調變頻譜正規化法使用於強健語音辨識之研究
Study of Modulation Spectrum Normalization Techniques for Robust Speech Recognition

王致程 Chih-Cheng Wang
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University, Taiwan
s95323553@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
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
jwhung@ncnu.edu.tw

摘要

自動語音辨識在實際系統應用中，語音信號經常受到環境雜訊的影響而降低其辨識率。爲了提升系統的效能，許多研究語音辨識的學者歷年來不斷地研究語音的強健技術，期望能達到語音辨識系統的最佳化表現。在本論文中，我們主要是受時間序列結構正規化法觀念所啓發，進而探討並發展出更精確有效的調變頻譜正規化技術。我們提出了三種新方法，包含了等連波時間序列濾波器法、最小平方頻譜擬合法與強度頻譜內插法。這些方法將語音特徵時間序列的功率頻譜密度正規化至一參考的功率頻譜密度，以得到新的語音特徵參數，藉此降低雜訊對語音之影響，進而提升雜訊環境下的語音辨識精確度。同時，我們也將這些新方法結合其他特徵強健化的技術，發現這樣的結合能帶來更顯著之辨識率的提升。

Abstract

The performance of an automatic speech recognition system is often degraded due to the embedded noise in the processed speech signal. A variety of techniques have been proposed to deal with this problem, and one category of these techniques aims to normalize the temporal statistics of the speech features, which is the main direction of our proposed new approaches here.

In this thesis, we propose a series of noise robustness approaches, all of which attempt to normalize the modulation spectrum of speech features. They include equi-ripple temporal filtering (ERTF), least-squares spectrum fitting (LSSF) and magnitude spectrum interpolation (MSI). With these approaches, the mismatch between the modulation spectra for clean and noise-corrupted speech features is reduced, and thus the resulting new features are expected to be more noise-robust.

Recognition experiments implemented on Aurora-2 digit database show that the three new approaches effectively improve the recognition accuracy under a wide range of noise-corrupted environment. Moreover, it is also shown that they can be successfully

combined with some other noise robustness approaches, like CMVN and MVA, to achieve a more excellent recognition performance.

關鍵詞：語音辨識、調變頻譜正規化、強健性語音特徵參數

keyword: speech recognition, modulation spectrum, robust speech features

一、緒論

自動語音辨識系統(automatic speech recognition systems, ASR)，藉由多年來各方學者的研究發展，逐漸達到實際應用的階段，而爲人類生活帶來更多方便與幫助，雖然還不能達到一個完美的地步，但是這方面的技術仍一直不斷地進步當中。

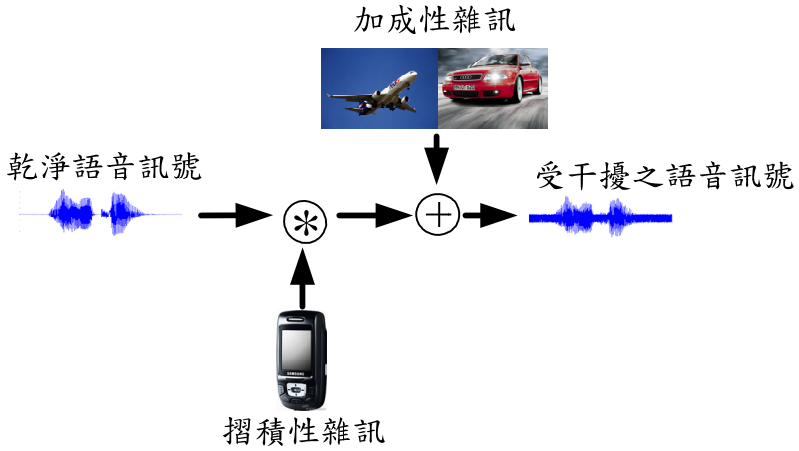
自動化語音辨認仍有許多相當具有挑戰性的研究課題，由於語音的變異性太多，例如每位語者說話的方式與口氣都不一樣、不同語言有不同的特性、語者當時說話的情緒、語者所處的環境是否有其他雜訊干擾等，這些變異對於語音辨識效果都有影響。在真實應用環境下，語音辨識系統所遇到的主要問題其中兩個，分別爲：

（一）語者不匹配(speaker mismatch)

語者不匹配的問題是因爲說話者先天條件（如口腔形狀）與後天習慣（如說話腔調）的差異所產生的變異性，因此當以特定語者所訓練出來的聲學模型來辨識不屬於此特定語者的語音時，辨識效果常會明顯下降，而要克服這一類問題的方法，通常是使用所謂的語者調適(speaker adaptation)技術。也就是將原本訓練出來的聲學模型調適成接近當下語者之語音特性的模型[1]，如此便可提高辨識率。

（二）環境不匹配(environment mismatch)

環境不匹配的問題是因爲語音辨識系統訓練環境與我們實驗或應用時的環境不同所致，其變異因子主要包含了加成性雜訊(additive noise)，如車站四周的雜訊、嘈雜街道的人聲或車聲等，及摺積性雜訊(convolutional noise)，如不同的有線或無線電話線路或麥克風所造成的通道效應等，語音辨識系統常會因這些雜訊的影響使辨識率降低。下圖一爲乾淨語音受雜訊干擾之示意圖。



圖一、乾淨語音受雜訊干擾之示意圖

在諸多降低雜訊影響、改進語音特徵的強健性技術中，有一大類的方法其目標是找出一強健語音特徵表示式(robust speech feature representation)，降低語音特徵對雜訊的敏感度，使雜訊產生的失真變小。此類著名的方法包括了倒頻譜平均消去法(cepstral mean subtraction, CMS)[2]、倒頻譜平均與變異數正規化法(cepstral mean and variance normalization, CMVN)[3]、相對頻譜法(RelAtive SpecTrAl, RASTA)[4]、倒頻譜平均與變異數正規化化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)[5]、倒頻譜增益正規化法 (cepstral gain normalization, CGN) [6]、資料導向時間序列濾波器法(data-driven temporal filter design)[7]等。以上這些方法皆是在語音特徵的時間序列域(temporal domain)作處理，根據語音訊號與雜訊在時間序列域上不同的特性，強調出語音的成分，而抑制雜訊的影響。近來，新加坡大學之李海洲博士研究團隊，新推出了一套時間序列濾波器設計的新方法，稱爲『時間序列結構正規化法』(temporal structure normalization, TSN)[8]，此方法的目的，在於將語音特徵序列之功率頻譜密度(power spectral density)正規化，使其輪廓逼近於一參考功率頻譜密度，此方法所得的時間序列濾波器，可以因應不同雜訊環境的語句特徵而加以調適，在其文獻[8]可知，當此新方法所得的時間序列濾波器作用於 CMVN 與 MVA 處理後的梅爾倒頻譜特徵參數時，在各種雜訊環境下所得到的語音辨識精確率都能有大幅改進。

雖然 TSN 法對語音特徵具有優異的強健化效果，且執行複雜度極低，但根據我們的觀察，此法仍然有幾點可以改進之處，首先，TSN 所得的初始濾波器係數是參考頻率響應之反傅利葉轉換求得，然後將這些係數會乘上一個漢寧窗(Hanning window)以減緩不當高頻成份的產生，此求取濾波器的方法未必是最佳化的，所得之濾波器係數其頻率響應可能與參考頻率響應之間的誤差較大。其次，在 TSN 法中，濾波器係數和被正規化爲 1，代表其直流增益爲一定值，此步驟使正規化後的特徵參數其功率頻譜密度並不會趨近參考功率頻譜密度，只在輪廓上大致相同。最後一點，則是 TSN 法皆是根據 MVN 或 MVA 處理後的梅爾倒頻譜特徵所設計，進而得到良好的效能，我們希望能探討 TSN 法單純作用於未經任何處理的梅爾倒頻譜特徵時，其效果是否也一樣明顯。

