強健性語音辨識中基於小波轉換之分頻統計補償技術的研究
A Study of Sub-band Feature Statistics Compensation Techniques Based on a Discrete Wavelet Transform for Robust Speech Recognition

范顯騰 Hao-teng Fan
國立暨南國際大學電機系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
s96323516@ncnu.edu.tw

杜文祥 Wen-Hsiang Tu
國立暨南國際大學電機系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
s96323905@ncnu.edu.tw

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

摘要

本論文主要是發展語音特徵強健化技術，來改進雜訊環境下語音辨識的效能。我們改良原始全頻帶式的特徵序列統計正規化技術，使用著名的離散小波轉換來對語音特徵時間序列進行分頻帶的處理，進而發展出兩種新的特徵統計補償法，分別爲分頻式平均值與變異數正規化法與分頻式統計圖等化法。在這兩種新方法中，我們將經由離散小波轉換所得之分頻帶的序列，分別以平均值與變異數正規化法與統計圖等化法處理，再將處理後的各分頻帶之特徵序列，藉由反離散小波轉換組合成新的特徵序列。如此處理的特點爲，可以將特徵序列作不等切的調變頻帶切割，進而對語音辨識較重要的低調變頻帶作個別的強健性處理。從 Aurora-2 連續數字資料庫的實驗結果證實，我們提出的分頻式新方法在各種雜訊環境下都優於傳統全頻帶式之方法，與基礎實驗結果相比較，其相對錯誤降低率皆在 50%以上，顯示了我們所提出之新方法能十分有效地提昇語音特徵在雜訊環境下的強健性。

Abstract

The environmental mismatch caused by additive noise and/or channel distortion often degrades the performance of a speech recognition system seriously. Various robustness techniques have been proposed to reduce this mismatch, and one category of them aims to normalize the statistics of speech features in both training and testing conditions. In general, these statistics normalization methods deal with the speech feature sequences in a full-band manner, which somewhat ignores the fact that different modulation frequency components

have unequal importance for speech recognition.

With the above observations, in this paper we propose that the speech feature streams be processed in a sub-band manner. The processed temporal-domain feature sequence is first decomposed into non-uniform sub-bands using discrete wavelet transform (DWT), and then each sub-band stream is individually processed by the well-known normalization methods, like mean and variance normalization (MVN) and histogram equalization (HEQ). Finally, we reconstruct the feature stream with all the modified sub-band streams using inverse DWT. 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 MVN and HEQ provide relative error rate reductions of 20.32% and 16.39% over the conventional MVN and HEQ, respectively. These results reveal that the proposed methods significantly enhance the robustness of speech features in noise-corrupted environments.

關鍵詞：離散小波轉換、語音辨識、強健性語音特徵參數

keywords: speech recognition, discrete wavelet transform, robust speech features

一、緒論

近年來，語音處理之領域的學者持續地開發研究，使語音處理相關理論與技術不斷精進成熟，逐漸趨於實際應用的目的，就語音辨識(speech recognition)而言，其系統常因所在環境之雜訊干擾或是傳輸通道的效應，而使辨識效能受到明顯影響。針對這樣的問題，近年來的研究學者提出了一系列的環境強健性(environmental robustness)技術，藉此降低雜訊或通道干擾或凸顯語音的獨特成份，而達到明顯的改進效果，本論文的研究方向，即爲開發出新的降低雜訊與通道干擾之相關的語音強健性演算法。然而，跟過去相關之強健性技術較爲不同的是，我們採用了小波轉換(wavelet transform)，對於語音特徵之時間序列(temporal trajectory)加以處理，來改善語音特徵的強健性。

小波相關理論在訊號處理的範疇中雖已發展數十年，然而相對於其他許多理論而言，應用於在語音強健性處理之領域中仍偏少數，而其應用的方向大致上主要包含了：語音強化(speech enhancement)、語音端點偵測(voice activity detection, VAD)、強健性語音特徵(robust speech feature)與聽覺濾波器設計(auditory filter design)等。我們將它們簡述如下：

（一）語音強化(speech enhancement)

語音強化主要目的，通常是在一段訊號中，將雜訊抑制，並將語音訊號成份強調出來，常用的方式是假設雜訊在頻譜(spectrum)上具有較爲穩態(stationary)的特性，在頻域上將雜訊成份減低，例如設計一濾波器來過濾雜訊等。而以目前基於小波的信號強化方法，其中之一爲 Donoho[1]學者所提出使用小波收縮(wavelet shrinkage)的方式，其方法是由小波轉換所得之係數，經由門檻值的設定將雜訊適度地抑制。在其相關論文之實驗結果顯示了，透過小波轉換處理的語音強化效能比起之前所提出的傳統語音強化方法[2]要來的好。

（二）語音端點偵測(voice activity detection, VAD)

由於一段錄音(recording)裡可能包含有非語音的區段，如果一併辨識整段錄音，將會影響辨識處理的速度，並可能造成辨識精確度明顯下降。語音端點偵測(voice activity detection, endpoint detection)相關技術即是於決定出一段訊號中真正語音存在的位置。在傳統的作法上，以時域(time domain)而言，透過計算一段語音信號的能量(energy)或過零

率(zero-crossing rate)來決定含有語音成分的位置；在頻域(frequency domain)上，則通常是計算語音頻譜的熵(entropy)來獲得語音成分的資訊[3]。而小波在此方向上所提出的技術相對較多，譬如在文獻[4]中提到了使用小波轉換的係數能量比例判定語音及非語音(non-speech)成分，或是在另一[5]文獻裡提出計算小波係數之變異數，將其視爲一組隨機變數(random variable) 經由機率理論之結果判定，所得分類方法相較於之前方式能更精確判別出語音跟非語音之成份。

（三）強健性語音特徵擷取(robust speech feature extraction)

此類的語音處理技術方法目的是擷取不容易受到雜訊干擾的語音特徵參數，傳統的強健性語音特徵擷取技術大多數是在探討語音特徵的頻譜性質進而發展而得，換句話說，其所使用的轉換法爲有名的傅立葉轉換(Fourier transform)。然而小波處理也相繼應用於強健性語音特徵擷取技術上，例如，在[6]提出將原始梅爾倒頻譜特徵(mel-frequency cepstral co efficients, M FCC)中的離散餘弦轉換(discrete cosine trans form, DCT) 程序改變爲離散小波轉換(discrete wavelet transform, DWT)，其論文呈現的實驗結果顯示所得到的特徵比原始 MFCC 更具有雜訊環境之強健性。

（四）聽覺濾波器設計(auditory filter design)

一般而言，語音辨識中特徵參數求取程序裡所應用的語音聽覺濾波器組爲梅爾尺度(mel-scaled)的濾波器組，這些濾波器其分佈特性爲：1 kHz 頻率以下爲線性分佈，1 kHz 以上頻率爲非線性分佈，彼此相互部分重疊，其可近似模擬人耳聽覺效應。相對而言，小波處理之研究學者[7]也提出了利用小波包(wavelet packet) 的特性來仿效人耳聽覺效應，其適當透過一連串小波包轉換所切割的部份頻帶，選擇出能趨近於人耳聽覺的濾波器組效應，而由於小波處理所得之彼此頻帶間都假設爲不相關，即爲互不影響，因此所切割出來的各頻率範圍的語音信號都涵蓋了獨立的辨識資訊，其中的實驗結果驗證了以上的處理可以優於傳統的梅爾濾波器組處理，達到將語音辨識精確度提升的目的。

