 和 γ，接著再利用式(4)一組非線性函數和一組線性函數進行加權平均(Weight Average)，欲使轉換後的語音特徵參數的統計分佈能夠和參考分佈愈相似，亦即受雜訊干擾而形成的非線性失真部份，可藉由 γ 項的使用進行補償。但針對每一轉換語句都必經由式(5)線上即時求得最佳的參數 α 和 γ，因此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。

3.統計圖等化法之改良

前面章節所描述的統計圖等化法雖然能有效補償因雜訊干擾而所產生的非線性失真情形，但無論是傳統查表式統計圖等化法或是分位差統計圖等化法，往往在執行等化的過程，需耗費大量的記

憶體使用空間或是處理器運算時間。為了能有效的解決此問題，我們提出利用數據擬合的概念求得累積密度函數的逆函數，藉由少量的多項式函數與多項式係數的使用，達到具有和統計圖等化法相同的補償效果[21]，整體概念和作法敘述如下。

3.1 多項式擬合統計圖等化法 (Polynomial-Fit Histogram Equalization, PHEQ)

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)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。通常迴歸公式$G(u_i)$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，換言之將所有u_i分別代入迴歸公式所求得的預測值$\tilde{v}_i$和實際觀測值v_i的誤差值平方和必須最小，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)。假設$G(u_i)$為M階的線性多項式函數：

$$\tilde{v}_i = G(u_i) = a_0 + a_1 u_i + a_2 u_i^2 + \cdots + a_M u_i^M = \sum_{m=0}^{M} a_m u_i^m \tag{6}$$

$a_0, a_1, \ldots, a_M$為多項式的係數(Coefficients)，則所對應的誤差平方合E^2定義成

$$E^2 = \sum_{i=1}^{N} \left(v_i - \sum_{m=0}^{M} a_m u_i^m \right)^2 \tag{7}$$

同理，此概念可延伸為求得參考分佈之累積密度函數的逆函數，我們稱為多項式擬合統計圖等化法 (Polynomial-Fit Histogram Equalization, PHEQ)。對於訓練語料語音特徵向量的每一維皆可利用一個多項式迴歸表示，以特徵值y_i為反應變數以及y_i對應的累積密度值$C_{Train}(y_i)$為解釋變數，則式(7)可重新表示成

$$G(C_{Train}(y_i)) = \tilde{y}_i = \sum_{m=0}^{M} a_m (C_{Train}(y_i))^m \tag{8}$$

並且誤差平方合E'^2定義為

$$E'^2 = \sum_{i=1}^{N} \left(y_i - \sum_{m=0}^{M} a_m (C_{Train}(y_i))^m \right)^2 \tag{9}$$

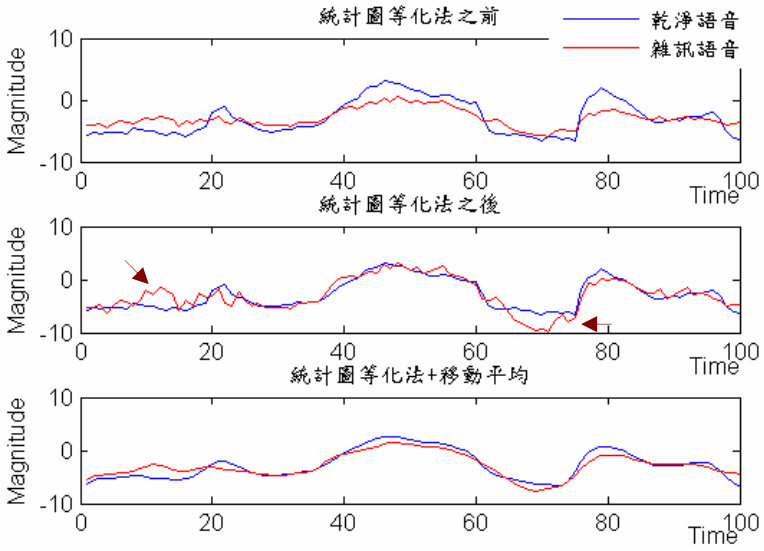
其中N為訓練語料中所有音框(Frame)的個數，若要使誤差平方合最小，則所有多項式係數$a_0, a_1, \ldots, a_M$會滿足式(10)的關係，只需透過解聯立方程式，即可求得$a_0, a_1, \ldots, a_M$係數。

$$\frac{\partial E'^2}{\partial a_m} = 0, \forall m = 1 \ldots M \tag{10}$$

在實作上，對轉換語句中語音特徵向量的每一維$Y = [y_1, y_2, \ldots y_N]$而言，每個時間點上，特徵值所對應的累積密度值可利用下列步驟近似而得：