根據以上對 TSN 法的分析與觀察，在本論文中，我們提出了三種語音特徵時間序列之調變頻譜正規化(modulation spectrum normalization)的新方法，分別爲等漣波時間序列濾波器法(equi-ripple temporal filtering, ERTF)、最小平方頻譜擬合法(least-squares spectrum fitting, LSSF)與強度頻譜內插法(magnitude spectrum interpolation, MSI)，這三種方法之目的與 TSN 類似，皆爲了正規化語音特徵時間序列的功率頻譜密度，但我們會在後面章節的實驗結果發現，這三種方法之效能皆比 TSN 法來得好，且並不需要與 MVN 或 MVA 法結合，即可以十分有效地處理梅爾倒頻譜特徵因雜訊干擾所造成的失真。然而，當它們與 MVN 或 MVA 相結合時，也可以得到更佳的辨識精確率，此代表它們與 MVN 或 MVA 有良好的加成性。

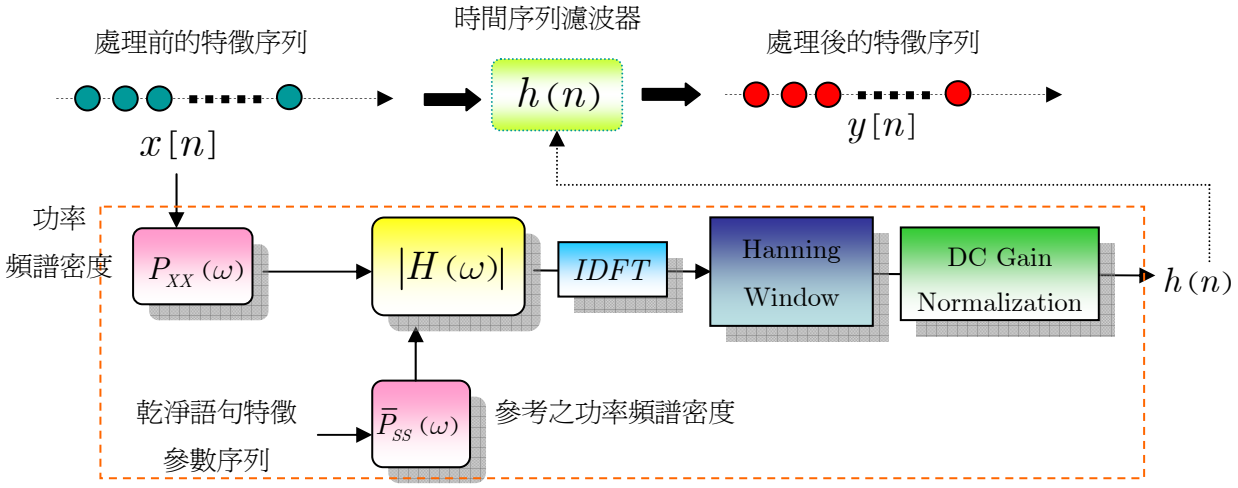
本論文其餘的章節概要如下：在第二章，我們將討論時間序列結構正規化法，包括其執行程序及初步效果，第三章爲本論文的重點，我們將在此章中針對時間序列結構正規化法作改進，而提出三種新的調變頻譜正規化法，並對其初步效果加以介紹。在第四章，我們將執行一系列的語音辨識實驗，來驗證所提之新方法足以有效提昇語音特徵在雜訊環境下的強健性，最後，第五章則爲結論及未來展望。

二、時間序列結構正規化法(temporal structure normalization, TSN)

（一）TSN 處理簡介

本章節主要介紹時間序列結構正規化法(temporal structure normalization, TSN)[8]，在下一章中，我們將以 TSN 法之觀念爲基礎，提出一系列的調變頻譜正規化的演算法。TSN 是屬於一種時間序列濾波器(temporal filter)設計之強健性語音技術，原始的 MFCC 語音特徵參數序列經過 CMVN 法[3]或 MVA 法[5]處理後，先求取其功率頻譜密度(power spectral density)，接著藉由此功率密度與事先定好的參考功率密度來決定一濾波器的強度響應(magnitude response)，此強度響應經反離散傅立葉轉換(inverse discrete Fourier transform, IDFT)、漢寧窗化(Hanning window)處理與直流增益正規化處理後，產生一組

濾波器係數，此即爲 TSN 法所求得的時間序列濾波器，將語音特徵序列通過此濾波器後，則預期可達到調變頻譜正規化的效果，而增加語音特徵之其強健性。圖二爲 TSN 法的處理程序示意圖：



圖二、TSN 法處理程序示意圖

在 TSN 法中，每一句訓練語料之某一維特徵序列 $\{s[n]\}$ 與測試語料同一維特徵序列 $\{x[n]\}$，先求取其功率頻譜密度，分別以 $\{P_{SS}(\omega_k)\}$ 與 $\{P_{XX}(\omega_k)\}$ 表示。接著將訓練語料所有句子同一維的功率頻譜密度作平均，所得即爲參考功率頻譜密度，如下所示：

$$\bar{P}_{SS}(\omega_k) = E\{P_{SS}(\omega_k)\}, \qquad \text{(式 2.1)}$$

在 TSN 法中所使用的濾波器，其初始的強度頻譜設定如下式所示：

$$|H(\omega_k)| = \sqrt{\bar{P}_{SS}(\omega_k)/P_{XX}(\omega_k)}, \qquad \text{(式 2.2)}$$

其上式明顯看出，當任一測試語料 $x[n]$ 通過上式之濾波器時，其原始功率頻譜密度 $P_{XX}(\omega_k)$ 會被正規化爲 $\bar{P}_{SS}(\omega_k)$。

爲了進一步求取濾波器的脈衝響應(impulse response)，上式(2.2)中的 $|H(j\omega_k)|$ 先經過反離散傅立葉轉換(inverse discrete Fourier transform, IDFT)，之後再乘上一個漢寧窗(Hanning window)，並將濾波器係數總和正規化爲 1，以達到直流增益正規化的目的。其數學表示式如以下數式所示：

1、反離散傅立葉轉換：

$$h[m] = \frac{1}{M}\sum_{k=0}^{M-1} H(j\omega_k)\, e^{-j\omega_k m}, \qquad 0 \le m \le M-1. \qquad \text{(式 2.3)}$$

2、漢寧窗化處理：

$$\hat{h}[m] = h[m]\cdot w[m], \qquad \text{(式 2.4)}$$

其中

$$w[m] = 0.5\left(1-\cos\left(2\pi\frac{m}{M-1}\right)\right), \qquad 0 \le m \le M-1.$$

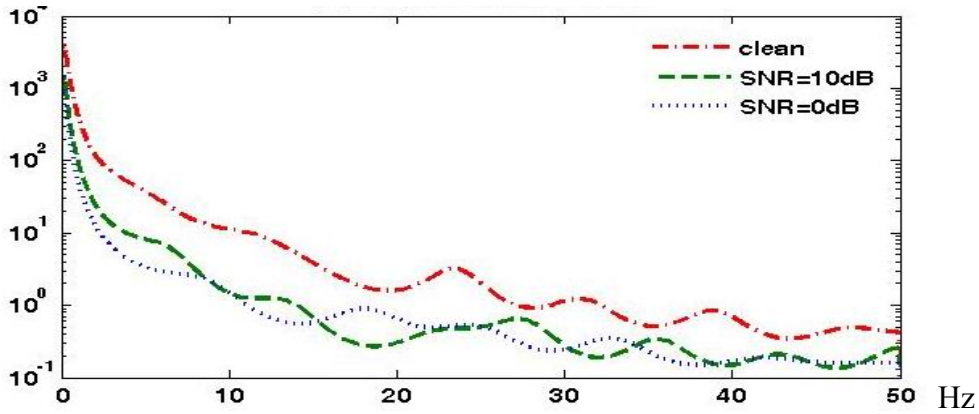
3、直流增益正規化：

$$\tilde{h}[m] = \frac{\hat{h}[m]}{\sum_{m'=0}^{M-1} h[m']}. \qquad \text{(式 2.5)}$$

