強健性語音辨識中分頻段調變頻譜補償之研究
A Study of Sub-band Modulation Spectrum Compensation for Robust Speech Recognition

黃勝源 Sheng-yuan Huang
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
s96323530@ncnu.edu.tw

杜文祥 Wen-hsiang Tu
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
aero3016@ms45.hinet.net

洪志偉 Jeih-weih Hung
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
jwhung@ncnu.edu.tw

摘要

雖然語音科技進步迅速，但自動語音辨識仍是一門值得繼續研究開發的課題。因爲目前多數的語音辨識系統應用於不受干擾的安靜環境，雖然能得到相當滿意的辨識效果，但若將其應用於實際的環境中，語音訊號往往會因爲環境雜訊的影響，導致辨識效能有明顯地衰減，發展多年的強健性技術即是針對此項缺點作改進。

在諸多強健性技術中，有一類方法爲對語音特徵作統計上的正規化，傳統上，這些方法都是對全頻段的語音特徵時間序列做正規化處理，然而，在分析此類方法的效能上，通常是以其調變頻譜的正規化程度作爲效能的依據，因此，如果直接在語音特徵之調變頻譜上作正規化，應亦可達到不錯的效果。另外，由於不同頻率的調變頻率成份具有不相等的重要性，但是傳統之特徵時間序列正規化法相對忽略了此性質，基於這些觀察，在本論文中，我們提出了一系列的分頻段調變頻譜統計正規化法，此類方法可以分別正規化不同頻段的統計特性，進而提升語音特徵在雜訊環境下的強健性能；在國際通用的 Aurora-2 連續數字資料庫之語音辨識上，我們所提出的新方法相對於基礎實驗的辨識率而言，可以達到高達 65%的相對錯誤降低率，而這些新的調變頻譜正規化法相對於時間序列正規化法而言，於相對錯誤降低率上也有 7%至 32%的進步空間，此足以驗證這些新方法能夠更有效地提昇語音辨識系統在雜訊環境下的辨識效能。

關鍵詞：語音辨識、調變頻譜、統計正規化、強健性語音特徵參數

Abstract

In this paper, we propose a novel scheme in performing feature statistics normalization techniques for robust speech recognition. In the proposed approach, the processed temporal-domain feature sequence is first converted into the modulation spectral domain. The magnitude part of the modulation spectrum is decomposed into non-uniform sub-band

segments, and then each sub-band segment is individually processed by the well-known normalization methods, like mean normalization (MN), mean and variance normalization (MVN) and histogram equalization (HEQ). Finally, we reconstruct the feature stream with all the modified sub-band magnitude spectral segments and the original phase spectrum using the inverse DFT. With this process, the components that correspond to more important modulation spectral bands in the feature sequence can be processed separately. For the Aurora-2 clean-condition training task, the new proposed sub-band spectral MN, MVN and HEQ provide relative error rate reductions of 18.66% and 23.58% over the conventional temporal MVN and HEQ, respectively.

一、簡介

雖然語音科技進步迅速，但自動語音辨識(automatic speech recognition, ASR)[1]仍是一門值得繼續研究開發的課題。目前多數的語音辨識系統若在不受干擾的安靜環境下，一般而言皆能得到相當滿意的辨識效果，然而若將其應用於實際的生活環境中，辨識效能便會有所衰減，主要是實際生活環境中有許多的變異性(variation)影響辨識效能，其中影響語音辨識的變異性有訓練環境與測試環境之間的環境不匹配(environmental mismatch)、語者變異性(speaker variation)及發音的變異性(pronunciation variation)等因素，這些因素都會明顯影響語音辨識系統的效能。因此在近幾十年來，持續不斷有許多學者研究努力改善上述幾類的語音變異性，進而使語音辨識系統能更有效地運用於真實的生活環境中。

針對環境不匹配所發展的許多強健性方法，大致上包含了特徵補償與模型補償兩大類型，而特徵補償方法中其中有一類別的方向是針對語音辨識所用的特徵參數之統計量作正規化處理，這些處理通常是作在特徵之時間序列域(temporal domain)上，例如倒頻譜平均值正規化法(cepstral mean normalization, CMN)[2]、倒頻譜平均值與變異數正規化法(cepstral mean and variance normalization, CMVN)[3]與統計圖等化法(histogram equalization, HEQ)[4]等。

以上各種方法主要是執行在語音特徵的時間序列域上，但在其效能的分析上，我們通常會去探討雜訊及通道效應對於原始特徵之調變頻譜的失真，及這些方法對於此失真的改善程度，因此近年來，開始有學者提出直接於特徵之調變頻譜域上使用特徵統計正規化法，如調變頻譜統計圖等化法(spectrum histogram equalization, SHE)[5]，此方法是針對調變頻譜的強度頻譜之機率分佈(probability distribution)作正規化處理，驗證了直接針對調變頻譜的強度成份作機率分佈的正規化的確帶來了明顯的特徵強健性效果。但是以上各種技術，皆是直接或間接將語音特徵序列的全調變頻帶資訊作整體的處理，並未對各不同的頻帶有不同的考慮。然而，根據許多的研究[6][7]證實，對語音辨識而言，不同頻率的調變頻譜成份具有不相等的重要性；在文獻[8]中更明確地提到，調變頻譜的偏低頻率成份資訊對於語音辨識有較大的助益，其中又以 1~16 Hz 之調變頻帶範圍的成份最爲重要。藉由以上之各觀點，在本論文中，我們提出了基於強度頻譜之**分頻段**調變頻譜統計正規化法，一方面希望如 SHE 法，直接對於語音特徵序列之調變頻譜作正規化處理，另一方面，則是希望在新方法中能異於過去之全調變頻帶之資訊一併處理的方式，將調變頻帶作一系列的頻段切割，在每個子頻段中加以正規化其調變頻譜，進而更有效地凸顯正規化的效能；在後面章節之一系列的實驗中，我們將呈現所提出之新方法確實可以更有效地提昇語音特徵在雜訊環境的強健性，達到我們以上所提的目的。

本論文其他章節概要如下：在第二章中，我們介紹本論文所提出之分頻段的統計正規化法其背景、原理及其相關的步驟說明。第三章將呈現並討論一系列分頻段調變頻譜統計正規化法的實驗結果，並與其他時間序列域上的強健性技術結合，對此類結合方式

的辨識實驗加以探討與分析，以驗證此類結合方式是否具有良好的加成性。而在第四章裡，則爲一簡要的結論與未來展望。

二、基於強度頻譜之分頻段調變頻譜統計正規化法

在這一章中，我們將對所新提出的分頻段調變頻譜統計正規化法之背景與步驟作詳細的說明，並且將以一段受雜訊干擾的語句爲例，驗證這些新方法在降低雜訊干擾的效能，及與其他相類似方法的初步比較。

(一) 分頻段調變頻譜統計正規化法

在本論文所提的新方法中，我們嘗試將調變頻譜中的強度頻譜(magnitude spectrum)切割成許多子頻段，再分別對各自子頻段的統計值作正規化處理；我們所用的正規化演算法，包括了除了文獻[5]之 SHE 技術所用的統計圖等化法(HEQ)外，也額外使用了較簡易執行的平均值正規化法(MN)與平均值與變異數正規化法(MVN)，以期它們相較於傳統全頻帶式的正規化法而言，能帶來更明顯的效能，或是能有效減低執行的複雜度。我們所提的分頻段調變頻譜正規化法的詳細步驟分列於下：