步驟一、將Y以遞增的方式做排序得到資料序列S

步驟二、對於Y中的每個特徵值y_i而言，可由下式近似其所對應的累積機率密度函數值



圖四、非穩性噪音所造成的異常尖峰或波谷示意圖

$$C(y_i) \approx \frac{S_{pos}(y_i)}{N} \tag{11}$$

其中 $S_{pos}(y_i)$ 為一指示函數，表示特徵值 y_i 在排序後的資料序列 S 中的位置，N 為資料序列中所有音框個數。在訓練階段時，可以利用式(11)求得所有訓練語料的累積密度函數，再利用式(8)、(9)和(10)求得累積密度函數的逆函數的係數；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_i 的對應累積密度函數值 $C(y_i)$ 代入先前已於訓練階段中求得的多項式函數(式(8))即可進行等化動作。

因此，本論文所提出數據擬合的使用，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佈轉換至先前已從訓練語句中定義好的參考分佈，並且能擁有和統計圖等化法相同的補償效果。

3.2 時間序列上特徵值移動平均(Moving Average, MA)

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，但值得注意的是，由非穩性噪音(Non-stationary Noise)所造成的異常尖峰(Sharp Peak)或波谷(Valley)，可能造成在等化的過程中，某些特徵值被過度的放大或縮小，此異常情形如圖四所示，最上方的圖為尚未做統計圖等化法前乾淨語音與雜訊語音倒譜頻特徵向量的第二維；中間的圖為做完統計圖等化法後的倒譜頻特徵向量的第二維特徵值，可清楚看見有二個區域被過度強調；最下方的圖為做完統計圖等化法加移動平均後的倒譜頻特徵向量。

移動平均在語音辨識的研究上，已非一個全新的議題，例如[22]利用移動平均的概念提出一種特徵向量正規化(Feature Normalization)的方法，首先對語音特徵向量進行平均消去法(Mean Subtraction)和變異數正規化(Variance Normalization)，接著再利用自動迴歸移動平均(Auto-Regression Moving Average, ARMA)對特徵向量進行正規化的動作，其實驗結果亦證實移動平均的使用對於提升整體辨識率有很大的幫助。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[22]：

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

$$\hat{y}_t = \begin{cases} \dfrac{\sum_{i=-L}^{L} \tilde{y}_{(t+i)}}{2L+1} & if\ L < t \le T-L, \\ \tilde{y}_t & otherwise \end{cases} \tag{12}$$

- 因果關係自動迴歸移動平均(Casual Moving Average)

$$\hat{y}_t = \begin{cases} \dfrac{\sum_{i=0}^{L} \tilde{y}_{(t-i)}}{L+1} & if\ L < t \le T, \\ \tilde{y}_t & otherwise \end{cases} \tag{13}$$

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

$$\hat{y}_t = \begin{cases} \dfrac{\sum_{i=1}^{L} \hat{y}_{(t-i)} + \sum_{j=0}^{L} \tilde{y}_{(t+j)}}{2L+1} & if\ L < t \le T-L, \\ \tilde{y}_t & otherwise \end{cases} \tag{14}$$

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

$$\hat{y}_t = \begin{cases} \dfrac{\sum_{i=1}^{L} \hat{y}_{(t-i)} + \sum_{j=0}^{L} \tilde{y}_{(t-j)}}{2L+1} & if\ L < t \le T, \\ \tilde{y}_t & otherwise \end{cases} \tag{15}$$

其中 $\tilde{y}_i$ 為輸入的語音特徵值，$\hat{y}_i$ 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）。