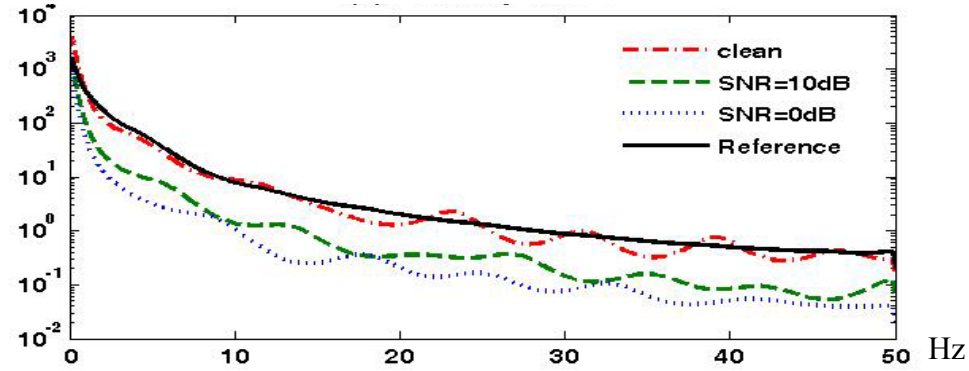
其中 M 爲濾波器長度。式(2.5)之 $\tilde{h}[m]$ 即爲 TSN 所求得之時間序列濾波器的脈衝響應。

（二）TSN 法效果相關討論

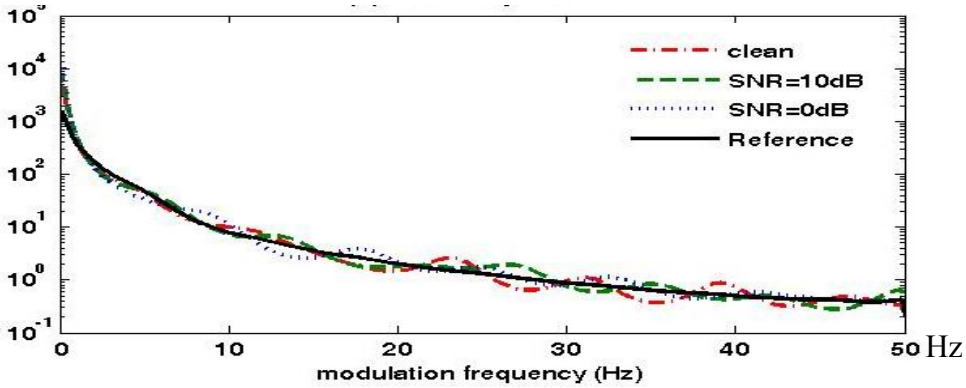
在 TSN 之文獻[8]中，所用的原始特徵參數皆爲經過 CMVN 法或 MVA 法所處理後之梅爾倒頻譜特徵參數(MFCC)。這裡我們特別將 TSN 法運用在未經處理之梅爾倒頻譜特徵參數上，觀察其改進效果。其中我們把原始 TSN 法命名爲 TSN-1，而把省略了直流增益正規化步驟的 TSN 法，命名爲 TSN-2。圖三爲原始第一維梅爾倒頻譜係數(c_1)序列的功率頻譜密度曲線圖，圖四爲原始 c_1 序列經 TSN-1 法處理後的功率頻譜密度曲線圖，圖五爲原始 c_1 序列經 TSN-2 法處理後的功率頻譜密度曲線圖。這些圖都使用了 AURORA 2 資料庫[9]裡的 MAH_4625A 語音檔，加入不同訊雜比的地下鐵雜訊。其中參考功率頻譜密度爲訓練語料庫之所有 c_1 序列之功率頻譜密度平均而得。



圖三、不同訊雜比之下，原始 c_1 序列之功率頻譜密度曲線圖



圖四、不同訊雜比之下，原始 c_1 序列經 TSN-1 處理後之功率頻譜密度曲線圖



圖五、不同訊雜比之下，原始 c_1 序列經 TSN-2 處理後之功率頻譜密度曲線圖

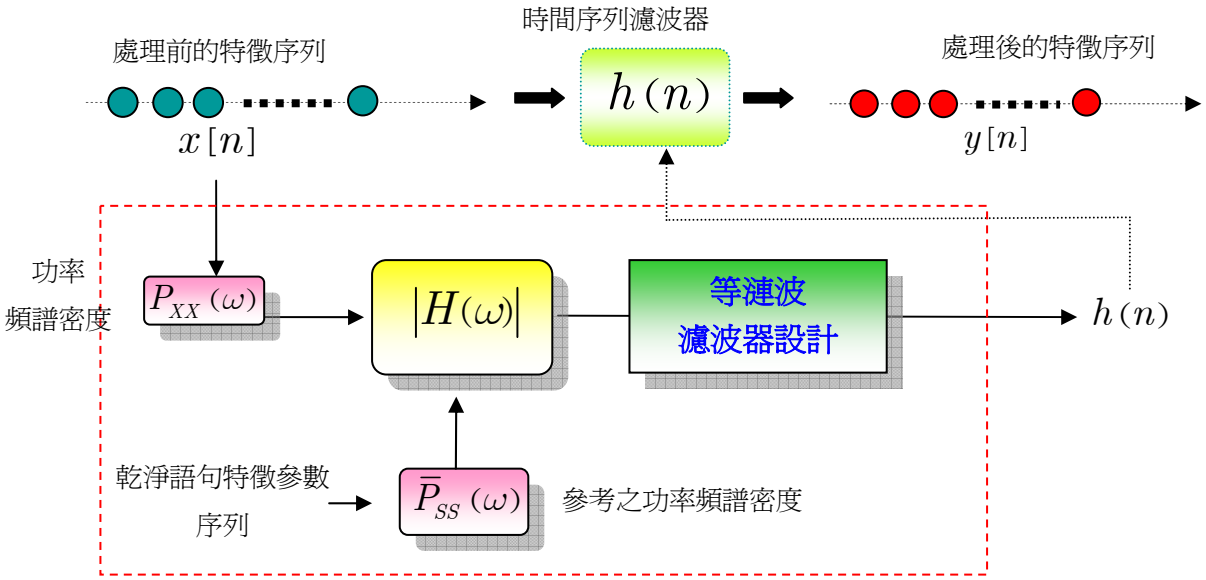
首先，從圖三可以明顯看出，雜訊會造成 c_1 特徵序列在功率頻譜密度上的失真，此是造成雜訊環境下，語音辨識精確率下降的原因之一。接著，我們從圖四觀察到原始 TSN 法（TSN-1）作用於原始 c_1 序列時，原本在圖三所看到之功率頻譜密度的失真並未被有效地改善，亦即其正規化效果並不理想，受到雜訊影響的 c_1 序列，當訊雜比(SNR)越低時，偏移參考功率頻譜密度的量越明顯。最後，從圖五可以看出，經過省卻直流增益正規化步驟的 TSN-2 法處理後，不同訊雜比下的 c_1 特徵序列其功率頻譜密度彼此十分接近，亦即 TSN-2 法可以有效正規化受雜訊干擾之原始 c_1 序列的功率頻譜密度，其降低失真的效能遠比 TSN-1 來的好。由此我們推論，原始 TSN 法中直流增益正規化的步驟並不是十分恰當，而其可能原因是，此步驟無法有效處理加成性雜訊對語音調變頻譜所造成的直流增益失真的效應。在下一章中，我們將提出一系列的方法，相較於 TSN 法而言，這些方法能更精確地正規化語音特徵的功率頻譜密度。

三、調變頻譜正規化的新方法

在第一章與第二章中，我們探討到時間序列結構正規化法(temporal structure normalization, TSN)可能有些可以改進的地方，同時藉由 TSN 法之觀念啓發，因此在本章節中，我們提出一系列的調變頻譜正規化的新方法。這些新方法分別爲等漣波時間序列濾波器法(equi-ripple temporal filtering, ERTF)、最小平方頻譜擬合法(least-squares spectrum fitting, LSSF)與強度頻譜內插法(magnitude spectrum interpolation, MSI)，這些方法分別在本章的前三節中作介紹，而最後第四小節則爲這些方法簡要的效能評估與特性討論。

（一）等漣波時間序列濾波器法(equi-ripple temporal filtering, ERTF)