1. 假設一段語音之梅爾倒頻譜特徵參數序列以下式(2-1)表示：

$$\left\{x^{(m)}[n];1\le n\le N\right\},\quad 1\le m\le M, \qquad \text{式(2-1)}$$

其中 M 爲一語音特徵向量中特徵個數， N 表示爲此單一語句的音框總數。每個特徵序列 $\left\{x^{(m)}[n]\right\}$ 經正規化處理後，以 $\left\{\tilde{x}^{(m)}[n]\right\}$ 表示，我們希望新的特徵序列 $\left\{\tilde{x}^{(m)}[n]\right\}$ 相對於原始特徵序列而言，更具有強健性，使辨識效果有明顯地提升。在之後的敘述，爲了精簡符號的標示，我們省略了上標"(m)"符號。

2. 將特徵序列 $\left\{x[n];1\le n\le N\right\}$ 經 N 點離散傅立葉轉換(discrete Fourier transform, DFT)後得到其調變頻譜 $\left\{X[k]\right\}$ ，如下式。

$$X[k]=\sum_{n=0}^{N-1}x[n]e^{-j\frac{2\pi nk}{K}},\quad 0\le k\le\left\lceil\frac{N}{2}\right\rceil \qquad \text{式(2-2)}$$

假設 $\{x[n]\}$ 的音框取樣頻率(frame rate)爲 F_s Hz，則在其調變頻譜域上 $\left\{X[k]\right\}$ 的頻率範圍爲 $\left[0,\frac{F_s}{2}\right]$ ；而由於 $X[k]$ 爲一複數，我們以極座標(polar form)表示 $X[k]$ 如下式：

$$X[k]=A[k]e^{j\theta_k} \qquad \text{式(2-3)}$$

其中 $A[k]$ 是 $X[k]$ 的強度成份， $\theta[k]$ 是 $X[k]$ 的相位成份，接下來我們只針對強度成份 $\left\{A[k]\right\}$ 作調整，而保留相位成份 $\left\{\theta[k]\right\}$ 不變。

3. 將上一步驟調變頻譜的強度成分 $\left\{A[k];0\le k\le\left\lceil\frac{N}{2}\right\rceil\right\}$ 以不等切(non-uniform)且倍頻(octave)的方式，切割成 L 個頻段，每個頻段的範圍如下式(2-4)所示：

$$\begin{cases}\left[0,\dfrac{1}{2^{L-1}}\left(\dfrac{F_s}{2}\right)\right], & \text{if } \ell=1.\\ \left[\dfrac{2^{\ell-2}}{2^{L-1}}\left(\dfrac{F_s}{2}\right),\dfrac{2^{\ell-1}}{2^{L-1}}\left(\dfrac{F_s}{2}\right)\right], & \text{if } \ell=2,3,\ldots,L.\end{cases} \qquad \text{式(2-4)}$$

由上式可以得知，調變頻譜低頻帶的部分被切割成較多個頻段，且每個頻段的長度較

短，相對地，高頻的部分被切割成較少的頻段，且每個頻段的長度較長。在將$\{A[k]\}$作上述的頻段切割後，我們以$\left\{A_\ell\left[k'\right]\right\}$表示其中的第 L 個頻段。此對於頻段不等切的原因，在於我們之前所提，低調變頻帶對於語音辨識較爲重要，理應分較多的頻段來個別處理，而高調變頻帶相對而言較不重要，所以可將較大的頻段範圍一併處理。

4. 我們將上一步驟所得之不同頻段的強度頻譜$\left\{A_\ell\left[k'\right]\right\}$作統計正規化處理。我們使用的正規化法分別爲：平均值正規化法(MN)、平均值與變異數正規化法(MVN)與統計圖等化法(HEQ)，處理後的特徵即以$\left\{\tilde{A}_\ell\left[k'\right]\right\}$表示。詳細地說，平均值正規化法(MN)在此的計算方式以下式(2-5)表示：

$$\tilde{A}_\ell\left[k'\right]=A_\ell\left[k'\right]-\mu_{\ell,s}+\mu_{\ell,a},\qquad \text{式(2-5)}$$

其中，$\mu_{\ell,s}$爲單一(single)語句之分頻段強度頻譜的平均值，$\mu_{\ell,a}$爲全部(all)訓練語句之分頻段強度頻譜的平均值。

平均值與變異數正規化法(MVN)在此的計算方式以式(2-6)表示：

$$\tilde{A}_\ell\left[k'\right]=\left(\frac{A_\ell\left[k'\right]-\mu_{\ell,s}}{\sigma_{\ell,s}}\right)\cdot\sigma_{\ell,a}+\mu_{\ell,a}\qquad \text{式(2-6)}$$

其中，$\mu_{\ell,s}$爲單一語句之分頻段強度頻譜的平均值，$\sigma_{\ell,s}$爲單一語句之分頻段強度頻譜的標準差，$\mu_{\ell,a}$爲全部訓練語句之分頻段強度頻譜的平均值，$\sigma_{\ell,a}$爲全部訓練語句之分頻段強度頻譜的標準差。

統計圖等化法(HEQ)在此的計算方式以式(2-7)表示：

$$\tilde{A}_\ell\left[k'\right]=F_{\ell,a}^{-1}\left(F_{\ell,s}\left(A_\ell\left[k'\right]\right)\right)\qquad \text{式(2-7)}$$

其中$F_{\ell,s}(\bullet)$爲單一語句之分頻段強度頻譜的機率分佈，$F_{\ell,a}(\bullet)$爲全部訓練語句之分頻段強度頻譜的機率分佈。

5. 在處理完每一頻段之後，我們將各頻段的強度頻譜$\left\{\tilde{A}_\ell\left[k'\right]\right\}$照其頻率大小順序重新串接起來，得到新的全頻段強度頻譜$\left\{\tilde{A}[k];0\le k\le\left\lceil\frac{N}{2}\right\rceil\right\}$，此即爲統計正規化法處理後的調變頻譜之強度成份，接著將$\left\{\tilde{A}[k]\right\}$補回式(2-3)中的原本相位成分$\{\theta[k]\}$，再經逆轉換離散傅立葉轉換(inverse discrete Fourier transform, IDFT)所得新的特徵$\tilde{x}[n]$，如下式(2-8)表示：

$$\tilde{x}[n]=\frac{1}{N}\sum_{k=0}^{N-1}\left(\tilde{A}[k]e^{j\theta[k]}\right)e^{j\frac{2\pi nk}{N}},\qquad 0\le n\le N-1.\qquad \text{式(2-8)}$$

由於特徵序列經傅立葉轉換後，具有左右對稱的特性，即$\tilde{A}[k]=\tilde{A}[N-k]$與$\theta[k]=-\theta[N-k]$，因此我們可藉此推得式(2-8)所需用到的$\left\{\tilde{A}[k]\right\}$與$\{\theta[k]\}$在$\left\lceil\frac{N}{2}\right\rceil<k\le N-1$的每一項。