3.3 結合等化前與等化後的特徵值資訊

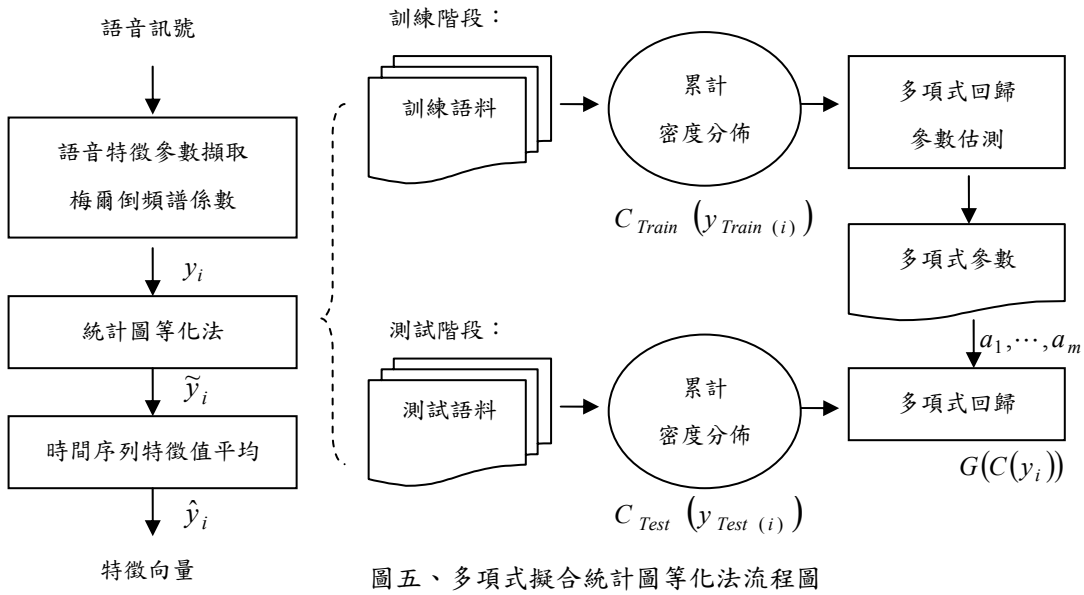
除了上述時間序列上進行特徵值移動平均的方法外，從實際雜訊語音特徵參數中，經由觀察發現等化前與等化後的特徵值，在特徵值受雜訊干擾不嚴重時，原本的特徵值與等化後的特徵值差異不大，相反地，特徵值可能會因等化動作被異常的放大或縮小，進而造成原本語音特徵所帶有資訊被破壞掉，因此可使用加權平均來補償此一異常情形，數學式如式(16)所示。當特徵值未受雜訊嚴重干擾時 $\bar{y}_i \approx \hat{y}_i$，相反地，可依造 α 的設定，決定 $\bar{y}_t$ 較保留等化前或等化後的資訊，此法概念與分位差統計統等法化[19][20]在概念上非常相似，藉由對加權數 α 值的控制，在線性與非線性補償中取得一平衡點，

$$\bar{y}_t = (1-\alpha) \times y_t + \alpha \times \hat{y}_t \tag{16}$$

y_i 為原本的特徵值，$\hat{y}_i$ 為做完等化後的特徵值。α 值的設定，除了可以利用窮舉的方法，逐一調動 α 值，以取得最佳的辨識效果外，亦可使用數據擬合的概念求算而得，誤差平方 E'' 定義如下：

$$E'' = \sum_{i=1}^{N} \left(\left(\bar{y}_{Noisy(i)} - \bar{y}_{Clean(i)} \right)^2 \right) \tag{17}$$

其中 $\bar{y}_{noisy(i)}$ 為雜訊語音第 i 個音框經由式(16)求得後的新的特徵值，$\bar{y}_{clean(i)}$ 為雜訊語音相對應的乾淨語音經由式(16)求得的特徵值，但值得注意的是，式(17)誤差平方計算程程中，為了避免整體誤差被部份異常的錯誤(Outlier)所支配，適當的利用門檻值(Threshold)將異常特徵值排除是須要的。



圖五、多項式擬合統計圖等化法流程圖

4. 實驗與討論

4.1 實驗架構與設定

本論文實驗所使用的語料庫 Aurora-2 是由歐洲電信標準協會(European Telecommunications Standards Institute, ESTI)所發行[23]，其本身為一套含有雜訊的連續英文數字語料，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)各自加入乾淨的語音裡，訊噪比包括 20dB、15dB、10dB、5dB、0dB 和-5dB；通道包含由國際電信聯合會所訂立的二個標準-G.712 和 MIRS。根據測試語料中加入之通道雜訊以及加成性雜訊之種類不同，Aurora-2 分為三組測試群組 Set A、Set B 和 Set C，Set A 所呈現的雜訊是屬於穩性(Stationary)雜訊，Set B 則是非穩性(Nonstationary)雜訊，Set C 除了穩性與非穩性雜訊外，還使用與訓練語料不同的通道效應。

在聲學模型(Acoustic Models)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(Continuous Density Hidden Markov Model, CDHMM)表示，其中包含 16 個狀態(State)，並且每個狀態是利用 3 個高斯混合分佈(Gaussian Mixture Distribution)表示。另外靜音模型的部份有二種模型，一個為靜音(Silence)模型包含三個狀態，用來表示語句開始跟結束時的靜音；另一個為間歇(Pause)模型包含六個狀態，表示語句內字與字之間的短暫停止，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[24]完成。

在前端處理方面(Front-End Processing)，本論文的基礎實驗是採用梅爾倒頻譜係數(Mel-Frequency Cepstral Coefficients, MFCCs)作為語音特徵參數，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)，同時會對13維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)。並且將所提出的多項式擬合統計圖等化法作用在梅爾倒頻譜係數與差量倒頻譜係數上，整體前端處理流程如圖五所示。

表一、多項式擬合統計圖等化法平均字錯誤率實驗結果

		多項式階數			
		3	5	7	9
乾淨語料訓練模式 (Clean-Condition)	所有訓練語料	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
複合情境訓練模式 (Multi-Condition)	所有訓練語料	10.80	10.34	10.43	10.54
	1000組	10.48	10.32	10.40	10.45
	100組	10.73	10.45	10.36	10.45
	10組	11.65	10.61	10.79	11.58