在本論文中，所發展出的新技術，並不同於上述所提的幾個傳統小波處理所應用的方向，而是著重於將小波處理其特殊的分頻技術適當地運用於語音特徵時間序列(temporal trajectory) 上，結合各種統計正規化的技術，來處理小波轉後各子頻帶的特徵時間序列，在之後的章節中我們將會逐步介紹此新技術，分析其主要觀念、作法與可能優於傳統技術的原因，並以一系列的實驗證實此新技術相對於傳統相近的技術而言，更能有效提昇語音辨識在雜訊干擾環境下的精確性。

本論文其餘的章節概要如下：在第二章裡，介紹目前常用之強健性特徵統計正規化法並探討傳統統計正規化法之可能缺失。在第三章，我們將簡要介紹離散小波轉換之分頻技術的實現，第四章爲本論文的重點，我們將在此章中介紹我們所提出的新方法，即兩種調變頻譜域的分頻統計特徵補償法：分頻帶平均值與變異數正規化法與分頻帶統計圖等化法，並對其初步效果加以介紹。在第五章，我們將執行一系列的語音辨識實驗，來驗證所提之新方法足以有效提昇語音特徵在雜訊環境下的強健性，最後，第六章則爲簡要結論，及未來可進一步研究的方向。

二、各種強健性技術介紹

在這裡我們首先目前常用之強健性特徵統計正規化法，之後探討傳統統計正規化法之可能缺失，並說明爲何使用小波轉換(discrete wavelet transform, DWT)改善這些問題。

由於語音辨識系統容易受到雜訊環境影響使得其辨識效能降低，因此語音處理相關研究的學者針對此雜訊干擾的問題，提出諸多的強健性技術，這些技術中有一大類是藉由正規化語音特徵的統計特性，來降低雜訊對語音特徵造成的失真。以下將介紹近年來在強健性語音辨識中常用的幾種語音特徵正規化技術。其中包含了：倒頻譜平均消去法

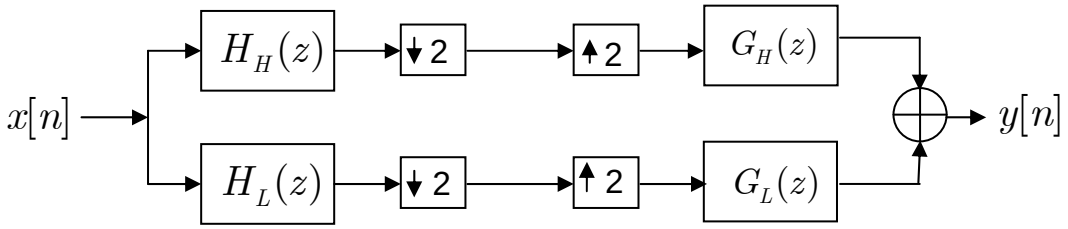
(cepstral mean subtraction, CMS)[8]、倒頻譜平均值與變異數正規化法(cepstral mean and variance normalization, MVN)[9]、倒頻譜平均與變異數正規化法結合自動回歸動態平均濾波器法(cepstral mean and variance normalization plus auto-regressive moving average filtering, MVA)[10]與統計圖正規化法(histogram equalization, HEQ)[11]等。

上述各種的正規化技術中，皆是把單一維特徵序列之所有特徵視爲同一個隨機變數的取樣(sample)，進而直接估測此隨機變數之統計參數，譬如期望值(mean)、變異數(variance)與機率分佈(probability distribution)等。雖然程序上易於實現，卻相對忽略了一段語句之中，其特徵隨時間變化的特性，例如調變頻譜的資訊。從另一觀點來看，這些作法等同於將全部調變頻率之成份一併做處理。然而根據過去許多的研究發現，不同的調變頻譜成份對於語音辨識擁有不同的重要性，更精確地說，在 N.Kanedera 學者[12]詳細指出大部分的語音辨識資訊分布在 1 Hz 和 16 Hz 的調變頻率之間，且主要集中在 4 Hz 附近。因此，許多知名且成功的時間序列濾波器(temporal filters)[13,14] ，都是特別強調出這些重要的調變頻率成分，進而顯示能有效改善雜訊環境下語音辨識的效能。

而前面介紹的各種特徵統計正規化演算法，可能缺失在於無法有效突顯不同調變頻率成份對於語音辨識的重要性，因此我們希望能把一特徵時間序列中的不同頻率成份分離出來，進而個別處理，初步的構想是能對於調變頻率較重要之低頻的部份較精細的處理，相對比較不重要之高頻的部份則使用較粗略的方式處理。基於此目的，我們發現小波轉換是個十分有用的工具，優點爲其能對一頻率區域作不等分的切割，即將訊號其較低頻率部分使用較窄的濾波器過濾出來，而高頻部分則用較寬的濾波器得之，之後對於每個子頻帶的特徵序列作統計正規化法。這樣的程序，相較於傳統的全頻帶式的特徵統計正規化法，理應可以進一步提昇處理後之特徵的強健性。之後一系列的章節，我們將逐步介紹小波轉換之分頻理論以及所提出的分頻特徵統計正規化法，最後以實驗結果證實此分頻式正規化法優於傳統之全頻式正規化方法。

三、小波轉換之分頻技術理論的概述

在這一章中，我們將專門討論小波轉換運用於離散時間訊號(discrete-time signal)的分頻(frequency division) 技術，此應算是小波轉換最常被用以處理訊號的方向。首先我們考慮一組典型雙通道的正交鏡像濾波器(quadrature-mirror filter bank, QMF)[15]，如圖三中所示：



圖三 雙通道 QMF 濾波器組

其中 $H_H(z)$ 與 $H_L(z)$ 表示爲分析(analysis)濾波器之高通與低通的轉換函數(transfer function)， $G_H(z)$ 與 $G_L(z)$ 則爲合成(synthesis)濾波器之高通與低通的轉換函數，且它們須符合以下的條件：

$$G_L(z) = H_H(z) \quad , \qquad G_H(z) = -H_L(-z) \tag{3-1}$$

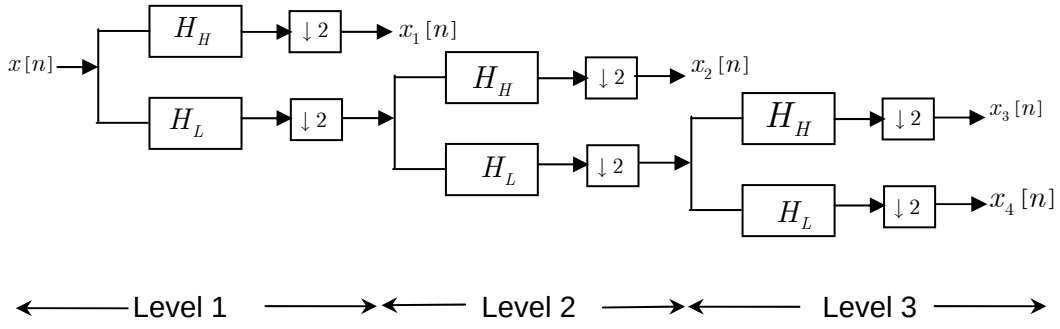
而在高通與低通分析濾波器之間存有以下關係：

$$H_H(z) = H_L(-z) \qquad \text{式(3-2)}$$

意即其頻率特性爲 $H_H\left(e^{j(\omega-\pi)}\right) = H_L\left(e^{j\omega}\right)$ ，其意義在於高通與低通濾波器之頻率響應會