在等漣波時間序列濾波器法(ERTF)中，我們使用等漣波濾波器設計法 (equi-ripple filter design)[10]來設計濾波器的脈衝響應，以取代原始 TSN 法中，反傅立葉轉換與窗化處理的步驟，同時，我們也挪去原始 TSN 法中正規化濾波器直流增益的步驟。圖六爲等漣波時間序列濾波器法處理程序圖，在 ERTF 法中，我們所提出的兩更新步驟之目的正是求取更精確的濾波器係數，以趨近正規化特徵序列之調變頻譜強度成份的目標。



圖六、等漣波時間序列濾波器法處理程序圖

ERTF 法中所使用的 $P_{XX}(\omega_k)$、$\bar{P}_{SS}(\omega_k)$ 和 $H(\omega_k)$ 求取方式都和前一章所述之原始

TSN 法相同，但是濾波器係數$\{h[n]\}$是以等漣波濾波器設計法[10]求得，此方法是利用所謂的最小化最大誤差準則(minimax criterion)來求取一最佳的濾波器頻率響應，如下式所示：

$$\breve{H}(\omega_k) = \arg\min_{H(\omega_k)}\left(\max_{\omega} W(\omega_k)\left|H(\omega_k) - D(\omega_k)\right|\right) \ , \qquad \text{(式 3.1)}$$

其中$W(\omega_k)$爲權重值，$\breve{H}(\omega_k)$爲最佳化濾波器之頻率響應，$D(\omega_k)$爲參考的頻率響應，$D(\omega_k)$可表示如下式：

$$D(\omega_k) = \sqrt{\frac{\bar{P}_{SS}(\omega_k)}{P_{XX}(\omega_k)}} \qquad \text{(式 3.2)}$$

由此法得到的濾波器係數$\{h[n]\}$，會自動符合前後對稱(symmetric)的性質，因此其相位響應是線性的(linear phase)[10]，並不會使原始特徵序列的調變頻譜產生相位失真的情形，同時，因爲濾波器本身是根據最佳化準則設計，所以我們預期它會比 TSN 法所得之濾波器效果來的好。

（二）最小平方頻譜擬合法(least-squares spectrum fitting, LSSF)

在這方法裡，我們針對每一個待正規化的N點特徵時間序列$\{x[n]|0 \le n \le N-1\}$先定義一$2P$點的參考調變頻譜，作爲此特徵序列的調變頻譜正規化的目標：

$$\hat{Y}(\omega_k) = \left|Y(\omega_k)\right|\exp\left(j\theta_X(\omega_k)\right), \quad 0 \le k \le 2P-1, \qquad \text{(式 3.3)}$$

其中的強度成份$\left|Y(\omega_k)\right|$以下式表示：

$$\left|Y(\omega_k)\right| = \left|X(\omega_k)\right|\sqrt{\bar{P}_{SS}(\omega_k)/P_{XX}(\omega_k)} \qquad \text{(式 3.4)}$$

其中，$\bar{P}_{SS}(\omega_k)$與如前章的式(2.1)中所定義，即$\bar{P}_{SS}(\omega_k)$爲所有訓練語料特徵與$\{x[n]\}$同一維序列的功率頻譜密度平均而得，$P_{XX}(\omega_k)$爲原始特徵序列$\{x[n]\}$的功率頻譜密度。而強度成份$\left|X(\omega_k)\right|$和相角成份$\theta_X(\omega_k)$爲$\{x[n]\}$經過$2P$點之離散傅立葉轉換(discrete Fourier transform, DFT)所得到。值得注意的是，特徵長度N會隨著不同的語句而不同，但是這裡的 DFT 取樣點數$2P$則設爲一固定值，也就是參考調變頻譜的長度對於每一個語句都是相同的。

由式(3.3)與式(3.4)可知，我們希望每一個更新後的特徵序列，其調變頻譜的強度成份能趨於一致，而相位成份則由原始的特徵序列$\{x[n]\}$而來。接下來，我們利用最小平方化(least-squares)[10]的最佳化準則求取一新的特徵參數序列，使新的特徵序列$\{y[n]\}$的調變頻譜逼近如式(3.3)的參考調變頻譜，如下式所示：

$$y[n] = \min_{\{\tilde{y}[n]|0\le n\le N-1\}} \sum_{k=0}^{2P-1}\left|\sum_{n=0}^{N-1}\tilde{y}[n]e^{-j\frac{2\pi nk}{2P}} - \hat{Y}(\omega_k)\right|^2, \qquad (2P \ge N) \qquad \text{(式 3.5)}$$

其中$2P$爲 DFT 取樣點數，N爲此特徵序列的點數。

藉由矩陣與向量表示法，我們可將式(3.5)改寫爲下式：

$$\mathbf{y} = \min_{\tilde{\mathbf{y}}}\left\|W\tilde{\mathbf{y}} - \hat{\mathbf{Y}}\right\|^2 \qquad \text{(式 3.6)}$$

其中W是$2P \times N$的矩陣，其第(m,n)項如下所示：

$$W_{mn} = \exp\left(-j\frac{2\pi mn}{2P}\right),$$

而$\mathbf{y}$、$\tilde{\mathbf{y}}$與$\hat{\mathbf{Y}}$則定義爲：

$$\mathbf{y} = \begin{bmatrix} y[0] & y[1] & \cdots & y[n-1] \end{bmatrix}^T,$$

$$\hat{\mathbf{y}} = \left[\hat{y}[0] \quad \hat{y}[1] \quad \cdots \quad \hat{y}[N-1]\right]^T,$$
$$\hat{\mathbf{Y}} = \left[\hat{Y}(\omega_0) \quad \hat{Y}(\omega_1) \quad \cdots \quad \hat{Y}(\omega_{2P-1})\right]^T,$$

由於$\hat{\mathbf{y}}$為實數向量，故式(3.6)可改寫為：

$$\begin{aligned}\mathbf{y} &= \min_{\tilde{\mathbf{y}}} \left\| \left(W_R\hat{\mathbf{y}} - \hat{\mathbf{Y}}_R\right) + j\left(W_I\hat{\mathbf{y}} - \hat{\mathbf{Y}}_I\right)\right\|^2 \\ &= \min_{\tilde{\mathbf{y}}} \left(\left\|W_R\hat{\mathbf{y}} - \hat{\mathbf{Y}}_R\right\|^2 + \left\|W_I\hat{\mathbf{y}} - \hat{\mathbf{Y}}_I\right\|^2 \right)\end{aligned} \qquad \text{(式 3.7)}$$

其中矩陣W_R與W_I分別爲矩陣W的實部與虛部，而向量$\hat{\mathbf{Y}}_R$與$\hat{\mathbf{Y}}_I$則分別爲向量$\hat{\mathbf{Y}}$的實部與虛部。

由式(3.7)明顯看出，此爲一典型的最小平方化(least-squares)的求解問題，故其精確的封閉解(closed-form solution)可由下式表示：

$$\mathbf{y} = \left(W_R^T W_R + W_I^T W_I\right)^{-1}\left(W_R^T\hat{\mathbf{Y}}_R + W_I^T\hat{\mathbf{Y}}_I\right) \qquad \text{(式3.8)}$$

所以，式(3.8)中的$\mathbf{y}$即爲 LSSF 法所求得之新特徵參數序列$\{y[n]\}$，其$2P$點之 DFT 和式(3.3)的參考調變頻譜之間具有最小平方誤差的良好性質。

（三）強度頻譜內插法(magnitude spectrum interpolation, MSI)

在此方法中，我們爲每一個待正規化的N點特徵序列$\{x[n] | 0 \le n \le N-1\}$，定義了一個$N$點的參考調變頻譜，作爲此特徵序列之調變頻譜正規化的目標，如下式所示：

$$\tilde{Y}(\omega_{k'}) = \left|\tilde{Y}(\omega_{k'})\right| \exp\left(j\theta_X(\omega_{k'})\right), \quad 0 \le k' \le N-1 \qquad \text{(式 3.9)}$$