4.1 多項式擬合統計圖等化法實驗

本文第一個實驗是利用多項式迴歸模型描述參考分佈的累積密度函數分佈情形，欲探討使用所有訓練語料與否以及不同多項式階數(Polynomial Order)對於整體辨識效能影響結果如何。其中參考分佈的資訊是由所有訓練語料統計而成的累積統計圖求得，其中累積統計圖所使用的分組組數(Histogram Bins)包括1000組、100組和10組，每一分組皆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(Order)的多項式進行等化動作。辨識結果如表一所示，表格內所呈現的數據皆為平均字錯誤率(Word Error Rate, WER)，是由Aurora-2中三組實驗群組(Sets A, B及C)中不同訊噪比(20dB至0dB)的辨識結果加總平均而得。

值得注意的是，當多項式階數結束行為(End Behavior)的特性，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，所以本文不考慮偶數階數的使用，由表一可清楚看到，辨識效能隨著多項式階數增加有所進步，但並非使用階數愈高愈好，因為資料的散佈情形，可能使高階多項式為了要更符合資料分佈情形而造成過度擬合(Overfit)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。由於使用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，因此下列所有有關多項式迴歸的實驗將使用7階多項式迴歸，並且參考分佈是利用100組的累積統計圖中統計而得。

4.2 時間序列上特徵值移動平均之使用

本小節將探討移動平均的使用，對於減輕由噪音或等化過程中所造成的異常情形，進而提高辨識能的效果如何，實驗結果如表二所示，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。

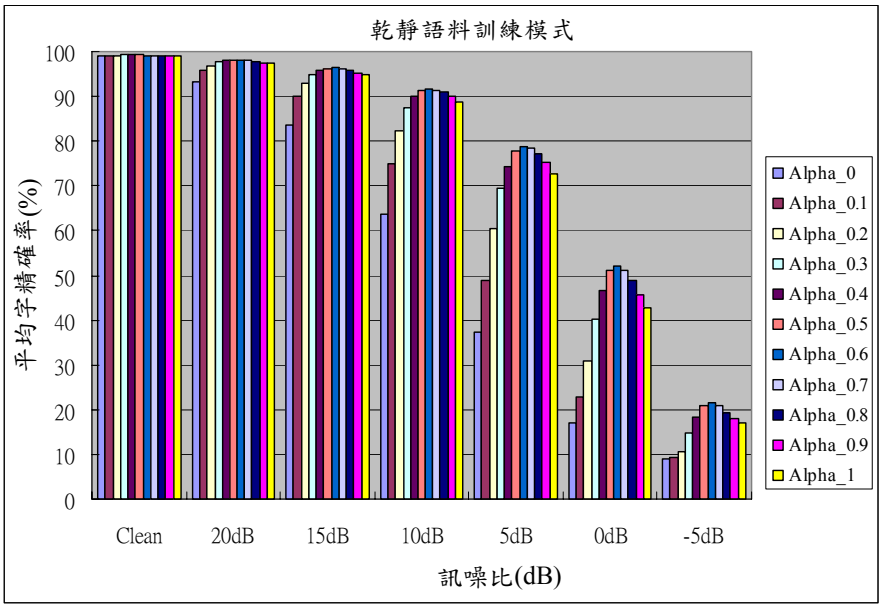
實驗結果和[22]呈現的相同，使用非因果關係自動迴歸移動平均(Non-Casual ARMA)會有較佳的辨識結果，相較於單純使用多項式擬合統計圖等化法而言，乾淨語料訓練模式，字錯誤率可達約 20%的相對進步(Relative Improvement)，對複合情境訓練模式也可有約 5%的相對進步。但是，若移動平均項的階數若使用太高，可能會造成原本帶有鑑別資訊的特徵值，因此被移除掉，使得辨識效果下降。