以$\omega = \frac{\pi}{2}$爲中心形成左右對稱的圖形，在小波轉換中，即利用此形式的濾波器來對訊號作分頻處理。

圖四所表示了一訊號藉由上述之濾波器處理的分解程序(decomposition process) ，即離散小波轉換的分頻處理。其中，一連串的兩倍頻(octave-band)分析濾波組與之後的降低取樣(down-sampling)的組合通常被稱作二元樹(binary tree)結構，單一輸入序列經由分頻處理與降低取樣器(down-sampler)的轉換，輸出變爲各子頻帶序列(sub-sequences)的集合。在圖四中，我們看到了一個三階(three-level)的二元樹分析濾波器組結構，其中高通($H_H(z)$)與低通($H_L(z)$)濾波器都具有完全重構(perfect reconstru ction)的雙通道(two-channel)特性，即訊號通過此兩濾波器之後，並未喪失任何資訊或引進未知的干擾訊號，而得以將分頻後的訊號完美重建回原始訊號。另外，如果輸入此濾波器組的訊號$x[n]$長度爲N，在第一階高通分析濾波器之輸出$x_1[n]$即約爲$N/2$，而再下一階高通分析濾波器輸出$x_2[n]$約爲$N/4$，如此重複這程序，就可以得到所有階層之濾波器的輸出。表一列出了各層濾波器的其頻帶範圍及輸出訊號的長度。以上所述之兩倍頻(octave)完全重構 QMF 濾波器組對輸入訊號的處理程序，即爲離散小波轉換(discrete wavelet transform, DWT)，由上述可知，如果所用之濾波器組的階層數爲L（相當於L層的離散小波轉換），則輸出的總分頻訊號的數目則爲$L+1$個。



圖四　離散小波轉換的分解程序圖（階層數爲 3）

表一、三層離散小波轉換(DWT)每一階層的輸出訊號點數及相對應的頻率範圍
（$x[n]$取樣頻率爲F_s Hz ）

訊號	總點數	頻率範圍
$x[n]$	N	$\left[0, F_s/2 \text{ Hz}\right]$
$x_1[n]$	$N/2$	$\left[F_s/4 \text{ Hz}, F_s/2 \text{ Hz}\right]$
$x_2[n]$	$N/4$	$\left[F_s/8 \text{ Hz}, F_s/4 \text{ Hz}\right]$
$x_3[n]$	$N/8$	$\left[F_s/16 \text{ Hz}, F_s/8 \text{ Hz}\right]$
$x_4[n]$	$N/8$	$\left[0, F_s/16 \text{ Hz}\right]$

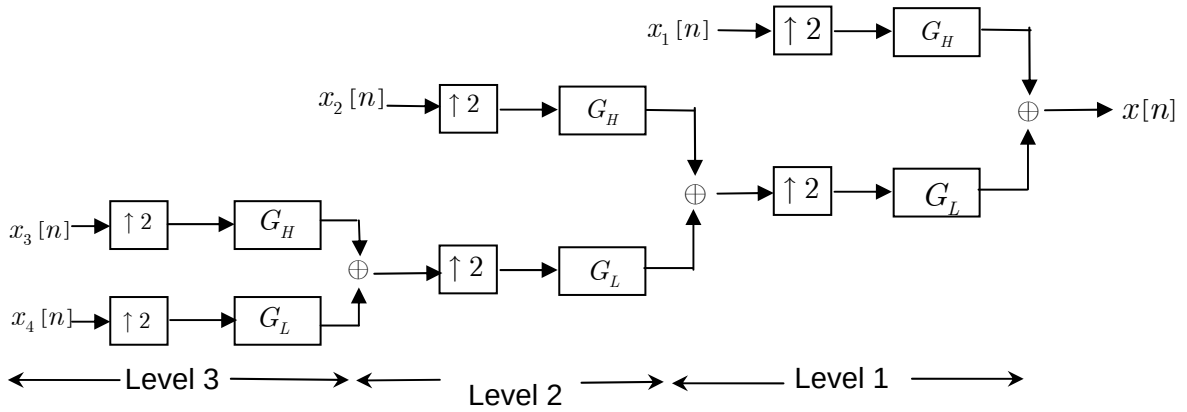
從上表一可知，如果序列$x[n]$涵蓋的頻率範圍爲[0, $F_s/2$ Hz]，其中F_s爲$x[n]$的取樣頻率，則經由第一階正交鏡像濾波器組之高頻輸出$x_1[n]$，頻率範圍爲$\left[F_s/4 \text{ Hz}, F_s/2 \text{ Hz}\right]$，依此類推，逐步往低頻率部份做不等分切割，隨著頻率越高，其

頻寬則越大。由上所述，離散小波轉換的第 k 個輸出 $x_k[n]$，相當於是原始序列 $x[n]$ 與第 k 個帶通濾波器之脈衝響應(impulse response)相互摺積(convolution)的結果，如式(3-3)所示：

$$x_k[n]=\begin{cases}\sum_{m=-\infty}^{\infty} h_{k,1}\left[2^{k+1}n-m\right]x[m], & 0\le k\le L-1,\\ \sum_{m=-\infty}^{\infty} h_k\left[2^{k}n-m\right]x[m], & k=L.\end{cases} \qquad \text{式(3-3)}$$

其中 $h_{k,1}\left[2^{k+1}n\right]$ 與 $h_k\left[2^k n\right]$ 爲原始脈衝響應 $h_{k,1}[n]$ 與 $h_k[n]$ 降低取樣而得，而高通濾波器之輸出，稱爲細節(detail)係數；低通濾波器之輸出則稱爲近似(approximation)係數。

若要藉由所有子頻帶訊號的集合得到原始序列 $x[n]$，其過程稱爲重建程序(reconstruction process)，此恰爲前述之分解程序的反程序(inverse process)，即使用所得之 $\{x_k[n]\}$ 經 L 階兩倍頻完全重構 QMF 合成濾波器組逐層處理，此過程即爲反離散小波轉換(inverse discrete wavelet transform, IDWT)，如下圖五所示：



圖五　反離散小波轉換的重建程序圖（階層數爲 3）

還原程序其數學式如式(3-4)：

$$x[n]=\sum_{k=0}^{L-1}\sum_{m=-\infty}^{\infty} g_{k,1}\left[n-2^{k+1}m\right]x_k[m]+\sum_{m=-\infty}^{\infty} g_k\left[n-2^{L}m\right]x_L[m], \qquad (3\text{-}4)$$

其中 $g_{k,1}\left[2^{k+1}n\right]$ 與 $g_k\left[2^k n\right]$ 分別爲原始脈衝響應 $g_{k,1}[n]$ 與 $g_k[n]$ 提升取樣而得。圖五之還原程序，即是將各子頻帶的訊號以提升取樣(up-sampling)的方式增加序列點數，再經過高通($G_H(z)=H_H(z)$)與低通($G_L(z)=H_L(z)$)之合成濾波器處理，如果第三階輸入訊號點數爲 $N/8$，則在第三階輸出訊號點數約爲 $N/4$，而第二階輸出訊號點數約爲 $N/2$，如此重覆此程序，則最後所得之訊號爲原始 N 點之訊號 $x[n]$:

以上所述爲小波轉換之分析(analysis)與合成(synthesis)程序，經由此轉換後，訊號被分解成各個子頻帶之訊號，如表一所示，低頻部分的子頻帶頻寬較小，而高頻部分的子頻帶頻寬較大。藉由以上所述的離散小波轉換程序，我們可以將語音特徵時間序列作分頻的處理，進而針對不同調變頻帶成分的語音特徵序列分別作處理，在下一章裡，我們將介紹其對應的的分頻式特徵統計補償法。

四、分頻帶特徵統計正規化法

在這一章中，我們首先在第一節介紹所新提出之分頻帶特徵統計補償法的步驟及特性，接著在第二節中，我們將以一語句爲例，驗證所提之新方法足以有效降低雜訊對語音調變頻譜之干擾。

(一) 分頻帶特徵統計正規化法的步驟說明

假設一段語句(utterance)的某一維梅爾倒頻譜語音特徵以下式(4-1)表示:

$$\left\{x^{(m)}[n];1<n\le N\right\},\ 0\le m\le M-1\ , \tag{4-1}$$

其中 N 爲此特徵序列的總音框數， M 表示每一音框中的特徵總數。此特徵序列相當於涵蓋了全調變頻帶(full-band)的語音資訊，然而，由前面章節所述，不同的頻帶成份，對於語音辨認的重要性有所不同，基於此項理由，這裡我們使用分頻的技術，將此特徵序列分解成各不同頻率的成分，如以下步驟（爲了簡易說明起見，我們在之後的討論中，將省略式(4-1)中代表不同維特徵的上標"(m)"，因爲我們是對每一個不同維的特徵序列皆作同樣處理)：

首先，我們將原始特徵序列 $\{x[n]\}$ 切割成 L 個分頻帶且假設每一分頻帶都爲各自獨立，而每一頻帶中的序列表示爲 $\left\{x_\ell[n]\right\},1\le\ell\le L$ ，此切割頻帶的方法是將原始特徵通過一倍頻(octave-band)帶通濾波器組，每一子頻帶訊號再作降低取樣(down-sampling)處理，此步驟等效於執行 $(L-1)$ 階的離散小波轉換(discrete wavelet transform, DWT) 於特徵序列 $x[n]$ 上。另外，假設特徵序列 $\{x[n]\}$ 音框取樣率爲 F_s (Hz) ，則其調變頻譜頻率範圍爲 $\left[0,F_s/2\right]$ ，因此，第 ℓ 個分頻帶序列的頻率範圍，可被近似表示成式(4-2)：

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

在 DWT 程序中，其方式是將一主頻帶依頻寬先等切爲兩個副頻帶，然後保持高頻帶不動，將低頻帶再等切成兩個副頻帶，如此反覆進行，因此相當於低頻部份會使用較多個頻寬較小的濾波器，而高頻部份則用較少個頻寬較大的濾波器，而因爲 DWT 程序中的降低取樣(down-sampling)的運算，所以每一分頻帶的序列 $\left\{x_\ell[n]\right\}$ 長度約正比於頻寬的大小。

接著，將上步驟所得的分頻帶序列 $\left\{x_\ell[n]\right\}$ 做特徵統計正規化，得到新的分頻帶序列，表示爲 $\left\{\tilde{x}_\ell[n]\right\}$ ，其特徵統計正規化的方式是將每一語句之子頻帶特徵 $\left\{x_\ell[n]\right\}$ 的統計量，譬如平均值(mean)、變異數(variance)或是更高階的動差(moments)作處理，使新的特徵參數 $\left\{\tilde{x}_\ell[n]\right\}$ 的統計量等同或逼近一目標(target)統計量，而此目標統計量是由乾淨訓練語料庫中，所有語句之子頻帶特徵 $\left\{x_\ell[n]\right\}$ 估測計算而得。在這裡我們使用的特徵統計正規化法有兩種，分別爲倒頻譜平均值與變異數正規化法(MVN)與統計圖等化法(HEQ)，以 MVN 法而言，所得新的分頻帶序列 $\left\{\tilde{c}_\ell[n]\right\}$ 可表示爲下式(4-3)：

$$\tilde{x}_\ell[n]=\left(\frac{x_\ell[n]-\mu_{\ell,s}}{\sigma_{\ell,s}}\right)\times\sigma_{\ell,t}+\mu_{\ell,t} \tag{4-3}$$

其中 $\mu_{\ell,s}$ 與 $\sigma_{\ell,s}^2$ 分別爲目前處理的單一(single)分頻帶序列 $\left\{x_\ell[n]\right\}$ 的平均值與變異數，而 $\mu_{\ell,t}$ 與 $\sigma_{\ell,t}^2$ 爲目標(target)平均值與變異數，此目標平均值與變異數是由原始乾淨訓練語料庫中所有分頻帶特徵序列 $\left\{x_\ell[n]\right\}$ 估測而得。同樣地，如以 HEQ 作爲統計補償法，則 $\left\{\tilde{x}_\ell[n]\right\}$ 與 $\left\{x_\ell[n]\right\}$ 彼此關係爲下式(4-4)：

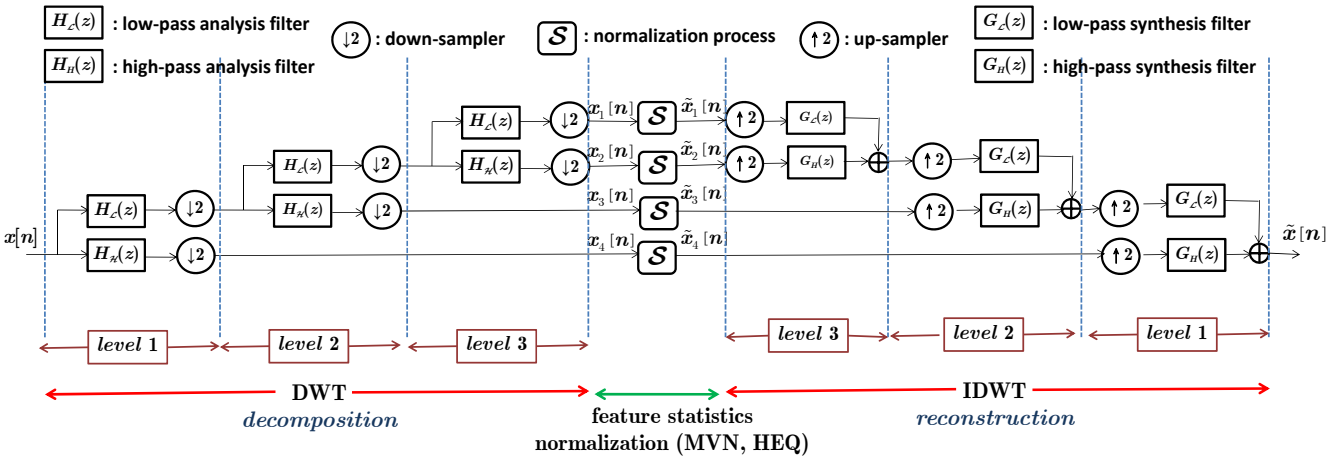
$$\tilde{x}_\ell[n]=F_{X,t}^{-1}\left(F_{X,s}\left(x_\ell[n]\right)\right) \tag{4-4}$$

其中 $F_{X,s}(.)$ 爲目前處理的單一分頻帶序列 $\left\{x_\ell[n]\right\}$ 所估測的機率分佈函數(probability distribution function)，而 $F_{X,t}(.)$ 是由原始乾淨訓練語庫中所有分頻帶特徵序列 $\left\{x_\ell[n]\right\}$ 所