在步驟 2 中，若我們未對調變頻譜的語音特徵作分頻段處理，即分段數$L=1$，接著在步驟 3 作統計圖等化法(HEQ)的正規化運算，這樣的運算方式相當於[5]中的調變頻譜統計圖等化法(SHE)；爲了之後討論方便起見，我們將上述式(2-5)、式(2-6)、式(2-7)的正規化方法處理，分別命名爲：分頻段調變頻譜平均值正規化法(sub-band spectral mean normalization, SB-SMN)、分頻段調變頻譜平均值與變異數正規化法(sub-band

spectral mean and variance normalization, SB-SMVN)與分頻段調變頻譜統計圖等化法(sub-band spectral histogram equalization, SB-SHE)，而文獻[5]中所用的全頻帶(full-band)之 SHE 技術，我們則以 FB-SHE 來表示。以下將針對這些分頻段正規化法的特點加以討論：

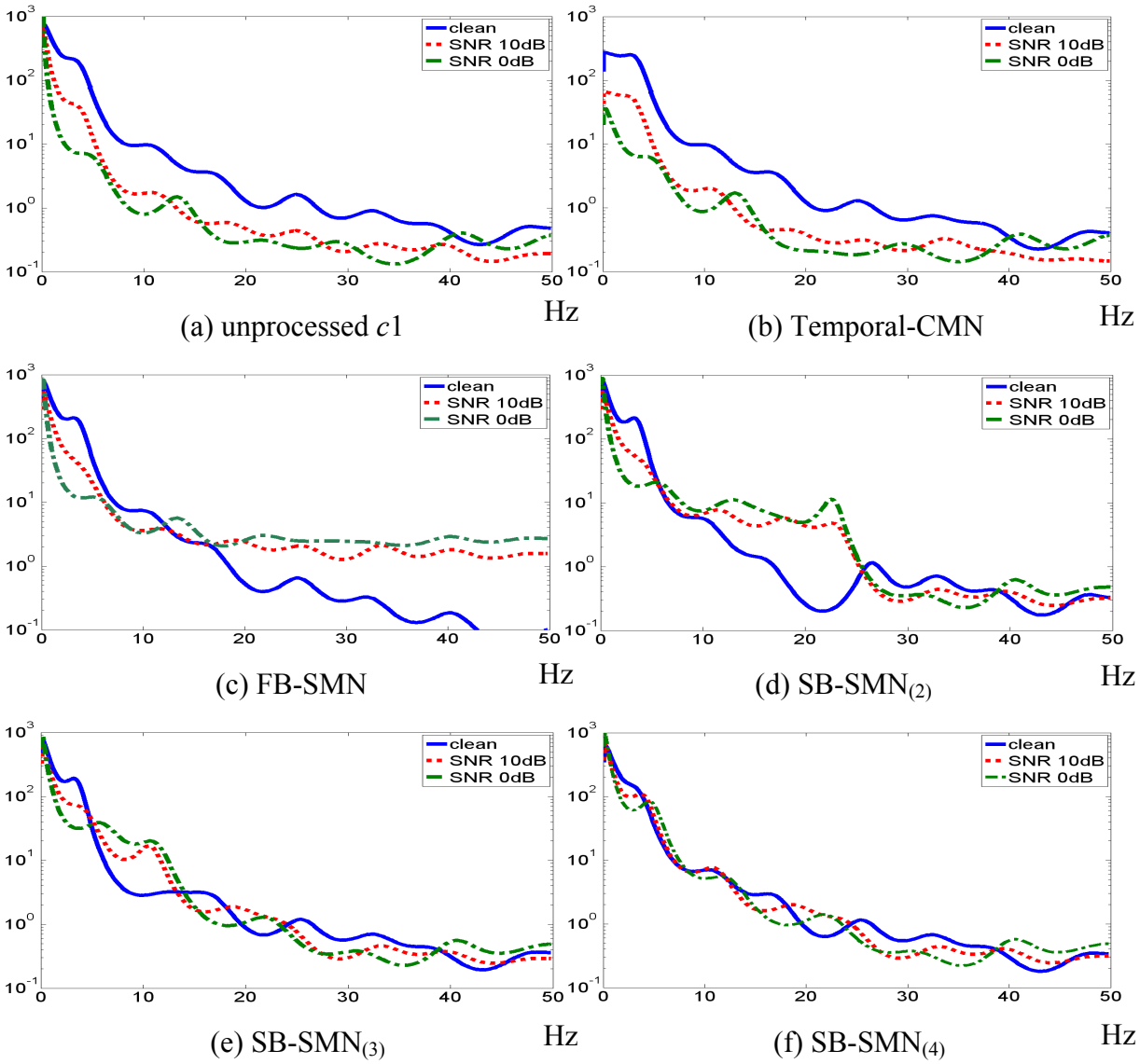
(1) 經由 SB-SMN 與 SB-SMVN 的方法處理之後，所得的調變頻譜強度之部份數值可能爲負值，此明顯違反頻譜強度必然非負的條件，因此當負值的情形出現時，我們將其值重設爲 0。

(2) 在 SB-SMN 與 SB-SMVN 方法中，不同頻段具有各自的目標平均值或目標變異數，同樣地，在 SB-SHE 中，不同頻段使用不同的目標機率分佈作正規化運算。這樣的作法，可以保留不同頻段的頻譜強度之間差異性。

(3) 在這些分頻段調變頻譜正規化法中，全頻段的長度等於各分頻段之長度的和，所以增加分頻段的數目並不會明顯增加運算上的複雜度。然而子頻段的數目不能過多，否則在低頻的子頻段的 $A[k]$ 項數將過少甚至爲零，如此明顯會影響單一頻段所求取之統計值（如平均值、變異數與機率分佈等）的精確性。舉例說明，假設單一語音特徵序列的總點數爲 N 點時，由於此（實數）特徵序列經 N 點傅立葉轉換後，其頻譜的強度成份具有左右對稱的特性，因此實際使用的頻譜點數爲總數的一半，即爲 $\left\lfloor \frac{N}{2} \right\rfloor$ 點，如果我們以不等切的方式切割整個頻帶，所得的頻段不能無限制地增多；例如當我們切 L 個子頻段時，所得的每個子頻段點數由多到少分別爲：$\left\lfloor \frac{N}{4} \right\rfloor, \left\lfloor \frac{N}{8} \right\rfloor, \ldots, \left\lfloor \frac{N}{2^{(L+1)}} \right\rfloor$，所以爲了滿足最少的那一個子頻段的點數不爲零，即每個頻段的資料量至少有一點，我們須滿足 $N \geq 2^{(L+1)}$ 的條件，由此推知，若 $N = 60$，最多只能切 5 個頻段，而若 $N = 30$，則最多只能切 4 個子頻段，若進一步要求若要求每個子頻段點數不能太少，則子頻段數目限制將會更嚴格。