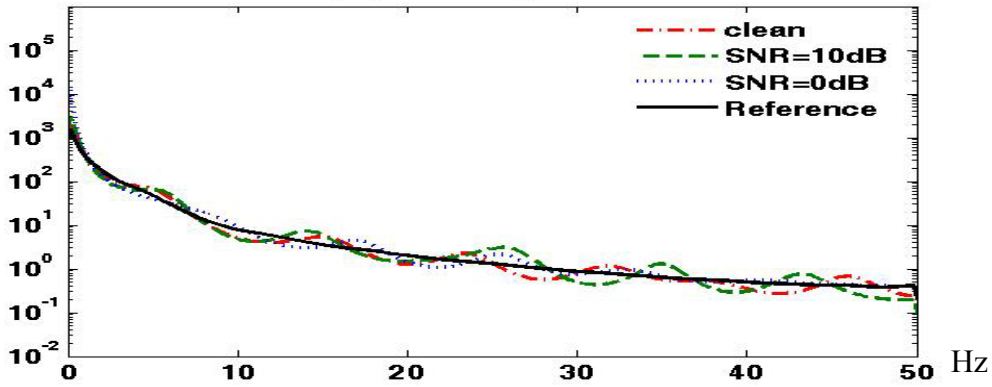
其中相位成份$\theta_X(\omega_{k'})$爲$x[n]$取N點的 DFT 所得。MSI 法跟前節之 LSSF 法的最大不同之處，在於此時我們是使用一個跟原始特徵序列長度相同的參考調變頻譜，而由於不同語句的特徵序列，其點數N也隨之不同，我們不能如前面的 LSSF 法中，直接拿$2P$點的參考功率頻譜密度$\left\{\overline{P}_{SS}(\omega_k) \middle| 0 \le k \le 2P-1\right\}$(如式(2.1)所示)來求取式(3.9)中的$N$點頻譜強度$\left|\tilde{Y}(\omega_{k'})\right|$。然而，由於原始$2P$點的參考頻譜其涵蓋頻率範圍與欲求的$\left|\tilde{Y}(\omega_{k'})\right|$頻率範圍相同，在這裡，我們使用線性內插(linear interpolation)[10]的方法，藉由式(3.4)中所示的之$2P$點的$\left\{\left|Y(\omega_k)\right| 0 \le k \le 2P-1\right\}$來求取式(3.9)中$N$點的$\left\{\left|\tilde{Y}(\omega_{k'})\right| 0 \le k' \le N-1\right\}$之近似值。但是式(3.9)的$\left\{\tilde{Y}(\omega_{k'})\right\}$爲一實數序列之離散傅立葉轉換，其強度成份$\left\{\left|\tilde{Y}(\omega_{k'})\right|\right\}$必須符合左右對稱的性質，即$\left|\tilde{Y}(\omega_{k'})\right| = \left|\tilde{Y}(\omega_{N-k'})\right|$，因此我們先利用$\left\{\left|Y(\omega_k)\right|\right\}$的左半部執行內插法，求取$\left\{\left|\tilde{Y}(\omega_{k'})\right|\right\}$的左半部$\left\{\left|\tilde{Y}(\omega_{k'})\right| 0 \le k' \le \left\lfloor\frac{N}{2}\right\rfloor\right\}$，再利用左右對稱的性質，求取$\left\{\left|\tilde{Y}(\omega_{k'})\right|\right\}$右半部$\left\{\left|\tilde{Y}(\omega_{k'})\right| N-1-\left\lfloor\frac{N}{2}\right\rfloor \le k' \le N-1\right\}$。在得到$\left\{\left|\tilde{Y}(\omega_{k'})\right| 0 \le k' \le N-1\right\}$後，我們就可以直接對式(3.9)的$\left\{\tilde{Y}(\omega_{k'})\right\}$做$N$點的反離散傅立葉轉換(inverse discrete Fourier transform, IDFT)，以求得新的特徵序列$\{y[n]\}$，如下式所示：

$$y[n] = \frac{1}{N}\sum_{k'=0}^{N-1}\tilde{Y}(\omega_{k'})e^{j\frac{2\pi nk'}{N}}, \quad 0 \le n \le N-1. \qquad \text{式(3.10)}$$

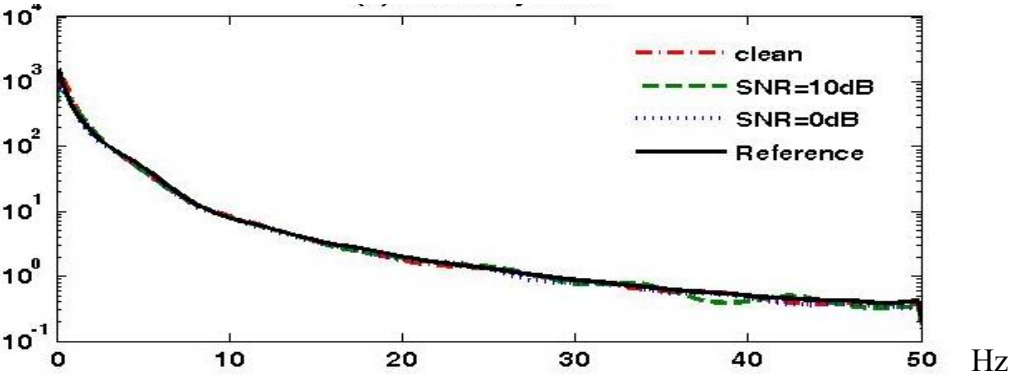
以上的方法，即稱爲強度頻譜內插法(magnitude spectrum interpolation, MSI)。

（四）調變頻譜正規化之新方法的效果討論

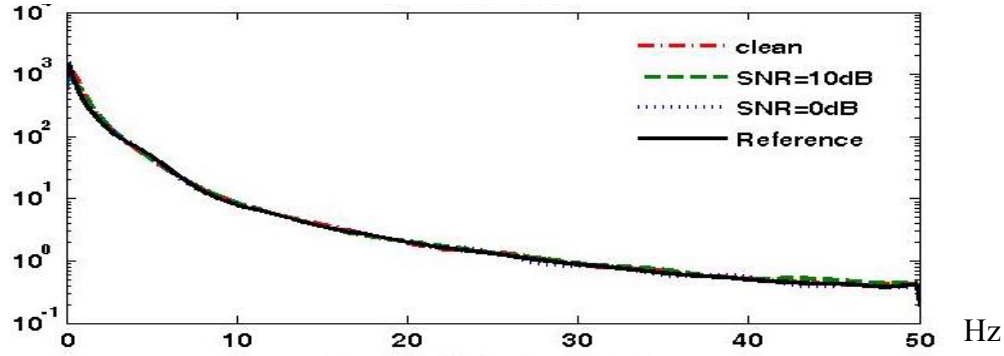
這小節將簡單展示本章節所提出的三種新方法對原始 MFCC 特徵序列之調變頻譜正規化的效果，圖七、圖八與圖九分別爲原始第一維梅爾倒頻譜係數(c_1)序列分別經 ERTF 法、LSSF 法與 MSI 法處理後的功率頻譜密度曲線圖。與前一章的圖三、圖四和圖五相同，這裡我們所使用的是 AURORA 2 資料庫[9]裡的 MAH_4625A 語音檔，然後加入不同訊雜比(SNR)的地下鐵(subway)雜訊。



圖七、不同訊雜比之下，原始 c_1 序列經 ERTF 法處理後之功率頻譜密度曲線圖



圖八、不同訊雜比之下，原始 c_1 序列經 LSSF 法處理後之功率頻譜密度曲線圖



圖九、不同訊雜比之下，原始 c_1 序列經 MSI 法處理後之功率頻譜密度曲線圖

將圖七、圖八、與圖九配合前一章之圖三、圖四與圖五相比較，我們有以下兩點討論：

① 由於 ERTF 法和 TSN 法同樣是設計一時間序列濾波器，作用於特徵參數序列上，我們先比較這兩種方法的效能。從圖七中可看出 ERTF 法能同時使得乾淨語音與受雜訊干擾的語音的功率頻譜密度曲線，逼近參考的功率頻譜密度曲線，有效降低圖三所顯示之不同訊雜比下特徵序列之功率頻譜密度的失真，相較於圖四所顯示之原始 TSN 法的效果有明顯改善，且與圖五之 TSN-2 法的效果十分接近，此代表我們使用等漣波濾波器設計法 (equi-ripple filter design)來設計時間序列濾波器，可以有效地正規化不同