估測而得的機率分佈函數。

最後，將所有的分頻帶序列$\{\tilde{x}_\ell[n]\}$（包含了更新過後與未更新的分頻帶序列）透過$(L-1)$階反離散小波轉換(inverse discrete wavelet transform, IDWT)，重建爲新的特徵時間序列，此即爲我們最後使用之語音特徵序列$\{\tilde{x}[n]\}$。

上述分頻帶統計正規化法的流程圖繪於下圖六：



圖六 分頻帶特徵統計正規化法的運作程序圖

爲了在之後的討論中，有效區隔傳統方法與所提出的新方法，對傳統全頻帶(full-band)的特徵統計正規化法 MVN 與 HEQ，我們分別稱之爲 FB-MVN 與 FB-HEQ，而如式(4-3)與(4-4)中分頻(sub-band)處理的特徵統計正規化法，我們則分別稱爲 SB-MVN 和 SB-HEQ。相較於傳統的全頻帶統計補償法，我們所提出之分頻帶統計補償法有以下幾點相異之處：

1. 傳統的全頻帶 MVN(FB-MVN)法中，任一特徵序列的平均值與變異數通常分別被正規化爲 0 與 1，但對於 SB-MVN 而言，不同分頻帶的特徵序列並不擁有相同的目標平均值與變異數，因此不同分頻帶特徵序列即使在正規化後，仍保有彼此統計特性的差異。相同地，SB-HEQ 也是具有此特性，不同的分頻帶特徵序列對應至不同的目標機率分佈函數。
2. 在 SB-MVN 與 SB-HEQ 中，可任意選擇某些分頻帶序列來作正規化。一般而言，對於語音辨識來說，低(調變)頻率的成分，包含的語音鑑別資訊較多，因此我們通常優先選擇低頻率的分頻帶特徵加以正規化。但是，如果有些非穩態雜訊(non-stationary noise)存在於高調變頻率的區域，爲了降低此類雜訊干擾，就須將高頻的分頻帶考慮進去一同處理。
3. 由於 DWT 程序中的降低取樣(down-sampling)步驟，我們所需處理之所有分頻帶序列的特徵總數近似等同於原始序列的特徵總數，因此處理上並不會因爲增加分頻帶的數目而使計算複雜度大幅提升。但若以傳統的分頻濾波器組(filter-bank)之方法，所需處理的總特徵數會明顯隨分頻帶的個數而增加，相對而言，其運算的複雜度會因此大幅提高。

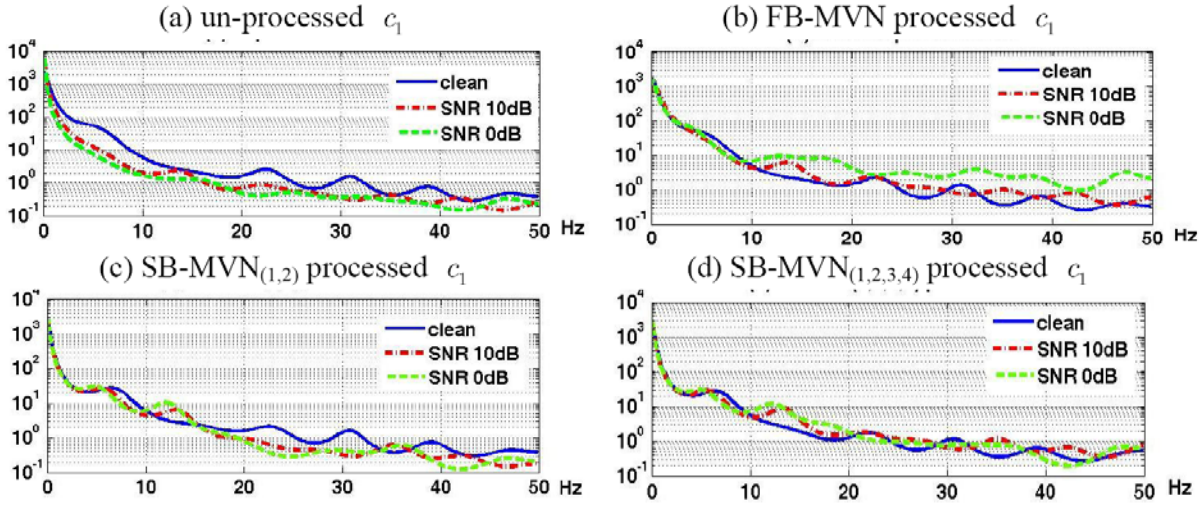
（二）分頻帶特徵統計正規化法的初步效能討論

在這裡，我們將所提出的分頻帶統計正規化法跟原始之全頻帶統計正規化法作初步的效能比較，根據這些方法在一語音特徵序列之調變頻譜的失真改善程度，來評估這些方法的效能。我們使用 AURORA-2 資料庫[20]裡的 MAH_27O6571A 語音檔，然後加入不同訊雜比(SNR)的地下鐵(subway)雜訊，繼而加以處理。

在我們所提出之方法中，初步使用了三階的 DWT 轉換，將整個調變頻帶[0, 50 Hz]切割出四種分頻帶範圍，分別是[0, 6.25 Hz]、[6.25 Hz, 12.5 Hz]、[12.5 Hz, 25 Hz]和[25 Hz, 50 Hz]，（由於特徵音框取樣率爲 100 Hz，因此特徵序列涵蓋之頻率範圍爲[0, 50 Hz]）。在之後討論的每個頻帶之正規法中，我們在方法名稱右下方使用下標數字來表示被正規化的頻帶，例如 SB-MVN$_{(1,2)}$與 SB-HEQ$_{(1,2)}$表示了第一個分頻帶（[0, 6.25 Hz] ）與第二個分頻帶([6.25 Hz, 12.5 Hz])使用了 MVN 或 HEQ 處理，剩餘的兩個高頻帶([12.5 Hz, 25 Hz]和[25 Hz, 50 Hz])則維持不動，而 SB-MVN$_{(1,2,3,4)}$與 SB-HEQ$_{(1,2,3,4)}$ 表示了全部四個分頻帶皆個別以 MVN 或 HEQ 處理。

首先，我們對於全頻帶與各種分頻帶之 MVN 法的處理結果加以討論。圖七(a)(b)(c)(d)分別表示爲原始未處理之第一維 MFCC(c_1)特徵序列、FB-MVN、SB-MVN$_{(1,2)}$與 SB-MVN$_{(1,2,3,4)}$處理後之 c_1 序列之功率頻譜密度(power spectral density, PSD)曲線。