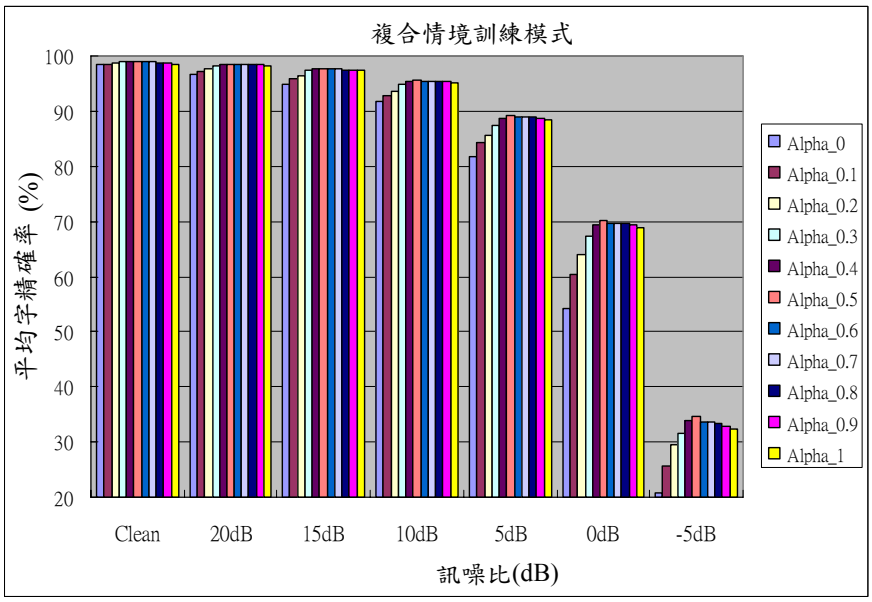
4.3 保留等化前與等化後的特徵值資訊

此小節實驗是根據式(16)，分別以不同的加權值α代入，辨識結果如圖六所示。當$\alpha=1$表

表二、多項式擬合統計圖等化法結合不同移動平均法之平均字錯誤率實驗結果

字錯誤率(Word Error Rate, WER)		移動平均項					
		0	1	2	3	4	5
乾淨語料訓練模式	Non-Casual MA	20.75	17.75	16.83	17.26	18.15	19.66
	Casual MA	20.75	19.23	18.28	17.44	17.12	17.28
	Non-Casual ARMA	20.75	17.83	16.90	16.38	16.99	17.34
	Casual ARMA	20.75	17.93	16.84	19.20	17.44	19.20
複合情境訓練模式	Non-Casual MA	10.36	9.88	9.88	10.24	10.94	11.69
	Casual MA	10.36	10.13	9.74	9.76	9.78	10.12
	Non-Casual ARMA	10.36	9.88	9.78	9.84	9.94	10.11
	Casual ARMA	10.36	9.95	9.71	10.84	9.76	10.68





圖六、保留等化前與等化後的特徵值資訊之平均字精確率實驗結果

表三、本論文所提出的多項式擬合統計圖等化法與與其他正規化補償技巧之比較

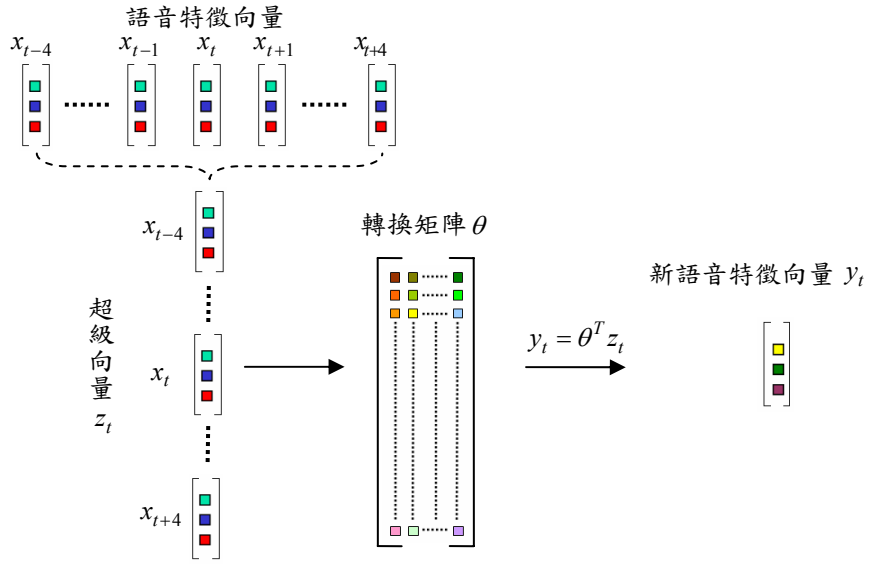
		平均字錯誤率 WER(%)			
		Set A	Set B	Set C	平均
乾淨語料訓練模式	MFCC	41.06	41.52	40.03	41.04
	AFE	38.69	44.25	28.76	38.93
	CMVN	27.73	24.60	27.17	26.37
	MS+VN+ARMA(3)	18.38	16.14	21.81	18.17
	THEQ	19.72	18.57	19.24	19.16
	QHEQ	23.53	21.90	22.36	22.64
	PHEQ	20.98	20.17	21.43	20.75
	PHEQ+MA	16.83	15.10	20.02	16.78
	PHEQ+α+MA	16.19	15.17	19.72	16.49
複合情境訓練模式	MFCC	14.78	16.01	19.33	16.18
	AFE	10.64	10.76	12.85	11.13
	CMVN	12.70	12.45	14.52	12.98
	MS+VN+ARMA(3)	9.49	10.37	10.06	9.95
	THEQ	10.02	10.41	10.34	10.24
	QHEQ	10.20	10.75	10.76	10.53
	PHEQ	9.91	9.41	13.14	10.36
	PHEQ+MA	9.41	9.53	11.21	9.82
	PHEQ+α+MA	9.15	9.08	11.53	9.60