(二) 分頻段調變頻譜正規化法其初步效能的討論：

在這裡，我們將探討本章所提出的三種分頻段調變頻譜正規化法在特徵序列之功率頻譜密度(power spectral density, PSD)降低失真的效果，同時，把這些方法呈現的結果，和第二章所介紹之特徵時間序列域(temporal domain)之正規化技術：倒頻譜平均值正規化法(CMN)、倒頻譜平均值與變異數正規化法(CMVN)、統計圖等化法(HEQ)分別作比較。我們利用了 AURORA-2 資料庫[9]裡 MIP_28826Z4A 語音檔，加入不同訊雜比(SNR)的人聲(babble)雜訊，再經各種正規化法加以處理，最後求取其功率頻譜密度。首先，圖一爲一系列之平均值正規化法（MN）作用於第一維倒頻譜特徵(the first cepstral coefficient, c_1)序列所得之功率頻譜密度圖。在圖一中，藉由圖(a)我們發現，雜訊的存在使乾淨語音與雜訊語音產生明顯的 PSD 失真，而圖(b)中所用 CMN 法，即時域型 MN 法(temporal CMN)，可稍微降低此失真，而頻域型的 MN 法中，由圖(c)至圖(f)發現,將全頻帶逐漸細分至 2 到 4 個子頻段（分別以 SB-SMN$_{(2)}$, SB-SMN$_{(3)}$與 SB-SMN$_{(4)}$表示，下標括號中的數字表示分頻段的個數），此 PSD 失真逐漸降低，其中以圖(f)經 SB-SMN$_{(4)}$ 處理後，PSD 的失真程度最小。由此說明分頻段調變頻譜平均值正規化法對於降低因雜訊所造成的 PSD 失真有明顯的幫助。



(a) unprocessed $c1$ (b) Temporal-CMN

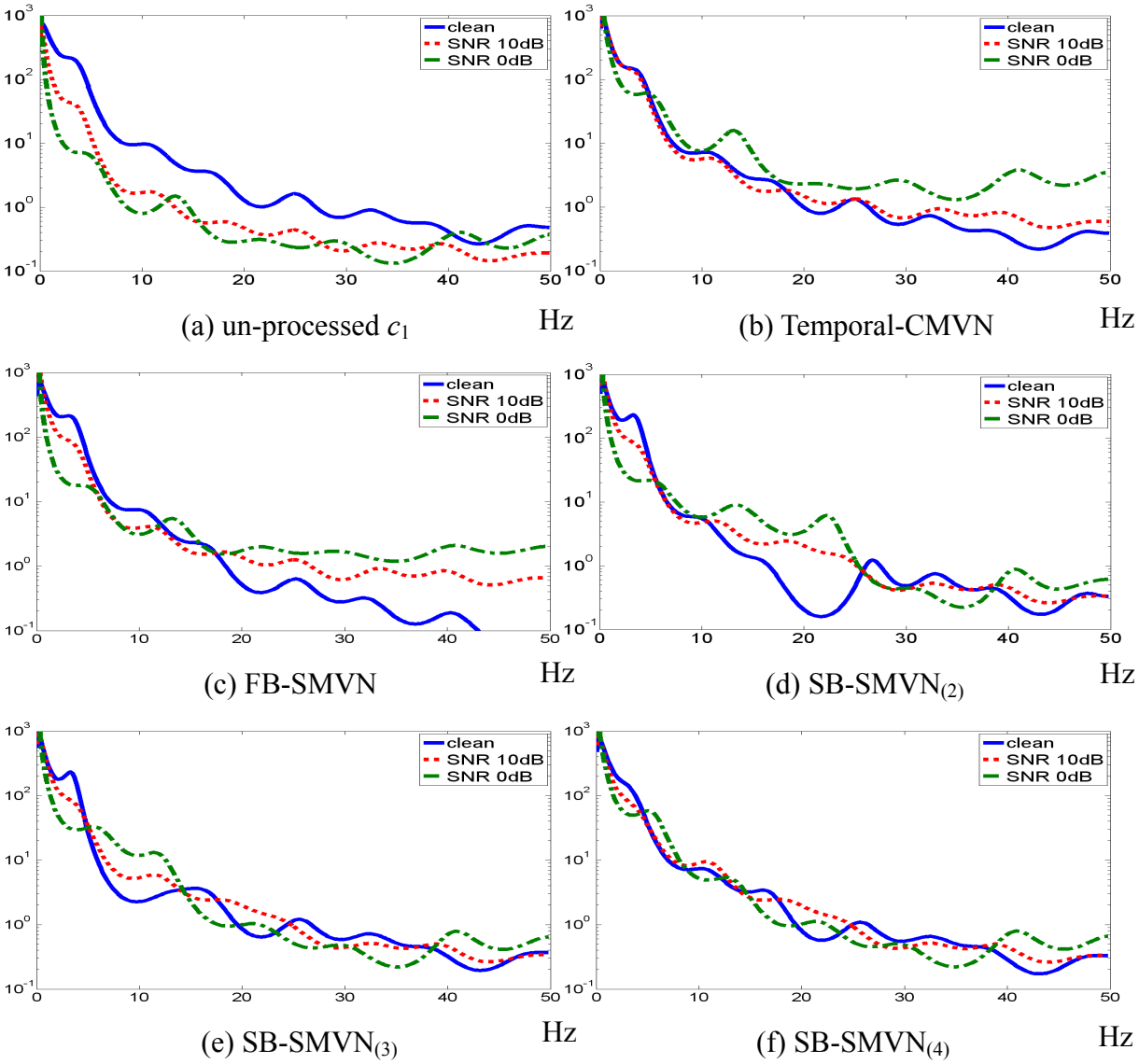
(c) FB-SMN (d) SB-SMN$_{(2)}$

(e) SB-SMN$_{(3)}$ (f) SB-SMN$_{(4)}$

圖一 平均值正規化法作用於不同訊雜比下語音之原始 $c1$ 特徵序列，其調變頻譜曲線圖：(a)原始 $c1$ 特徵序列，(b)時域型 MN 法—CMN，(c)頻域型之全頻帶 MN 法—FB-SMN，(d)頻域型之分頻段 MN 法—SB-SMN$_{(2)}$，(e)頻域型之分頻段 MN 法—SB-SMN$_{(3)}$，(f)頻域型之分頻段 MN 法—SB-SMN$_{(4)}$

接著，圖二爲一系列之平均值與變異數正規化法（MVN）作用於第一維倒頻譜特徵(the first cepstral coefficient, c_1)序列所得之功率頻譜密度圖。在圖二中，藉由圖(b)我們發現，傳統的 CMVN 法，即時域型 MVN 法(temporal CMVN)，相較於圖一的圖(b)之 CMN 而言，降低 PSD 失真的效應更好，意味了額外處理特徵的變異數確實是有幫助的。而在各種頻域型的 MVN 法中，由圖(c)至圖(f)發現，類似 MN 法的效果，當我們將全頻帶逐漸細分至 2 到 4 個子頻段（分別以 SB-SMVN$_{(2)}$, SB-SMVN$_{(3)}$與 SB-SMVN$_{(4)}$表示，下標括號中的數字表示分頻段的個數），PSD 失真也逐漸降低，其中以圖(f)經 SB-SMVN$_{(4)}$ 處理後，對於 PSD 的失真的降低效果最好。由此亦說明了分頻段調變頻譜平均值與變異數正規化法對於降低因雜訊所造成的 PSD 失真也有明顯幫助，同時將圖二與圖一比