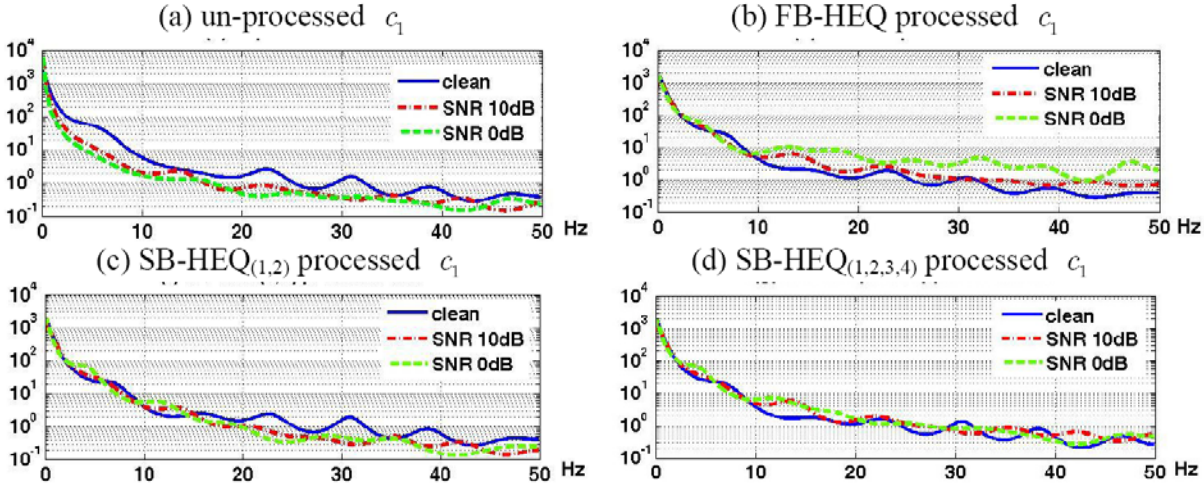
在圖七(a)中，可看出不同 SNR 值下(clean, 10 dB 與 0dB)之未處理過的 c_1 序列，其 PSD 曲線，受到加成性雜訊(additive noise) 的影響，存在嚴重的失真情形。而經由圖七(b)可看出，FB-MVN 處理後之 c_1 序列，在較低的調變頻率[0, 10 Hz]之間，其 PSD 失真的情況很明顯降低，但在高調變頻率範圍[10Hz, 50 Hz]，PSD 失真的情形並沒有太大的改善。圖七(c)爲 SB-MVN$_{(1,2)}$所得之特徵序列之 PSD 圖，其所處理的頻帶分別爲[0, 6.25 Hz]和[6.25 Hz, 12.5 Hz] ，從此圖可以發現，約在調變頻率 20 Hz 以下，其 PSD 失真情形相對減低，但在未處理的調變頻率範圍[12.5 Hz, 50 Hz]，同樣存有明顯的失真情況。圖七(d)爲 SB-MVN$_{(1,2,3,4)}$所得之特徵序列之 PSD 圖，其所處理的頻帶分別爲[0, 6.25 Hz]、[6.25 Hz, 12.5 Hz]、[12.5 Hz, 25 Hz]與[25 Hz, 50 Hz]，很明顯可看出在全部的調變頻率範圍，其 PSD 失真的情況皆有效降低。



圖七 (a) 原始 c_1 特徵序列及(b)FB-MVN、(c)SB-MVN$_{(1,2)}$與(d)SB-MVN$_{(1,2,3,4)}$ 作用在不同訊雜比下之 c_1 特徵序列之功率頻譜密度曲線圖

接下來，圖八(a)(b)(c)(d)分別表示爲原始未處理之第一維 MFCC(c_1)特徵序列、FB-HEQ、SB-HEQ$_{(1,2)}$與 SB-HEQ$_{(1,2,3,4)}$處理後之 c_1 序列之 PSD 曲線，其中括弧中的數字表示所處理的頻帶。比較圖八(a)與圖八(b)可知，對於較低的調變頻率範圍[0, 10 Hz] ，FB-HEQ 可有效降低 PSD 之失真，但對於其他調變頻率範圍[10 Hz, 50 Hz] ，PSD 失真的情形並沒有獲得太大的改善。圖八(c)爲 SB-HEQ$_{(1,2)}$所得之特徵序列之 PSD 圖，其所處理的頻帶分別爲[0, 6.25 Hz]與[6.25 Hz, 12.5 Hz] ，在此圖中，可以發現約在調變頻率

20 Hz 以下之 PSD 失真現象相對被減低，但在其他調變頻率範圍，仍有明顯的失真情況。跟之前圖七(c)SB-MVN$_{(1,2)}$的效果比較，可看出 SB-HEQ$_{(1,2)}$優於 SB-MVN$_{(1,2)}$ ，更有效降低約在頻率 20 Hz 以下的 PSD 失真度。圖八(d)爲 SB-HEQ$_{(1,2,3,4)}$所得之特徵序列之 PSD 圖，其所處理的頻帶個別爲[0, 6.25 Hz]、[6.25 Hz, 12.5 Hz]、[12.5 Hz, 25 Hz]與[25 Hz, 50 Hz] ，從此圖很明顯可看出全部的調變頻率範圍之 PSD 曲線，其失真的情況皆已有效降低。類似之前的狀況，當比較圖八(d)與圖七(d)時，可看出 SB-HEQ$_{(1,2,3,4)}$在降低 PSD 失真的性能上優於 SB-MVN$_{(1,2,3,4)}$。



圖八 (a) 原始c_1特徵序列、(b)FB-HEQ、(c)SB-HEQ$_{(1,2)}$與(d)SB-HEQ$_{(1,2,3,4)}$ 作用在不同訊雜比下之c_1特徵序列之功率頻譜密度曲線圖

五、調變頻譜分頻帶正規化法的辨識實驗結果與討論

本章主要內容爲呈現並分析一系列的強健性特徵技術所得之語音辨識的效果，這些技術包括了傳統的全頻式特徵統計正規化法、我們所新提出的分頻式 MVN(SB-MVN)法與分頻式 HEQ(SB-HEQ)法。

（一）實驗環境與架構設定

本辨識實驗所採用的語音資料庫爲歐洲電信標準協會(European Telecommunication Standard Institute, ETSI) 所發行的語料庫 AURORA-2[16]，內容是以美國成年男女所錄製的一系列連續的英文數字字串，測試語音本身加上各種加成性雜訊或通道效應的干擾。加成性雜訊共有八種，分別是地下鐵(subway)、人聲(babble)、汽車(car)、展覽會館(exhibition)、餐廳(restaurant)、街道(street)、飛機場(airport)和火車站(train station)雜訊等；而通道效應有兩種，分別爲 G712 和 MIRS。雜訊比例的大小包含了乾淨無雜訊的狀態(clean)，以及六種不同雜訊比(signal to noise ratio, SNR)，分別是 20 dB、15 dB 、10 dB 、5 dB 、0 dB 與-5 dB ，因此我們可以觀察分析不同雜訊環境下對於語音辨識的影響。由於雜訊的不同，測試環境可分爲 Set A、Set B 與 Set C 三組。

在辨識中所使用的聲學模型是由隱藏式馬可夫模型工具(Hidden Markov Model Tool Kit, HTK)[17] 訓練而得，包括了 11 個數字模型(zero, one, two,…, nine 及 oh)以及靜音(silence)模型，每個數字模型則有 16 個狀態，各狀態包含 20 個高斯密度混合。

（二）全頻帶補償法與各種分頻帶正規化法之實驗結果

本章節實驗採用梅爾倒頻譜係數(mel-frequency cepstral coefficients, MFCC)

特徵參數13維(c0~c12)，加上一階與二階差量，總共爲39維特徵參數。在表三中，我們呈現了基礎實驗(baseline)、各種分頻式 SB-MVN 與 SB-HEQ、全頻式 FB-MVN 和 FB-HEQ 作用在原始 MFCC 特徵上所得的平均辨識結果（不同種辨識環境的平均辨識率及相對改善率），其中 RR1和 RR2分別爲相較於基礎實驗和全頻帶法之相對錯誤降低率(relative error rate reductions)。表四列出在各種不同的 SNR 值下的各種方法的平均辨識率，而圖九簡要畫出各方法平均辨識率的比較圖。