示完全使用多項式擬合統計圖等化法所得到的辨識效果，相反地，當$\alpha=0$時，為未使用多項式擬合統計圖等化法所得到的辨識效果，從圖中可得知隨著訊噪比降低，若只單純使用多項式擬合統計圖等化法可能會有部份特徵值因等化過程而被異常放大或縮小的情況，導致辨識效果降低，因此適當保留等化前的特徵值對辨識效能能有所提昇。

因為此方法跟前面小節所敘述的時間序列上特徵值移動平均皆具有特徵值平滑(Smoothing)的效果，因此若要同時要使用此二種平均法，可能會跟使用高階移動平均項存在著相同的問題，原本帶有鑑別資訊的特徵可能因此被平滑掉，所以本文建議先以式(16)進行等法前與等化後的特徵值加權平均，再搭配式(12)~(15)低階的移動平均使用。

4.4 與其他正規化補償方法之比較

此章節將本論文所提出的方法與其他現有的正規化補償技巧進行比較，包括歐洲電信標準協會的標準前端特徵擷取(Advanced Front-End Processing, AFE)、倒頻譜正規化法(CMVN)、查表式統計圖等化法(THEQ)、分位差統計圖等化法(QHEQ)、3 階移動平均的使用(MS+VN+ARMA)以及本論文所提出的方法(PHEQ)、搭配 3 階非因果關係自動迴歸移動(PHEQ+MA)與採用α設定為 0.6 搭配 1 階非因果關係自動迴歸移動平均(PHEQ+α+MA)，其中查表式統計圖等化法和分位差統計圖等化法的實驗結果分別直接採用[18]和[14]的實驗數據結果，實驗結果如表三所示，本論文所提出的多項式擬合統計圖等化法若與單純的梅爾倒頻譜係數或是歐洲電信標準協會的標準前端特徵擷取或是倒頻譜正規化法都有明顯進步，並且和傳統查表式統計圖等化法或是分位差統計圖