雜訊比下的語音特徵之調變頻譜。

② LSSF 法和 MSI 法都是直接在特徵的調變頻譜域(modulation spectral domain)上正規化其強度成份，從圖八和圖九可看出這兩種方法與 ERTF 法類似，能將受雜訊干擾的語音之功率頻譜密度曲線，逼近參考的功率頻譜密度曲線，使這些曲線之間的差異明顯較低，代表了這兩個調變頻譜強度正規化法也能有效地強健語音特徵。其中，MSI 法是三個方法中計算複雜度最低的技術，因此有更大的應用價值。

從上述三個圖中，可看出我們所提的三個新方法都能有效地降低雜訊所造成之語音特徵在調變頻譜上失真的現象，我們在下個章節，將會以辨識實驗數據證實這些方法的效能。

四、調變頻譜正規化法與各種特徵時間序列正規化技術法之辨識實驗結果與討論

本章節主要是將我們提出的三種調變頻譜正規化法：等漣波時間序列濾波器(equi-ripple temporal filtering, ERTF)、最小平方頻譜擬合法(least-squares spectrum fitting, LSSF)和強度頻譜內插法(magnitude spectrum interpolation, MSI)運用於雜訊環境下的語音辨識，藉此觀察分析其結果，同時我們也會將它們與其他特徵時間序列正規化法的效果作比較。最後，我們嘗試將這些新方法與其他方法互相結合，來觀察這樣的結合是否能來更進一步的效能提升。

（一）實驗環境與實驗架構設定

本論文中所採用的語音資料庫爲歐洲電信標準協會(European Telecommunication Standard Institute, ETSI)所發行之語料庫：AURORA 2.0[9]，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，語音本身並加上各種加成性雜訊與通道效應的干擾。加成性雜訊共有八種，分別爲地下鐵、人聲、汽車，展覽會館、餐廳、街道、飛機場和火車站雜訊等，通道效應則有兩種，分別爲 G712 與 MIRS[11]。雜訊含量的大小包含了乾淨無雜訊的狀態，以及六種不同訊雜比(signal-to-noise ratio, SNR)狀態，分別是 20dB、15dB、10dB、5dB、0dB 與-5dB，因此我們可以觀察不同的雜訊環境對於語音辨識的影響。因雜訊特性的不同，測試環境可分爲 Set A、Set B 與 Set C 三組[9]。

聲學模型是執行隱藏式馬可夫模型工具(hidden Markov model tool kit, HTK)[12]訓練所得，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。

（二）調變頻譜正規化法作用於梅爾倒頻譜特徵參數之實驗結果

本章節所有實驗所使用的語音特徵爲梅爾倒頻譜係數(mel-frequency cepstral coefficients, MFCC)，我們採用的 MFCC 特徵參數爲13維(c0~c12)，加上其一階差量(delta)和二階差量(delta-delta)，總共爲39維特徵參數。基本實驗(baseline experiment)是以原始 MFCC 特徵參數作爲訓練與測試，TSN-1法爲第二章中所介紹之原始 TSN 法，而 TSN-2 法則是將原始 TSN 法中直流增益正規化步驟省略所得的修正法，TSN-1與 TSN-2所得之時間序列濾波器長度皆設爲21，此值是直接參考 TSN 法的文獻[8]而來。ERTF 法所得的時間序列濾波器長度爲21，而 LSSF 與 MSI 法所用的 DFT 點數$2P$(如式(3.3)所示)則固定爲1024。下表一中，我們綜合了 TSN-1、TSN-2、ERTF、LSSF、MSI，及著名的特徵正規化技術 CMVN[3]和 MVA[5]，其各別作用於原始 MFCC 特徵參數所得的平均辨識率（20dB、15dB、10dB、5dB 與0dB 五種訊雜比下的辨識率平均），其中 AR 與 RR 分別爲相較於基本實驗結果之絕對錯誤降低率(absolute error rate reduction)和相對錯誤降低率(relative error rate reduction)。

由表一的數據，我們可看出以下幾點現象：

表一、各種特徵序列處理技術之辨識率(%)

Method	Set A	Set B	Set C	average	AR	RR
Baseline	72.46	68.31	78.82	73.20	-	-
TSN-1	73.61	70.44	77.19	73.75	0.55	2.05
TSN-2	80.29	82.36	75.82	79.49	6.29	23.47
ERTF	85.45	86.92	85.34	85.90	12.70	47.39
LSSF	84.37	86.21	84.72	85.10	11.90	44.40
MSI	83.61	85.36	84.28	84.42	11.22	41.87
CMVN	85.03	85.56	85.60	85.40	12.20	45.52
CMVN+ARMA(MVA)	88.12	88.81	88.50	88.48	15.28	57.01

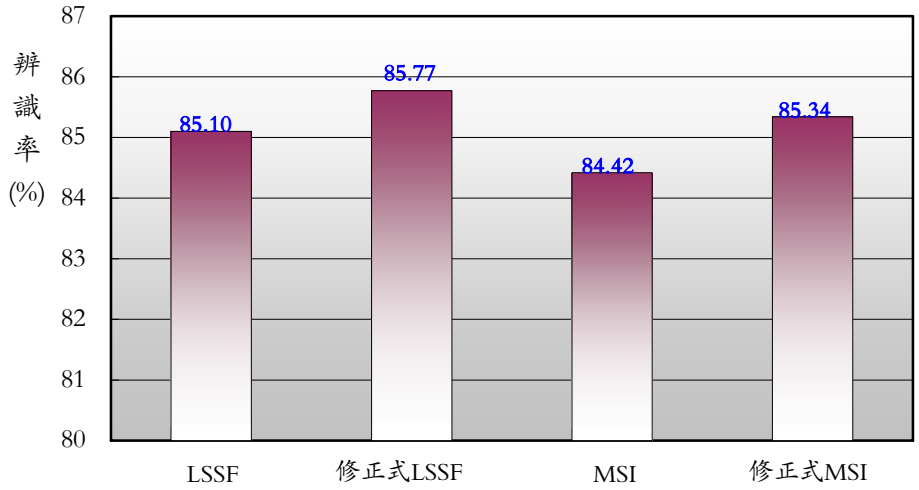
①原始 TSN 法(TSN-1)對 MFCC 特徵在雜訊環境下的辨識率的改進並不是很明顯，只進步0.55%，然而 TSN-2法帶來十分明顯的辨識率提升(Set C 除外)，在 Set A 和 Set B 環境下，平均辨識率相對於 TSN-1而言分別改進了8%與14%左右。如此看出，藉由省略直流增益正規化的步驟，TSN-2 比 TSN-1 具有更佳的特徵調變頻譜正規化的效果，這也呼應了在第二章的圖三，原始 TSN 法無法有效降低各雜訊環境下，原始語音 MFCC 特徵時間序列之功率頻譜密度曲線的不匹配現象。

②ERTF、LSSF 與 MSI 法三種新方法在各種不同的雜訊環境下皆能明顯提升辨識率，對 Set A 環境而言，它們分別使辨識率提升了 12.99%、11.91%與 11.15%，對 Set B 環境而言，辨識率分別提升了 18.61%、17.90%與 17.05%，在 Set C 環境下，辨識率分別提升了 6.52%、5.90%與 5.46%。這三種方法中，又以 ERTF 法的表現最好，明顯優於 LSSF 法與 MSI 法，但它們所能達到的相對錯誤降低率都高達 40%以上，明顯優於 TSN-1 法與 TSN-2 法。另外，值得一提的是，TSN-2 法在 Set C 中的效果比 TSN-1 與基礎實驗差，但 ERTF、LSSF 與 MSI 法卻未有這樣的不良結果。

③ 兩種目前廣爲人用的特徵正規化技術：CMVN 法與 MVA 法，對辨識率的提升都十分明顯，CMVN 的效能與我們所提的三種新方法大致相同，但結合了 CMVN 與 ARMA 濾波處理的 MVA 法其效能又比 CMVN 法來的好，基於這樣的觀察，在下兩小節中，我們將試著把各種調變頻譜正規化法與 CMVN 法或 MVA 法加以整合，探討是否能帶來辨識率上更顯著的進步。