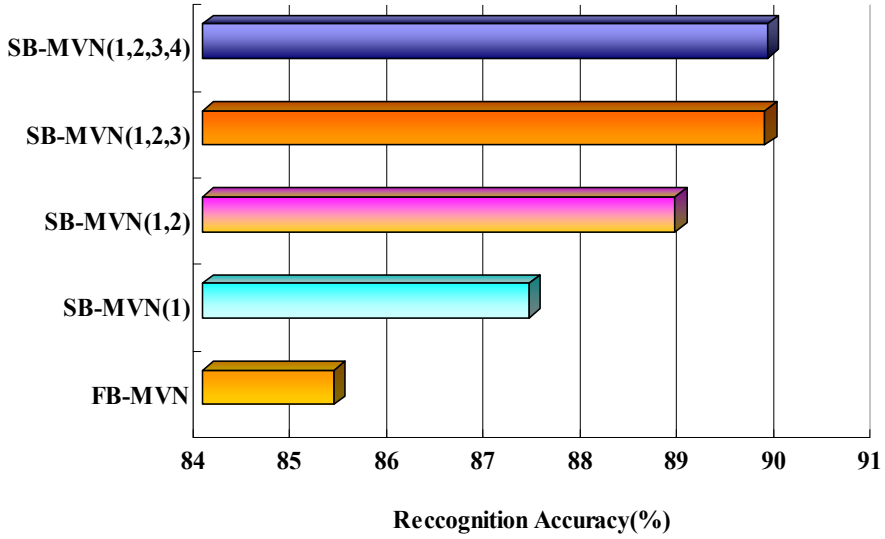
表三、各分頻帶方法與全頻帶方法的平均辨識率(%)與相對錯誤降低率(%)

Method	Set A	Set B	Set C	Avg.	RR1	RR2
Baseline	71.92	68.22	77.61	71.58	—	—
FB-MVN	85.03	85.56	85.60	85.36	48.49	—
SB-MVN$_{(1)}$	86.87	87.90	87.37	87.38	55.59	13.80
SB-MVN$_{(1,2)}$	87.28	90.23	89.44	88.89	60.91	24.11
SB-MVN$_{(1,2,3)}$	89.44	90.31	89.61	89.82	64.18	30.46
SB-MVN$_{(1,2,3,4)}$	89.47	90.31	89.62	**89.84**	**64.25**	**30.60**
FB-HEQ	87.59	88.84	87.64	88.10	58.13	—
SB-HEQ $_{(1)}$	87.70	89.31	87.81	88.37	59.08	2.27
SB-HEQ $_{(1,2)}$	89.22	90.55	90.23	89.95	64.64	15.55
SB-HEQ$_{(1,2,3)}$	89.51	90.75	89.54	90.01	64.85	16.05
SB-HEQ $_{(1,2,3,4)}$	89.51	90.83	89.57	**90.05**	**64.99**	**16.39**

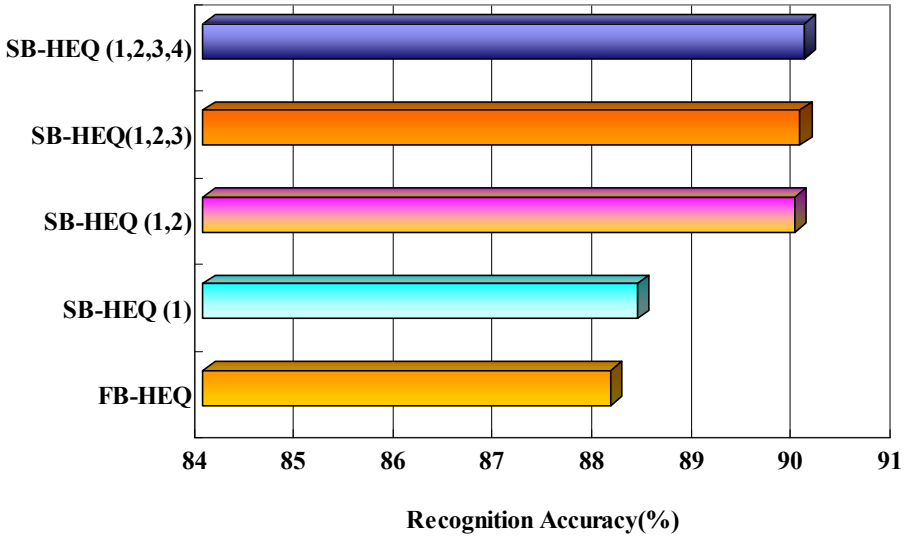
表四、所有不同 SNR 值雜訊環境下的平均辨識率(%)

Method	clean	20dB	15dB	10dB	5dB	0dB	-5dB
Baseline	99.79	95.80	88.15	73.81	56.32	43.82	40.13
FB-MVN	99.82	98.73	96.83	91.88	79.52	59.80	46.70
SB-MVN$_{(1)}$	99.79	98.67	96.98	92.24	82.42	66.60	51.95
SB-MVN$_{(1,2)}$	99.80	98.97	97.76	94.33	86.40	70.99	53.80
SB-MVN$_{(1,2,3)}$	99.78	98.99	97.75	94.57	86.59	71.19	53.69
SB-MVN$_{(1,2,3,4)}$	99.81	98.97	97.75	94.51	86.60	71.34	53.72
FB-HEQ	99.77	99.01	97.76	94.22	84.30	65.21	48.96
SB-HEQ $_{(1)}$	99.72	98.72	97.31	93.34	83.98	68.49	52.70
SB-HEQ $_{(1,2)}$	99.64	98.84	97.64	94.50	87.52	71.28	53.65
SB-HEQ$_{(1,2,3)}$	99.66	98.84	97.70	94.68	87.09	71.74	53.78
SB-HEQ $_{(1,2,3,4)}$	99.64	98.85	97.69	94.74	87.15	71.83	54.01

(a) 不同形式之平均值與變異數正規化法的辨識率比較



(b) 不同形式之統計圖等化法的辨識率比較



圖九 各分頻帶方法與全頻帶方法的平均辨識率(%)之綜合比較圖

從表三、表四和圖九可發現，我們所新提出的分頻帶正規化法，確實能有效提昇其雜訊環境下的強健性，其詳細現象如以下幾點：

1. 無論全頻帶與分頻帶正規化方法，相較於基本實驗而言，都有良好的改善效能，相對錯誤降低率都在 48%以上，除此之外，每一種 HEQ 的效果都比其相同形式的 MVN 來的好。相較於 MVN，HEQ 額外對於特徵高階動差做補償處理，所以整體來說，HEQ 更有助於改善雜訊環境所造成的特徵失真。
2. SB-MVN 的四種分頻模式效能都優於原始全頻式的 FB-MVN，此情況在 SB-HEQ 與 FB-HEQ 之間的比較也是如此。而 SB-MVN 和 SB-HEQ 相較於原始 FB-MVN 和 FB-HEQ 的相對錯誤降低率分別高達 30.60%與 16.39%，此結果顯示所提出的新分頻處理技術優於傳統全頻帶的處理，因此我們成功的驗證了之前章節的推論，即不同的調變頻譜成份對於語音辨識有不同的重要性，對不同頻帶分別作補償可帶來更好的效能。
3. 從表四中觀察在不同 SNR 值情況下的平均辨識率，我們可知在不受任何雜訊干擾