較後，可以明顯看出 MVN 法在降低 PSD 失真的效能上優於 MN 法，此吻合我們一般對這兩類方法之效能的認知。



(a) un-processed c_1 (b) Temporal-CMVN

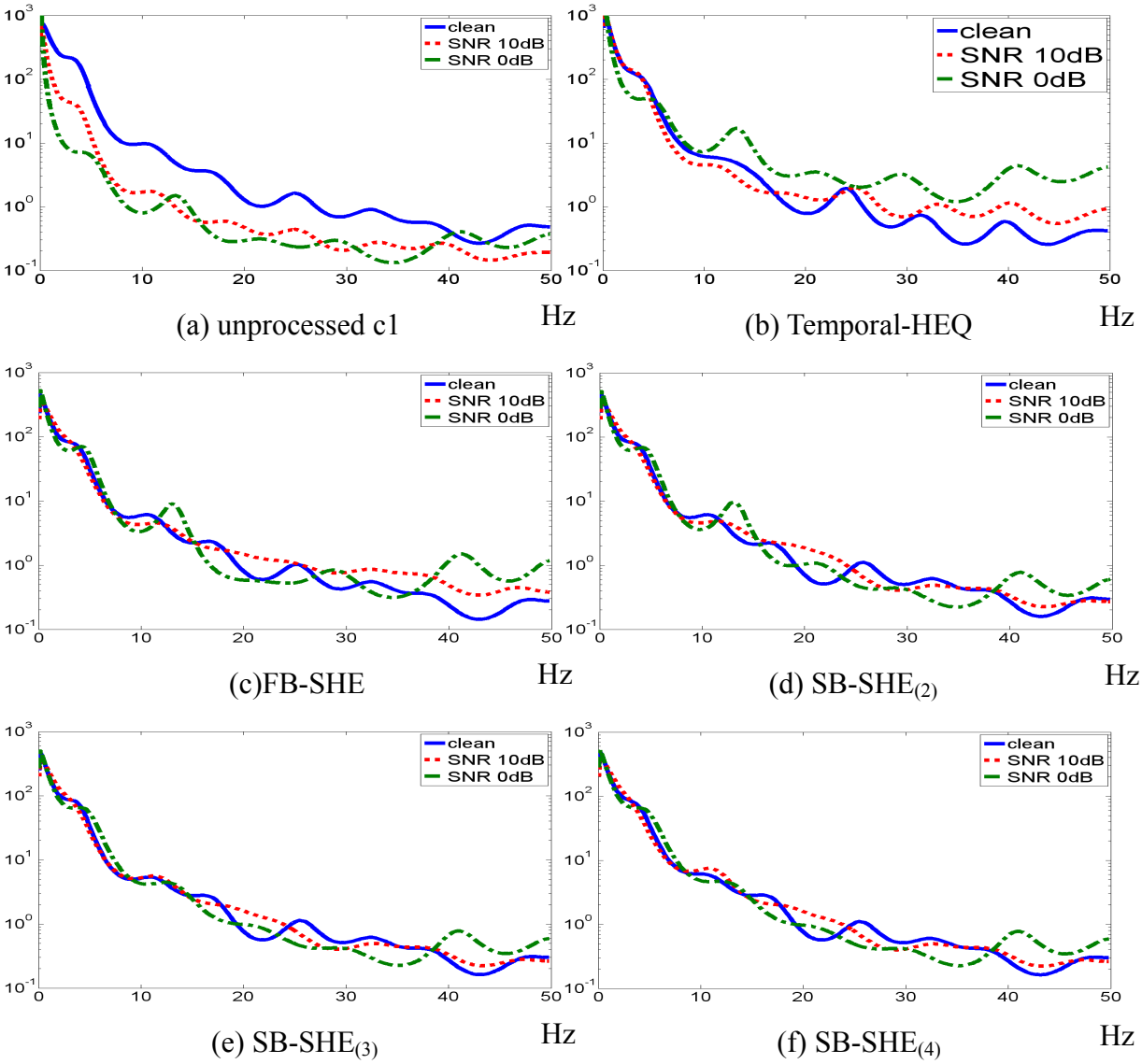
(c) FB-SMVN (d) SB-SMVN$_{(2)}$

(e) SB-SMVN$_{(3)}$ (f) SB-SMVN$_{(4)}$

圖二 平均值與變異數正規化法作用於不同訊雜比下語音之原始$c1$特徵序列，其調變頻譜曲線圖：(a)原始$c1$特徵序列，(b)時域型 MVN 法—CMVN，(c)頻域型之全頻帶 MVN 法—FB-SMVN， (d)頻域型之分頻段 MVN 法—SB-SMVN$_{(2)}$，(e)頻域型之分頻段 MVN 法—SB-SMVN$_{(3)}$，(f)頻域型之分頻段 MVN 法—SB-SMVN$_{(4)}$

最後，圖三爲一系列之統計圖等化法（HEQ）作用於第一維倒頻譜特徵(the first cepstral coefficient, $c1$)序列所得之功率頻譜密度圖。我們將圖三與圖一和圖二比較，可明顯看出正規化整個機率分佈的 HEQ 法，明顯在降低 PSD 的失真上優於只正規化平均值的 MN 法與正規化平均值與變異數的 MVN 法（無論是時域型或頻域型的皆是如此），此外，我們若比較三種全頻式的方法(圖一(c)的 FB-SMN, 圖二(c)的 FB-SMVN 與圖三(c)FB-SHE)，可發現 FB-SHE 相對於 FB-SMN 與 FB-SMVN 而言，從低頻到高頻的 PSD 失真都有明顯降低，而不是像 FB-SMN 與 FB-SMVN 相對只有減少低頻成分的 PSD 失

真，此現象也間接驗證了文獻[5]所提之 FB-SHE 的良好效能。然而在我們所提出的各種分頻段 SHE(SB-SHE)法中，明顯看出它們皆比 FB-SHE 在減低 PSD 失真的效能來得好，由圖(c)至圖(f)發現，類似之前 MN 與 MVN 法的效果，當我們將頻段從全頻段逐漸細分至 2 到 4 個頻段（分別以 SB-HEQ$_{(2)}$, SB- HEQ $_{(3)}$與 SB- HEQ $_{(4)}$表示，下標括號中的數字表示分頻段的個數），PSD 失真逐漸降低，其中以圖(f)經 SB-HEQ$_{(4)}$處理後，對於 PSD 的失真的降低效果最好。由此明顯說明了分頻段調變頻譜統計圖等化法足以有效降低因雜訊所造成的 PSD 失真，在下一章的辨識實驗中，我們將更明顯地看出這些新方法對於提昇語音辨識精確度的效能。



(a) unprocessed c1　　(b) Temporal-HEQ

(c)FB-SHE　　(d) SB-SHE$_{(2)}$

(e) SB-SHE$_{(3)}$　　(f) SB-SHE$_{(4)}$

圖三 統計圖等化法作用於不同訊雜比下語音之原始$c1$特徵序列，其調變頻譜曲線圖：(a)原始$c1$特徵序列，(b)時域型 HEQ 法—Temporal HEQ，(c)頻域型之全頻帶 HEQ 法—FB-HEQ，(d)頻域型之分頻段 HEQ 法—SB-HEQ$_{(2)}$，(e)頻域型之分頻段 HEQ 法—SB-HEQ$_{(3)}$，(f)頻域型之分頻段 HEQ 法—SB-HEQ$_{(4)}$