圖七、鑑別性特徵擷取法示意圖

等化法的補償效果不分軒輊，若適當的加入時間序列上特徵值平均的使用(PHEQ+MA與PHEQ+α+MA)，辨識效果則有更顯著的進步。

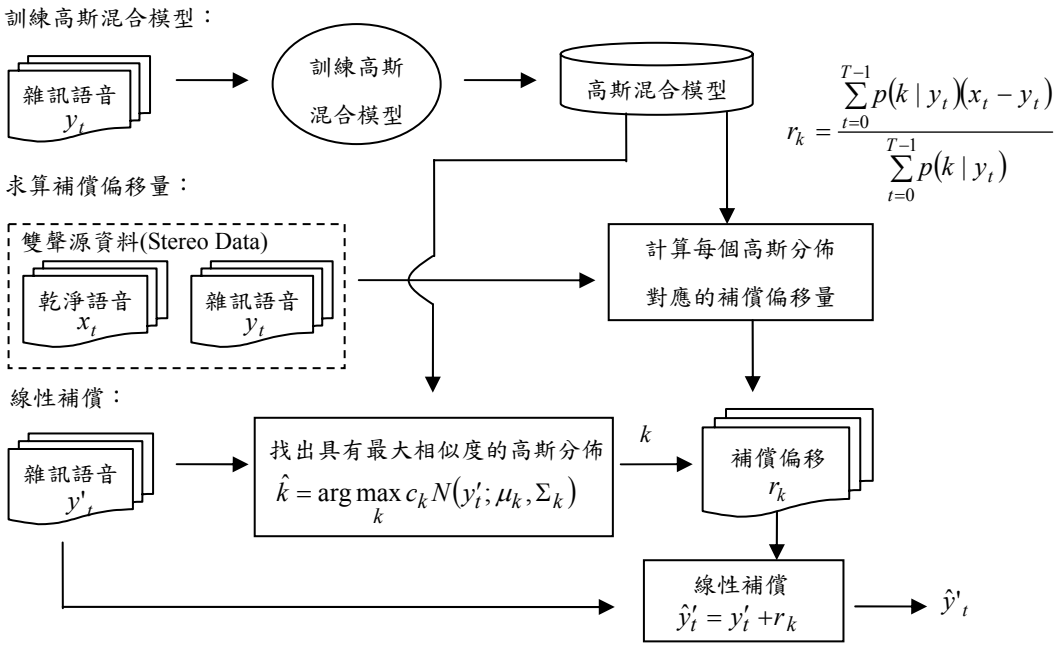
4.5 結合其他特徵擷取或補償方法

最後本論文嘗試將所提出的多項式擬合統計圖等化法與鑑別性特徵擷取法和雙聲源為基礎之分段線性補償(Stereo-based Piecewise Linear Compensation, SPLICE)[25]進行結合。在鑑別性特徵擷取法我們使用異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA) [26]加上最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)[27]並且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(Discrete Cosine Transform, DCT)達到各維度特徵向量部份解相關(Partial Decorrection)的效果，詳細數學推導可參考[28]，整體語音特徵擷取示意圖如圖七所示，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4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 經由異質性線性鑑別分析與最大相似度線性轉換的基底矩陣 θ 進行線性轉換後，可得新語音特徵向量 y_t ，數學式表示如下：

$$y_t = \theta^T z_t \tag{18}$$

最後再以多項式擬合統計圖等化法進行等化動作，實驗結果如表四所示第一列和第二列所示，無論是乾淨語料訓練模式或是複合情境訓練模式都比倒頻譜正規化有更明顯的補償效果，其中以乾淨語料訓練模式而言，與倒頻譜正規化法相比約有25%的相對進步。

另外在與雙聲源為基礎之分段線性補償結合的實驗，因此法有個前題是需擁有雙聲源的語音訊號，正好依照Aurora-2的設定，此雙聲源可從自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，整個方法流程如圖八所示，首先雜訊語料須先訓練出一個高斯混合模型，在論文設為512個高斯分佈，對於每個高斯分佈會計算相對應的補償偏移量 r_k ，運算如下[25]：



圖八、雙聲源為基礎之分段線性補償流程圖

$$r_k = \frac{\sum_{t=1}^{N} p(k \mid y_t)(x_t - y_t)}{\sum_{t=1}^{N} p(k \mid y_t)} \tag{18}$$

其中 N 為所有訓練語料音框個數，y_t 為時間點 t 含雜訊的訓練語音特徵向量，x_t 為相對應的乾淨訓練語音特徵向量，k 表示高斯混合模型中第 k 個高斯分佈。而測試語音特徵向量 y'_t 的補償後測試語音特徵向量 $\hat{y}'_t$ 可由下式二個步驟求得：

$$\begin{aligned} \hat{k} &= \arg\max_k c_k N(y'_t; \mu_k, \Sigma_k) \\ \hat{y}'_t &= y'_t + r_{\hat{k}} \end{aligned} \tag{19}$$

第一個步驟是找出 y'_t 和高斯混合分佈中具有最大相似度(Likelihood)的高斯分佈 k ，因為高斯混合模型是經由雜訊語料訓練而成，因此我們可以視每一個高斯分佈是代表某一種類型與訊噪比的噪音，此步驟即找出最相似的高斯分佈來進行補償，接著再以由式(18)所事先求得的補償偏移量進行補償。實驗結果如表四第三列和第四列所示，其中因為Set C是包含與訓練語料不同通道效應的測試語料，在求算補償偏移量時沒有被考慮到，因此補償效果較有限，實驗結果亦證明雙聲源為基礎之分段線性補償結合本論文所提出的方法會比結合倒頻譜正規化有更佳的補償效果。

5. 結論

本論文成功利用數據擬合的方法創造一逆函數，能有效且快速的將測試語句累積密度函數近似至參考資料的累積密度函數，藉由逆函數的使用，成功地改善傳統統計圖等化法或分位差統計圖等化法需要耗費大量記憶體使用空間或是處理器運算時間的缺點，同時也探討時間序列上特徵值移動平均法對於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。實驗結果清楚的呈現本論文所提出的特徵值正規化法對噪音環境下的語音有很顯著的幫助。此外，本論文也嘗試和其他特徵擷取或補償方法進行結合，實驗結果亦呈

表四、整合其他特徵擷取或補償方法之實驗結果

		平均字錯誤率 WER(%)			
		Set A	Set B	Set C	平均
乾淨語料訓練模式	HLDA-MLLT+CMVN	21.63	21.37	21.59	21.52
	HLDA-MLLT+PHEQ-MA	15.98	15.96	15.91	15.96
	SPLICE+CMVN	16.34	14.95	21.18	16.75
	SPLICE+PHEQ-MA	13.40	13.41	17.08	14.14
複合情境訓練模式	HLDA-MLLT+CMVN	9.49	9.51	10.40	9.68
	HLDA-MLLT+PHEQ-MA	9.06	8.87	8.55	8.88
	SPLICE+CMVN	10.40	11.00	13.80	11.32
	SPLICE+PHEQ-MA	9.54	10.88	12.18	10.60

現補償效果比倒頻譜正規化法有更顯著的效果，與 HLDA+MLLT 的結合，在複合情境訓練模式下，有最佳的辨識效果，平均字錯誤率達 8.88%；另外與 SPLICE 結合，在乾淨語料訓練模式下，平均字錯誤率達 14.14%。

6. 參考文獻

[1] Y. Gong, “Speech Recognition in noisy environments: A survey,” *Speech communication*, Vol.16, 1995.