當我們使用 LSSF 法與 MSI 法時，我們會將原始爲 N 點的特徵序列轉換成 $2P$ 點之功率頻譜密度或離散頻譜，然而由於通常 $2P > N$ ，我們會以補零的方式先將原始的 N 點的特徵序列變長爲 $2P$ 點，意即多補了 $2P - N$ 個零點，這樣的作法容易產生非零值的點與零值的點之間訊號值不連續的情形，而引進了不必要的高頻成份，這效應類似於直接於一訊號加上矩形窗所造成頻譜遺漏(leakage)[10]的缺點，因此，我們這裡在 LSSF 與 MSI 法之補零的程序前，先將原始的 N 點的特徵序列乘上一漢寧窗(Hanning window)[10]，來降低上述可能的不良效應，觀察這樣的操作是否可進一步提升 LSSF 法與 MSI 法的效果，我們稱這樣修改結果分別爲修正式 LSSF 法(modified LSSF)與修正式 MSI 法(modified MSI)。

圖十爲原始與修正式 LSSF 與 MSI 作用於原始 MFCC 特徵之平均辨識率長條圖。由此圖中可以看出修正式 LSSF 法相較於原始 LSSF 法而言，平均辨識率有 0.67%的提升，而修正式 MSI 相較於原始 MSI 而言，在平均辨識率上有 0.92%的提升。由此我們驗證了，在修正法中所作的窗化處理確實能有效改進 LSSF 法與 MSI 法的效能。



圖十、原始和修正式 LSSF 與 MSI 作用於原始 MFCC 特徵之平均辨識率

（三）調變頻譜正規化法結合倒頻譜平均與變異數正規化法之實驗結果

前面提到，倒頻譜平均與變異數正規化法(cepstral mean and variance normalization, CMVN)[3]對雜訊環境下的語音辨識率有明顯的改進，因此這裡我們嘗試將各種調變頻譜正規化法與 CMVN 法作結合，意即原始 MFCC 特徵先經過 CMVN 法處理後，再以各種調變頻譜正規化法分別作處理。以下我們測試這樣的結合是否有加成性的效果。在表二中，我們整理了 CMVN 法分別結合 TSN-1、TSN-2、ERTF、LSSF、MSI 及 ARMA 濾波法(MVA)[5]各方法所得的平均辨識率，其中 AR 與 RR 分別爲相較於單一 CMVN 結果之絕對錯誤降低率(absolute error rate reduction)和相對錯誤降低率(relative error rate reduction)。

表二、各調變頻譜處理法作用於 CMVN 處理後之 MFCC 特徵所得之辨識率(%)

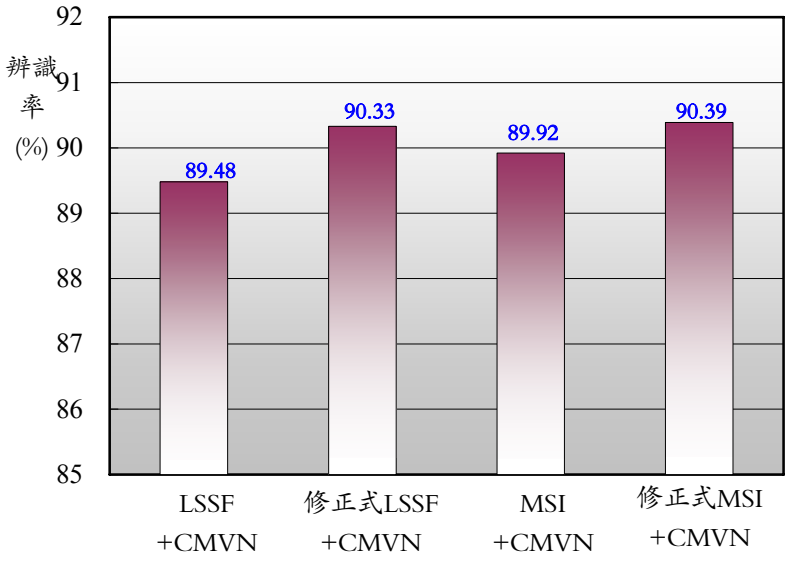
Method	Set A	Set B	Set C	average	AR	RR
CMVN	85.03	85.56	85.60	85.40	—	—
CMVN+TSN-1	89.42	90.03	89.03	89.49	4.10	28.05
CMVN+TSN-2	89.59	90.36	89.34	89.76	4.36	29.90
CMVN+ERTF	89.61	90.67	89.28	89.85	4.45	30.52
CMVN+LSSF	89.12	90.17	89.16	89.48	4.09	27.98
CMVN+MSI	89.59	90.56	89.60	89.92	4.52	30.95
CMVN+ARMA(MVA)	88.12	88.81	88.50	88.48	3.08	21.09

由表二的數據，我們可看出以下幾點現象：

① TSN-1 法作用於 CMVN 處理過的 MFCC 特徵，其改進辨識率的效能十分顯著，相較於單一 CMVN 法而言，在 Set A、Set B 與 Set C 環境下分別具有 4.39%、4.47%與 3.43%的辨識率改善，此結果十分吻合在 TSN 法的原始文獻[8]裡之結果，相較於表一所呈現之 TSN-1 並未明顯改善受雜訊影響之原始 MFCC 特徵的現象，在這裡，TSN-1 法能有明顯改進之效能的原因可能在於，CMVN 法已事先有效地降低原始 MFCC 特徵受雜訊影響所造成之調變頻譜上下偏移的失真，因此 TSN-1 能單純處理調變頻譜正規化的部份，而帶來辨識率的改善。另外，我們也發現到，TSN-1 和 TSN-2 所得結果之間的差距變得較小，但 TSN-2 的整體辨識率還是比 TSN-1 來的好，再一次驗證原 TSN 法中直

流增益正規化的步驟應該是不必要的。
②當我們所提出的新方法 ERTF、LSSF 與 MSI 法作用於 CMVN 處理後之 MFCC 特徵時，相較於單一 CMVN 所得的辨識率而言，皆帶來十分顯著的改善，例如在 Set A 環境下，這三種方法分別具有 4.58%、4.09%與 4.56%的辨識率提昇，此顯示了這三種新方法與 CMVN 有良好的加成性。而三個新方法中，ERTF 和 MSI 法表現的都比 TSN-1 或 TSN-2 法更好，雖然 LSSF 法表現稍不如預期，但是可能原因在於前一節所討論到的，原始 LSSF 法和 MSI 法可能會產生頻譜遺漏(leakage)現象之類的不良效應，因此在後面，我們將以修正式 LSSF 法與 MSI 法結合 CMVN 法來探討其可能的改進效果。
③在之前的表一數據顯示，當作用於原始 MFCC 特徵時，ERTF 表現比 LSSF 和 MSI 法更好。但是在這裡我們發現當這些方法和 CMVN 法結合時，其效果變得十分接近，這也意味著 CMVN 法確實已對原始 MFCC 特徵作了十分有效的強健性處理，而使後續的改進技術，其進步的空間相對變小。

如之前所提到的，原始 LSSF 法和 MSI 法可能有頻譜遺漏(leakage)的缺點，因此我們這裡使用之前所述的修正式 LSSF 與 MSI 法，作用於 CMVN 處理後的 MFCC 特徵，即在此兩方法補零的程序前先將原始 N 點的 CMVN 法處理後之 MFCC 特徵序列乘上一漢寧窗(Hanning window)，觀察這樣的操作是否可進一步提升原始 LSSF 法與 MSI 法結合 CMVN 法的效果。
圖十一爲原始與修正式 LSSF 與 MSI 作用於 CMVN 法處理後 MFCC 特徵之平均辨識率長條圖。由此圖可以看出，在結合 CMVN 法後，修正式 LSSF 法相較於原始 LSSF 法而言，有 0.85%之平均辨識率的提升，同樣地，修正式 MSI 法相對於原始 MSI 法而言，有 0.47%之平均辨識率的提升，二者平均辨識率皆超過 90%。此外，當與表二的數據比較，我們看到這兩種修正式方法結合 CMVN 法後在總平均辨識率上皆明顯優於與 CMVN 法結合的 TSN-1 法(89.49%)與 TSN-2 法(89.76%)，以上結果都顯示了這樣的修正確實能有效改進原方法的缺點，而提升其效能。