三、分頻段調變頻譜統計正規化法之實驗結果及分析討論

(一) 語音資料庫簡介

本論文使用的語音資料庫爲歐洲電信標準協會(European Telecommunication Standard Institute, ETSI)所發行的 Aurora-2 語音資料庫[9]，它是一套以人工方式錄製的連續英文數字字串，語者由美國成年男女所組成，加上八種來源不同的雜訊，分別爲：地下鐵、人聲、汽車、展覽館、餐廳、街道、機場、火車站等，並以不同程度的訊雜比(signal-to-noise ratio, SNR)加入雜訊，分別爲：clean、20 dB、15 dB、10 dB、5 dB、0 dB 與-5 dB；其通道效應分別爲 G.712 與 MIRS，其爲國際電信聯盟(International Telecommunication Union, ITU)[10]所訂立的兩個通道標準。

(二) 語音特徵參數設定及聲學模型

本論文之相關語音辨識實驗所使用特徵參數爲梅爾倒頻譜係數(MFCC)，附加上其一階差量與二階差量，其詳細語音特徵參數設定分別在表一表示。

取樣頻率	8 kHz
音框長度(frame size)	25 ms, 200 點
音框平移(frame shift)	10 ms, 80 點
預強調濾波器	$1-0.97z^{-1}$
視窗形式(window)	漢明窗(Hamming window)
傅立葉轉換點數(T)	256 點
濾波器組(filter bank)	梅爾刻度三角濾波器組，共 23 個三角濾波器
特徵向量(feature vector)	MFCC 13 維($c1$~$c12$, log-energy) + △MFCC 13 維+△△MFCC 13 維，共 39 維

表一 本論文所使用語音特徵參數設定

我們是以隱藏式馬可夫模型(hidden Markov model, HMM)[11]作爲聲學模型(acoustic models)的型式。包含11個數字模型(zero, one, two, …, nine 及oh)以及靜音(silence)模型，每個數字模型包含16個狀態，各狀態包含20個高斯密度混合。

(三) 語音辨識實驗結果

在這一節中，我們將各種調變頻譜正規化法之實驗結果綜合整理成表二，其中絕對錯誤降低率(absolute error rate reduction, AR)與相對錯誤降低率 1 (relative error rate reduction 1, RR_1)分別爲新辨識率與基礎實驗辨識率(baseline)比較下，所得到的絕對改善率與相對改善率，相對錯誤改善率 2 (relative error rate reduction 2, RR_2)，它是分頻段技術相較於全頻段技術而言所得到的相對錯誤改善率，其計算方式分別由式(3-1)、式(3-2)、式(3-3)所示：

$$AR(\%)=\left(\text{新辨識率}-\text{基礎實驗辨識率}\right)\times 100\% \qquad \text{式(3-1)}$$

$$RR_1(\%)=\left(\frac{\text{新辨識率}-\text{基礎實驗辨識率}}{100\%-\text{基礎實驗辨識率}}\right)\times 100\% \qquad \text{式(3-2)}$$

$$RR_2(\%)=\left(\frac{\text{分頻段法辨識率}-\text{全頻段法辨識率}}{100\%-\text{全頻段法辨識率}}\right)\times 100\% \qquad \text{式(3-3)}$$

由表二觀察中，我們可以得到以下幾點結果：

1. 我們所新提出之各種分頻段調變頻譜正規化法相較於基本實驗而言，皆能使辨識率明顯提升，從 RR_1 的數據看出，它們至少能有 19.00%的相對錯誤降低率；其中 SB-SHE 法的辨識效果比 SB-SMN 法及 SB-SMVN 法更優越，可能之原因如我們預期的，SB-SMN 法及 SB-SMVN 法只對一階動差或一階及二階動差作正規化，而 SB-SHE 法能同時對更高階的動差作正規化處理，使得 SB-SHE 法有較優異的表現。

2. 由於調變頻譜中低頻部分(1~16Hz)佔有較多重要的語音成份，所以我們著重於將低頻部分切割開來分別作正規化處理，從表二可以清楚發現，當低頻部份切割越細，能有效提升語音辨識效能，而三種分頻段調變頻譜補償技術皆以分割四個頻段的效果最爲優越。相對於全頻段式的方法而言，分頻段式的方法其相對錯誤改善率(RR_2)爲：$SB\text{-}SMN_{(4)}$ 的 8.31%，$SB\text{-}SMVN_{(4)}$的 32.64%，$SB\text{-}SHE_{(4)}$的 7.56%。

Method	Set A	Set B	Set C	average	AR	RR_1	RR_2
Baseline	71.98	67.79	78.28	71.56	—	—	—
FB-SMN	77.43	76.26	78.05	77.08	5.52	19.41	—
$SB\text{-}SMN_{(2)}$	77.87	77.26	78.36	77.72	6.16	21.66	2.79
$SB\text{-}SMN_{(3)}$	78.21	76.37	80.82	77.99	6.43	22.61	3.97
$SB\text{-}SMN_{(4)}$	79.12	77.26	82.20	**78.99**	**7.43**	**26.13**	**8.31**
FB-SMVN	79.03	81.19	78.29	79.75	8.19	28.80	—
$SB\text{-}SMVN_{(2)}$	80.06	81.97	79.28	80.67	9.11	32.03	4.54
$SB\text{-}SMVN_{(3)}$	80.84	82.59	80.89	81.55	9.99	35.13	8.89
$SB\text{-}SMVN_{(4)}$	85.94	87.06	85.79	**86.36**	**14.80**	**52.04**	**32.64**
FB-SHE	89.71	90.03	88.27	89.55	17.99	63.26	—
$SB\text{-}SHE_{(2)}$	89.76	90.09	88.40	89.62	18.06	63.50	0.67
$SB\text{-}SHE_{(3)}$	90.13	90.47	88.68	89.98	18.42	64.77	4.11
$SB\text{-}SHE_{(4)}$	90.59	90.69	89.13	**90.34**	**18.78**	**66.03**	**7.56**

表二 調變頻譜統計正規化法之實驗辨識率(%)綜合比較表

(四) 調變頻譜正規化法結合時域型特徵正規化法之實驗結果

在本節中，我們先將原始 MFCC 特徵經各式時域型特徵統計正規化法處理後，再作調變頻譜統計正規化法的處理。在以下各項將呈現並討論各式調變頻譜正規化結合時域型特徵正規化法之實驗結果。

1. 調變頻譜平均值正規化法結合時域型特徵統計正規化法之實驗結果

實驗結果討論：

(1) 表三中，調變頻譜平均值正規化法結合時域型特徵統計正規化法，與其中單一特徵正規化法比較，幾乎皆能有效提升語音辨識效能。舉例而言：$SB\text{-}SMN_{(4)}$結合 CMN 的辨識率爲 88.02%，比起 CMN 的辨識率 81.66%與 $SB\text{-}SMN_{(L=4)}$的辨識率 78.99%，都有相當明顯的改善。惟獨在 $SB\text{-}SMN_{(2)}$結合 CMN 情況下，無法進一步提升辨識率，這可能是該結合方式的分頻段之平均值無法有效逼近訓練語句之分頻段強度頻譜的平均