[2] S.F. Boll, “Supperssion of Acoutstic Noise in Speech Using Spectral Subtraction,” *IEEE Trans. on ASSP*, Vol.27, No.2, pp.133-120, 1979.

[3] X. Huang, A. Acero and H. Hon, “Spoken Language Processing: A Guide to Theory, Algorithm and System Development,” *Prentice Hall PTR Upper Saddle River*, NJ, USA, 2001.

[4] S. Furui, “Cepstral Analysis Techniques for Automatic Speaker Verification,” *IEEE Trans. on ASSP*, 1981.

[5] A. Viikki and K. Laurila, “Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition,” *Speech Communication*, Vol. 25, 1998.

[6] J.L. Gauian and C.H. Lee, “Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,” *IEEE Trans. on Speech and Audio Processing*, 1994.

[7] C.J. Leggetter and P.C. Woodland, “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models,” *Computer Speech and Language*, 1995.

[8] Y. H. Suk, S. H. Choi, H. S. Lee, “Cepstrum Third-order Normalization Method for Noisy Speech Recognition,” *Electronics Letters*, Vol. 35, no. 7, pp. 527-528, April 1999.

[9] C.W. Hsu and L.S. Lee, “Higher Order Cepstral Moment Normalization (HOCMN) for Robust Speech Recognition,” in *Proc. ICASSP 2004*.

[10] S. Dharanipargda and M. Padmanabhan, “A Nonlinear Unsupervised Adaptation Technique for Speech Recognition,” in *Proc. ICSLP 2000*.

[11] C. Y. Wan and L.S. Lee, “Joint Uncertainty Decoding (JUD) with Histogram-Based Quantization (HQ) for Robust and/or Distributed Speech Recognition,” in *Proc. ICASSP 2006*.

[12] C.Y. Wan and L.S. Lee, “Histogram-based quantization (HQ) for robust and scalable distributed speech recognition,” in *Proc. EUROSPEECH 2005*.

[13] F. Hilger, H. Ney, “Quantile based histogram equalization for noise robust speech recognition,” in *Proc. EUROPSEECH 2001*.

[14] F. Hilger et al., “Quantile Based Histogram Equalization for Noise Robust Large Vocabulary Speech Recognition,” *IEEE Trans. on Speech and Audio Processing*, Vol. 14(3), 2005.

[15] A. de la Torre et al., “Non-linear Transformation of the Feature Space for Robust Speech Recognition,” in *Proc. ICASSP 2002*.

[16] S. Molau et al., “Histogram Based Normalization in the Acoustic Feature Space,” in *Proc. ASRU 2001*.

[17] S. Molau et al., “Feature Space Normalization in Adverse Acoustic Conditions,” in *Proc. ICASSP 2003*.

[18] S. Molau et al., “Histogram Normalization in the Acoustic Feature Space,” in *Proc. ICASSP 2002*.

[19] J. C. Segura et al., “Cepstral domain segmental nonlinear feature transformations for robust speech recognition,” *IEEE Signal Processing Letters*, Vol. 11(5), 2004.

[20] A. de la Torre et al., “Histogram equalization of the speech representation for robust speech recognition,” *IEEE Trans. on Speech and Audio Processing*, Vol. 13(3), 2005.

[21] S.H. Lin, Y.M. Yeh and B. Chen, “Exploiting Polynomial-Fit Histogram Equalization and Temporal Average for Robust Speech Recognition,” in *Proc. ICSLP 2006*.

[22] C.P. Chen, J. Bilmes and K. Kirchhoff, “Low-Resource Noise-Robust Feature Post-Processing on Aurora 2.0,” in *Proc. ICSLP 2002*.

[23] H. G. Hirsch, D. Pearce, “The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions,” in *Proc. ISCA ITRW ASR 2000*.

[24] S. Young et al., “The HTK Book Version 3.3,” 2005.

[25] L. Deng, A. Acero, M. Plumpe and X. Huang. “Large-Vocabulary Speech Recognition under Adverse Acoustic Environments,” in *Proc. ICSLP 2000*.

[26] M. J. F. Gales, “Maximum Likelihood Multiple Projection Schemes for Hidden Markov Models,” *Cambridge University Technical Report RT-365,* 2001.

[27] G. Saon et al., “Maximum Likelihood Discriminant Feature Spaces,” in *Proc. ICASSP 2000*.

[28] 張志豪, “強健性和鑑別力語音特徵擷取技術於大詞彙連續語音辨識之研究,” *國立台灣師範大學資訊工程研究所碩士論文*, 2005.