圖十一、原始和修正式 LSSF 與 MSI 作用於 CMVN 法處理後 MFCC 特徵之平均辨識率

（四）調變頻譜正規化法結合倒頻譜平均與變異數正規化結合自動回歸動態平均濾波器法之實驗結果

前面提到，倒頻譜平均與變異數正規化結合自動回歸動態平均濾波器法(MVA)[5]

能夠對雜訊環境下的語音特徵有明顯的強健化效果，而帶來十分顯著的辨識率提升，且其效能優於 CMVN，因此在這裡，我們將各種調變頻譜正規化法與 MVA 法作結合，也就是把這些正規化法作用於經 MVA 法處理後之 MFCC 特徵上，以檢視這些正規化法與 MVA 法是否有加成性。實驗中我們設定 MVA 法中的 ARMA 濾波器階數爲 2(參照[5])。在下表三中，我們列出了 MVA 法分別結合 TSN-1、TSN-2、ERTF、LSSF 與 MSI 各方法所得的平均辨識率，其中 AR 與 RR 分別爲相較於單一 MVA 法之結果的絕對錯誤降低率(absolute error rate reduction)和相對錯誤降低率(relative error rate reduction)。

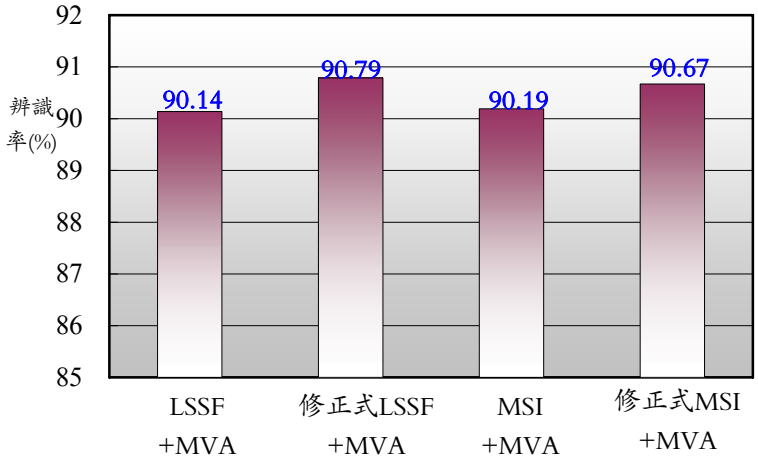
表三、各調變頻譜處理法作用於 MVA 處理後之 MFCC 特徵所得之辨識率(%)

Method	Set A	Set B	Set C	average	AR	RR
MVA	88.12	88.81	88.50	88.48	—	—
MVA+TSN-1	89.58	90.19	89.74	89.84	1.36	11.80
MVA+TSN-2	89.81	90.34	89.84	99.00	1.52	13.19
MVA+ERTF	89.75	90.81	89.64	90.07	1.59	13.80
MVA+LSSF	89.63	90.87	89.94	90.14	1.67	14.49
MVA+MSI	89.71	90.91	89.94	90.19	1.71	14.84

由表三可看出，TSN-1 在結合 MVA 後，其效能有明顯的提昇，而 TSN-1 和 TSN-2 之間的差異雖然不明顯，但是省略直流增益正規化步驟的 TSN-2 法仍然表現比較好，相對於單一 MVA 法的結果而言，結合了 MVA 法之 TSN-1 在總平均辨識率上提升 1.36%，而 TSN-2 提升了 1.52%。此外，結合 MVA 法之後，我們提出的 ERTF、LSSF 與 MSI 三個方法仍優於 TSN-1 與 TSN-2，而其中以 MSI 法最好，在辨識率上提升 1.71%，其次爲 LSSF 法，提升了 1.67%， ERTF 法則提升了 1.59%。儘管如此，我們可明顯看出，這些方法在結合 MVA 法後，所帶來的辨識率提升程度相對而言都已十分接近。

如同前節所描述之原始 LSSF 法與 MSI 法的可能缺點，在這裡，我們同樣地測試修正式 LSSF 法與 MSI 法結合 MVA 法的效果，即在原始 LSSF 法或 MSI 法之補零的程序前先將原始 N 點之 MVA 法處理後之 MFCC 特徵序列乘上一漢寧窗(Hanning window)，觀察這樣的操作能否帶來進步。

圖十二爲原始與修正式 LSSF 與 MSI 作用於 MVA 法處理後 MFCC 特徵之平均辨識率長條圖。由此圖可以看出，在結合 MVA 法的前提下，修正式 LSSF 法相較於原始 LSSF 法而言，有 0.65%平均辨識率的提升，而修正式 MSI 法相對於原始 MSI 法而言，有 0.48%平均辨識率的提升，因此，我們驗證了兩種修正式方法都能使原始方法進一步提升效能。



圖十二、原始和修正式 LSSF 與 MSI 法作用於 MVA 法處理後 MFCC 特徵之平均辨識率

五、結論

在作用於原始 MFCC 特徵時，我們發現，原始 TSN 法(TSN-1)的直流增益正規化步驟是造成其效果不彰的原因之一，挪去此步驟所得之 TSN-2 法即可有十分顯著的表現，而我們提出的三種新方法，相較於 TSN-1 與 TSN-2，都能有更佳的效果，而其中又以 ERTF 法之表現最好，由於 ERTF 與 TSN-2 只有在設計時間序列濾波器的程序上有差別，這表示我們 ERTF 設計出來的濾波器，比起 TSN-2 法的濾波器更精確地對特徵之調變頻譜作正規化。而當我們將這些方法作用於 CMVN 法或 MVA 法處理後的 MFCC 特徵時，發現它們相較於單一 CMVN 法或 MVA 法而言，能帶來更佳的辨識率，且我們所提出之三種新方法的表現幾乎仍然優於 TSN-1 法與 TSN-2 法。此外，我們探討 LSSF 法與 MSI 法可能存在之頻譜遺漏(leakage)的缺點，而提出相對應的修正方法，發現這些修正法能更進一步改善原始 LSSF 法與 MSI 法的效能。

若就三種新方法彼此作比較，ERTF 法與 LSSF 法運算複雜度較大，MSI 法則相對較小，雖然 ERTF 法對原始 MFCC 特徵而言，表現比 LSSF 法與 MSI 法來得好，但當它們作用於 CMVN 法或 MVA 法處理過後的 MFCC 特徵時，其效能的差異性已經很小，這意味著運算複雜度較小的 MSI 法相對於 ERTF 法與 LSSF 法而言，可能有更佳的應用性。

六、參考文獻

[1] 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

[2] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", IEEE Trans. on Acoustics, Speech and Signal Processing, 1981

[3] S. Tiberewala and H. Hermansky, "Multiband and Adaptation Approaches to Robust Speech Recognition", 1997 European Conference on Speech Communication and Technology (Eurospeech 1997)

[4] H. Hermansky and N. Morgan, "RASTA Processing of Speech", IEEE Trans. on Speech and Audio Processing, 1994

[5] C-P. Chen and J-A. Bilmes, "MVA Processing of Speech Features", IEEE Trans. on Audio, Speech, and Language Processing, 2006

[6] S. Yoshizawa, N. Hayasaka, N. Wada and Y. Miyanaga, "Cepstral Gain Normalization for Noise Robust Speech Recognition", 2004 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2004)

[7] 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, 2006

[8] X. Xiao, E-S. Chng, and Haizhou Li, "Temporal Structure Normalization of Speech Feature for Robust Speech Recognition", IEEE Signal Processing Letters, vol. 14, 2007

[9] H. G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", Proceedings of ISCA IIWR ASR2000, Paris, France, 2000

[10] S. K. Mitra, "Digital Signal Processing", The McGraw-Hill International, 3rd edition, 2006

[11] ITU recommendation G.712, "Transmission Performance Characteristics of Pulse Code Modulation Channels," Nov. 1996

[12] http://htk.eng.cam.ac.uk/