值，導致辨識率明顯下降。

(2) 從表三也可以清楚發現，當低頻部份切割越細，能有效提升語音辨識效能，而 SMN 法結合其他特徵正規化法皆以分割四個頻段的效果最爲優越。其中 SB-SMN$_{(4)}$結合 CMN 的 RR$_2$ 爲 30.02%，SB-SMN$_{(4)}$結合 CMVN 的 RR$_2$ 爲 15.95%，SB-SMN$_{(4)}$結合 MVA 的 RR$_2$ 爲 12.70%，SB-SMN$_{(4)}$結合 HEQ 的 RR$_2$ 爲 4.98%。

(3) SB-SMN$_{(L=4)}$分別與 CMVN、MVA 及 HEQ 結合，使辨識率幾乎達到 90.00%；而此代表我們用相對簡單的一階統計正規化法(SB-SMN)結合 CMVN、MVA 及 HEQ，即可達到十分突出的效果。

Method		Set A	Set B	Set C	average	AR	RR$_1$	RR$_2$
Baseline		71.98	67.79	78.28	71.56	—	—	—
CMN		80.69	83.41	80.09	81.66	—	—	—
CMN	FB-SMN	82.34	84.06	81.61	82.88	1.22	6.65	—
	SB-SMN$_{(2)}$	80.89	82.22	80.24	81.29	-0.37	-2.02	-9.29
	SB-SMN$_{(3)}$	83.67	84.63	82.69	83.86	2.20	12.00	5.72
	SB-SMN$_{(4)}$	88.09	88.64	86.63	**88.02**	**6.36**	**34.68**	**30.02**
CMVN		83.55	83.75	81.57	83.23	—	—	—
CMVN	FB-SMN	87.81	88.18	86.27	87.65	4.42	26.36	—
	SB-SMN$_{(2)}$	89.08	89.39	87.39	88.87	5.64	33.63	9.88
	SB-SMN$_{(3)}$	89.63	89.97	88.37	89.51	6.28	37.45	15.06
	SB-SMN$_{(4)}$	89.86	90.09	88.20	**89.62**	**6.39**	**38.10**	**15.95**
MVA		86.69	86.89	84.98	86.43	—	—	—
MVA	FB-SMN	88.89	89.19	87.54	88.74	2.31	17.02	—
	SB-SMN$_{(2)}$	89.87	90.17	88.77	89.77	3.34	24.61	9.15
	SB-SMN$_{(3)}$	90.08	90.51	88.94	90.02	3.59	26.46	11.37
	SB-SMN$_{(=4)}$	90.36	90.59	88.94	**90.17**	**3.74**	**27.56**	**12.70**
HEQ		86.90	87.73	87.56	87.36	—	—	—
HEQ	FB-SMN	89.10	89.70	89.20	89.36	2.00	15.82	—
	SB-SMN$_{(L=2)}$	89.20	89.81	89.30	89.46	2.10	16.61	0.94
	SB-SMN$_{(L=3)}$	89.15	89.89	89.35	89.48	2.12	16.77	1.13
	SB-SMN$_{(L=4)}$	89.54	90.28	89.82	**89.89**	**2.53**	**20.02**	**4.98**

表三 SMN 法結合時域型特徵正規化法之實驗綜合比較表

2. 調變頻譜平均值與變異數正規化法結合時域型特徵正規化法之實驗結果

實驗結果討論：

(1) 表四中，調變頻譜平均值與變異數正規化法結合時域上特徵正規化法與單一作特徵正規化法比較，皆能有效提升語音辨識效能。舉而言之：SB-SMVN$_{(4)}$結合 CMVN 的辨識率爲 89.87%，比起 CMVN 的辨識率 83.23%與 SB-SMVN$_{(4)}$的辨識率 86.36%，都有相當明顯的改善。

(2) 從表四也可以清楚發現，當低頻部份切割越細，更能有效提升語音辨識效能，而 SMVN 法結合其他特徵正規化法皆以分割四個頻段的效果最爲優越。其中 SB-SMVN$_{(4)}$ 結合 CMN 的 RR_2 爲 20.73%，SB-SMVN$_{(4)}$結合 CMVN 的 RR_2 爲 21.17%，SB-SMVN$_{(4)}$ 結合 MVA 的 RR_2 爲 16.99%，SB-SMVN$_{(4)}$結合 HEQ 的 RR_2 爲 9.80%。

Method		Set A	Set B	Set C	average	AR	RR_1	RR_2
Baseline		71.98	67.79	78.28	71.56	—	—	—
CMN		80.69	83.41	80.09	81.66	—	—	—
CMN	FB-SMVN	87.38	87.73	85.61	87.17	5.51	30.04	—
	SB-SMVN$_{(2)}$	88.60	88.89	86.83	88.36	6.70	36.53	9.28
	SB-SMVN$_{(3)}$	89.77	89.90	88.17	89.50	7.84	42.75	18.16
	SB-SMVN$_{(4)}$	90.01	90.35	88.43	**89.83**	**8.17**	**44.55**	**20.73**
CMVN		83.55	83.75	81.57	83.23	—	—	—
CMVN	FB-SMVN	87.33	87.80	85.47	87.15	3.92	23.38	—
	SB-SMVN$_{(2)}$	88.54	88.84	86.80	88.31	5.08	30.29	9.03
	SB-SMVN$_{(3)}$	89.72	89.90	88.03	89.45	6.22	37.09	17.90
	SB-SMVN$_{(4)}$	90.11	90.37	88.42	**89.87**	**6.64**	**39.59**	**21.17**
MVA		86.69	86.89	84.98	86.43	—	—	—
MVA	FB-SMVN	88.18	88.60	86.39	87.99	1.56	11.50	—
	SB-SMVN$_{(2)}$	89.27	89.49	87.58	89.02	2.59	19.09	8.58
	SB-SMVN$_{(3)}$	89.69	89.95	88.36	89.52	3.09	22.77	12.74
	SB-SMVN$_{(4)}$	90.27	90.45	88.71	**90.03**	**3.60**	**26.53**	**16.99**
HEQ		86.90	87.73	87.56	87.36	—	—	—
HEQ	FB-SMVN	89.05	89.55	89.26	89.29	1.93	15.27	—
	SB-SMVN$_{(2)}$	89.33	89.81	89.40	89.54	2.18	17.25	2.33
	SB-SMVN$_{(3)}$	89.95	90.35	89.91	90.11	2.75	21.76	7.66
	SB-SMVN$_{(4)}$	90.19	90.59	90.17	**90.34**	**2.98**	**23.58**	**9.80**