之匹配情況下，所有方法都有很高的辨識率，也就是說這些方法並不會降低與原始 MFCC 高鑑別度的特性。但在受到不同雜訊干擾之不匹配情況下，從表中可看出所有方法都能有效改善辨識效果，即增加原始 MFCC 特徵的強健性，在 SNR 值爲 20 dB 和 15 dB 時，分頻帶與全頻帶方法其效能差異並不顯著，如果訊雜比繼續下降時，可以發現到分頻式的 SB-MVN 與 SB-HEQ 平均辨識率明顯優於全頻式的 FB-MVN 和 FB-HEQ。

4. 對於不同分頻式的 SB-MVN 與 SB-HEQ，若只正規化最低的頻帶[0, 6.25 Hz](即 $\text{SB-MVN}_{(1)}$和 $\text{SB-HEQ}_{(1)}$)，其相對於基礎實驗就有顯著改善效果。當我們增加正規化的頻帶數目時，相對改善率都有明顯的成長，特別是在處理兩低頻帶([0, 6.25 Hz]與[6.25 Hz, 12.5 Hz])後，也就是 $\text{SB-MVN}_{(1,2)}$和 $\text{SB-HEQ}_{(1,2)}$時，幾乎其效能已是最佳，如進一步再處理較高的分頻帶，如 $\text{SB-MVN}_{(1,2,3,4)}$或 $\text{SB-HEQ}_{(1,2,3,4)}$，所改善的效果就有限。此結果顯示與過去文獻互相吻合，即在 1 到 16Hz 之間的調變頻率成分，對於語音辨識而言是相對重要的。

六、結論與未來展望

我們提出了兩種分頻式特徵統計正規化技術，分別爲分頻式平均與變異數正規化法(sub-band cepstral m ean and variance norm alization, SB-MVN) 與分頻式統計圖等化法(sub-band histogram equalization, SB-HEQ) ，我們使用了著名的離散小波轉換(discrete wavelet tra nsform, DW T)來對語音的特徵時間序列作分頻處理，其特點在於利用 DWT 可以將對語音辨識較有幫助之調變頻譜低頻成份做較細緻的切割，高頻部份則相對應的切割區間數較少。之後，對每個子頻帶的特徵序列個別作統計正規化處理，再將所有子頻帶的特徵序列以反離散小波轉換，組合成新的特徵序列。經由連續數字之語音辨識實驗，顯示了上述之分頻式的新方法相對於傳統全頻式的方法，更能提昇雜訊干擾環境下語音辨識的精確度，相當於這些新方法能更有助於增加語音特徵的強健性。除了此優點外，此分頻式的新方法並未增加所須處理之語音特徵的個數，因此並不會因所分頻帶數目的增加，而大幅增加執行的複雜度。

在未來研究中，我們期望相關的實驗不只限制在連續數字辨識中，而是進一步應用在較大字彙之語音資料庫，探究其效能爲何。其次，我們亦將執行其他種類的特徵統計正規化於此所述的分頻帶特徵序列上，譬如倒頻譜增益正規化法(CGN)[18]、高階倒頻譜動差正規化法(HOCMN)[19]與倒頻譜形狀正規化法(CSN)[20]等，進一步驗證在這些方法中，分頻處理上是否能得到更好的效能。另外，我們也會嘗試使用各種不同的小波函數，分析每一種小波函數的特性，並探討這些函數對於所述之分頻式技術的影響，或是使用有別於小波轉換的小波包(wavelet pack et)[21]，研究這兩者對於所提之新方法上效能的差異。希望在未來，小波轉換能更成熟地應用於語音相關的分析研究上，使語音強健性技術能趨於成熟與多樣化，使語音處理更具理論性與實用性。

參考文獻

[1] D.L. Donoho, "De-noising by soft-thresholding", *IEEE Trans. on Information Theory*, vol. 41, no. 3, pp. 613-627, May 1995.

[2] S. F . Boll, "Suppression of aco ustic noise in speech u sing spectral subtraction ", *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113-120, 1979

[3] B.-F . W u, K.-C. W ang, "Noise spectrum estim ation with entropy-based V AD i n non-stationary environm ents", *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E89-A, no. 2, Feb 2006

[4] J.K. Lee , C.D. Y oo, "Wavelet speech e nhancement based on voiced /unvoiced decision",

32nd Inter-Noise, pp. 4149-4156, Aug 2003
[5] X. Zhang, Z. Zhao and G. Zhao, "A speech endpoint detection method based on wavelet coefficient variance and sub-band amplitude variance", *International Conference on Innovative Computing, Information and Control*, vol. 3, pp. 83-86, 2006
[6] J.N. Gowdy and Z. Tufekci, "Mel-scaled discrete wavelet coefficients for speech recognition", *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), vol. 3, pp. 1351-1354, 2000
[7] M. Siafarikas, T. Ganchev and N. Fakotakis, "Objective wavelet packets features for speaker verification" , *in Proc. International Conference on Spoken Language Processing* (ICSLP), pp. 2365-2368, 2002
[8] S. Furui, "Cepstral analysis technique for automatic speaker verification", *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, Apr 1981
[9] C.-P. Chen, K. Filaliy and J. A. Bilmes, "Frontend post-processing and backend model enhancement on The Aurora 2.0/3.0 databases' ', *in Proc. International Conference on Spoken Language Processing* (*ICSLP*), 2002
[10] C.-P. Chen and J. -A. Bilmes, "MVA processing of speech features", *IEEE Trans. on Audio, Speech, and Language Processing*, vol.15, no. 1, pp.257-270, Jan 2006.
[11] F. Hilger and H. Ney , "Quantile based histogram equalization for noise robust speech recognition", *in European Conference on Speech Communication and Technology* (*Eurospeech*), 2001
[12] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the importance of various modulation frequencies for speech recognition ", *in European Conference on Speech Communication and Technology* (*Eurospeech*), 1997
[13] H. Hermansky and N. Morgan, "RASTA Processing of Speech", *IEEE Trans. on Speech and Audio Processing*, vol.2, no. 4, Oct. 1994.
[14] J- W. Hung and L-S. Lee, "Optimization of temporal filters for constructing robust features in speech recognition", *IEEE Trans. on Audio, Speech and Language Processing*, vol.4, no. 3, May 2006
[15] D. Esteban and C. Galand. "Application of quadrature mirror filters to spilt-band voice coding schemes", *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), vol. 2, pp. 191-195, May 1977.
[16] H. G . Hirsch and D. Pearce, "The A URORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions ", *in Proc. of ISCA IIWR ASR2000*, Paris, France, 2000
[17] http://htk.eng.cam.ac.uk/
[18] S. Yoshizawa et al., "Cepstral gain normalization for noise robust speech recognition," in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), vol. 1, pp. I-209-12, May 2004
[19] C.-W. Hsu and L.-S. Lee, "Higher order cepstral moment normalization (HOCMN) for robust speech recognition ", *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), vol. 17, no. 2, pp. 205-220, Feb 2004
[20] J. Du and R.-H W ang, "Cepstral shape normalization (CSN) for robust speech recognition", *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 4389-4392, April 2008
[21] R. R. Coifman and M. V . W ickerhauser. "Entropy-based algorithms for best basis selection", *IEEE Trans. on Information Theory*, vol. 38, no. 2, pp. 713-718, March 1992