表四 SMVN 法結合時域型特徵正規化法之實驗綜合比較表

3. 調變頻譜統計圖等化法結合時域型特徵正規化法之實驗結果

實驗結果討論：

(1) 類似表三、表四，從表五看出，調變頻譜統計圖等化法結合時域型特徵正規化法皆優於個別特徵正規化法。舉例而言：SB-SHE$_{(L=4)}$結合 CMVN 的辨識率爲 90.18%，比起 CMVN 的辨識率 83.23%有相當明顯的改善。但 SB-SHE 結合 CMVN 之實驗結果與 SB-SHE 作在 MFCC 特徵之實驗結果比較，在某些結合方式下，辨識結果會有些微的下降，這可能是實驗的誤差範圍，或是過度正規化的不良效應。

(2) 從表五也可以清楚發現，當切割的子頻段數目越多，越能有效提升語音辨識效能，而 SHE 法結合其他特徵正規化法皆以分割四個頻段的效果最爲優越。其中 SB-SHE$_{(4)}$ 結合 CMN 的 RR_2 爲 6.45%，SB-SHE$_{(4)}$結合 CMVN 的 RR_2 爲 7.97%，SB-SHE$_{(4)}$結合

MVA 的 RR_2 爲 2.60%，SB-$SHE_{(4)}$結合 HEQ 的 RR_2 爲 6.19%。

Method		Set A	Set B	Set C	average	AR	RR_1	RR_2
Baseline		71.98	67.79	78.28	71.56	—	—	—
CMN		80.69	83.41	80.09	81.66	—	—	—
CMN	FB-SHE	89.45	90.08	88.24	89.46	7.80	42.53	—
	SB-$SHE_{(2)}$	89.44	89.97	88.24	89.41	7.75	42.26	-0.47
	SB-$SHE_{(3)}$	89.90	90.29	88.68	89.81	8.15	44.44	3.32
	SB-$SHE_{(4)}$	90.20	90.63	89.06	**90.14**	**8.48**	**46.24**	**6.45**
CMVN		83.55	83.75	81.57	83.23	—	—	—
CMVN	FB-SHE	89.42	89.87	88.07	89.33	6.10	36.37	—
	SB-$SHE_{(2)}$	89.45	89.96	88.23	89.41	6.18	36.85	0.75
	SB-$SHE_{(3)}$	89.74	90.22	88.61	89.70	6.47	38.58	3.47
	SB-$SHE_{(4)}$	90.23	90.67	89.10	**90.18**	**6.95**	**41.44**	**7.97**
MVA		86.69	86.89	84.98	86.43	—	—	—
MVA	FB-SHE	89.97	90.50	88.98	89.99	3.56	26.23	—
	SB-$SHE_{(2)}$	89.98	90.49	88.98	89.98	3.55	26.16	-0.10
	SB-$SHE_{(3)}$	90.25	90.65	89.30	90.22	3.79	27.92	2.30
	SB-$SHE_{(4)}$	90.25	90.76	89.22	**90.25**	**3.82**	**28.15**	**2.60**
HEQ		86.90	87.73	87.56	87.36	—	—	—
HEQ	FB-SHE	89.24	90.09	89.61	89.66	2.30	18.20	—
	SB-$SHE_{(2)}$	89.22	90.08	89.44	89.61	2.25	17.80	-0.48
	SB-$SHE_{(3)}$	89.48	90.30	89.82	89.87	2.51	19.86	2.03
	SB-$SHE_{(4)}$	89.91	90.75	90.17	**90.30**	**2.94**	**23.26**	**6.19**

表五 SHE 法結合時域型特徵正規化法之實驗綜合比較表

四、結論與未來展望

在本論文中，我們提出了一系列分頻段調變頻譜統計正規化的演算法，以不等切的方式切割調變頻譜，再分別針對每個頻段的調變頻譜強度作統計正規化，分析其對語音特徵在雜訊環境下提昇強健性的效果。由實驗結果發現，相對於傳統不切割頻段的方式而言，這些新方法都有明顯的改進效果，我們也發現由於調變頻譜中偏低頻(約 1~16 Hz)的語音成份包含了大多數語音辨識所需的資訊，若我們將此低頻部份切割地越細，進而個別正規化處理，越有效提升語音辨識效能，而在我們所提的各種分頻段調變頻譜正規化法中，皆以切割四個頻段所得的辨識效果最爲優越。另外，我們也將各種分頻段調變頻譜正規化法分別與傳統時間序列域上之特徵統計正規化法作結合；由辨識實驗結果發現，二者組合其辨識精確率皆比使用單一強健性技術所得到的辨識率更好。由此可看出，我們所提的分頻段式之新方法，不僅能有效改善原先全頻段式的方法，更與其他語

音強健性技術有良好的加成性，得以明顯改善雜訊環境下的語音辨識效能。

在未來展望中，我們將進一步研究分頻段調變頻譜統計正規化法中的理論基礎，並希望能藉由更嚴謹的數學分析與推導，求取這些方法中最佳的分頻段數目。此外，我們也希望相關實驗不僅在數字辨識上處理，也擴展至其他較大字彙量的語音辨識，探討這一系列分頻段調變頻譜統計正規化法在不同複雜度之語音辨識系統的效能，或是應用於其他類型的干擾失真環境，進一步驗證這些新方法的效能與實用性，以上各點都是未來能夠嘗試研究發展的方向。期盼將來語音辨識之效能能夠更加提升，並且普遍應用於日常生活，讓人們輕鬆地利用語音與電腦或 3C 產品進行互動，令生活能夠更便利，使語音辨識之發展兼具理論性與實用性。

參考文獻

[1] 王小川, "語音訊號處理", 全華科技圖書, 2004

[2] Sadaoki Furui, "Cepstral analysis technique for automatic speaker verification", *IEEE Trans. on Acoustics, Speech and Signal Processing*, pp.254-272, 1981

[3] Olli Viikki and Kari Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition ", *Speech Communication*, vol. 25, pp.133-147, 1998

[4] Ángel de la Torre, Antonio M. Peinado, José C. Segura, José L. Pérez-Córdoba, Ma Carmen Benítez, and Antonio J. Rubio, "Histogram equalization of speech representation for robust speech recognition", *IEEE Trans. on Speech and Audio Processing*, pp.355-366, 2005

[5] Liang-che Sun, Chang-wen Hsu and Lin-shan Lee, "Modulation spectrum equalization for robust speech recognition", *in Proc. IEEE Workshop on Automatic Speech Recognition & Understanding* (ASRU), pp.81-86, 2007

[6] Hynek Hermansky and Nelson Morgan, "RASTA processing of speech", IEEE Trans. on Speech and Audio Processing, pp.578-589, 1994

[7] Hynek Hermansky and Petr Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR", *2005 International Conference on Spoken Language Processing* (*Interspeech*), pp.361-364

[8] Noboru Kanedera, Takayuki Arai, Hynek Hermansky, and Misha Pavel, "On the importance of various modulation frequencies for speech recognition", *1997 European Conference on Speech Communication and Technology* (*Eurospeech*)**,** pp. 1079-1082

[9] David Pearce and Hans-Günter Hirsch, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", *in Proc. of ISCA IIWR ASR2000*, Paris, France, pp.181-188, 2000

[10] ITU recommendation G.712, "Transmission performance characteristics of pulse code modulation channels", Nov. 1996

[11] Henry Stark, John W. Woods, "Probability and random processes with applications to signal processing", *3rd Edition*, Prentice-Hall, 2002