最小變異數調變頻譜濾波器於強健性語音辨識之研究

A Study of Minimum Variance Modulation Filter for Robust Speech Recognition

謝仁豪 Ren-hau Hsieh
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University
s97323560@ncnu.edu.tw

范顯騰 Hao-teng Fan
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University
s96323516@ncnu.edu.tw

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
jwhung@ncnu.edu.tw

摘要

本論文所探討的是語音特徵強健性技術，藉此改善雜訊環境下語音辨識的效能。我們利用原始最小變異數調變濾波器法設計的環境失真目標函數，應用至求取濾波器之最佳頻率響應上，進而發展出兩種特徵時間序列濾波器求取演算法，分別爲基於最小變異數準則之最小平方頻譜擬合法(MV-LSSF)及基於最小變異數準則之強度頻譜內插法(MV-MSI)。在這兩種方法中，利用我們所求得的濾波器之最佳頻率響應取代原始最小平方頻譜擬合法(LSSF)與強度頻譜內插法(MSI)中所使用的濾波器，來得到欲逼近的目標功率頻譜密度。從 Aurora-2 連續數字資料庫的實驗結果證實，這兩種基於最小變異數準之調變頻譜正規化法，在各種雜訊環境下都優於傳統的兩種調變頻譜正規化法，而得到更佳的辨識精確度。與基礎實驗結果相比較，MV-LSSF 與 MV-MSI 所達到之相對錯誤降低率分別爲在 55.41%與 51.20%，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 modulation spectra of speech features are often distorted due to environmental interferences. In order to reduce the distortion, in this paper we apply the minimum variance (MV) criterion to obtain the optimal frequency response of the temporal filter, and then two approaches, least-squares spectral fitting (LSSF) and magnitude spectrum interpolation (MSI) are used to obtain the filtered feature sequence. Accordingly, two new temporal processing approaches are proposed, which are named MV-LSSF and MV-MSI, respectively.

In the Aurora-2 clean-condition training task, we show that the new MV-LSSF and MV-MSI give more than 50% relative error rate reduction over the baseline, and provide relative error rate reductions of 8.18% and 2.73% over the conventional LSSF and MSI, respectively.

Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010), Pages 86-97, Puli, Nantou,Taiwan, September 2010.

These results reveal that the proposed methods significantly enhance the robustness of speech features in noise-corrupted environments.

關鍵詞：自動語音辨識、最小變異數、調變頻譜、強健性語音特徵

Keywords: speech recognition, minimum variance, modulation spectra, robust speech features

一、簡介

縱使語音科技日新月異，自動語音辨識(automatic speech recognition, ASR)[1]依舊是眾多專家、學者研究開發的標的。主要原因在於實際生活環境中存在著多方面的變異性(variation)影響辨識效果，這當中影響語音辨識的變異性包含了訓練環境與應用環境之間的環境不匹配(environmental mismatch)、訓練語者與應用語者之間的語者差異性(speaker variation)及不同語者或同一語者在發音上的變異性(pronunciation variation)等許多因素，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，有許許多多的學者持續不斷朝著努力改善以上幾種的語音差異性，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。

針對環境不匹配的狀況發展出許多強健性方法，綜觀而言，大致包含特徵補償[2]與模型補償[3]兩大類型，而特徵補償方法當中，有一類型是針對語音辨識所用的特徵參數之統計量作正規化處理，這些處理方式通常是作在特徵之時間序列域(temporal domain)上，目的是將強調特徵中語音的成分，將雜訊的成分壓抑下來，或是使不同環境下的語音特徵之統計量都能趨於一致，藉此提高語音辨識率，這些方法例如倒頻譜平均值正規化法(cepstral mean normalization, CMN)[4]、倒頻譜平均值與變異數正規化法(cepstral mean and variance normalization, CMVN)[5]、相對頻譜法(RelAtive SpecTra, RASTA)[6]、倒頻譜平均值與變異數正規化結合自回歸動態平均濾波器法(cepstral mean and variance normalization plus auto-regressive-moving average filtering, MVA)[7]、與統計圖等化法(histogram equalization, HEQ)[8]、時間序列結構正規化法(temporal structure normalization, TSN)[9]、等漣波時間序列濾波器法(equi-ripple temporal filter, ERTF)[10]、最小平方頻譜擬合法(least squares spectrum fitting, LSSF) [10]、強度頻譜內插法(magnitude spectrum interpolation, MSI) [10]等。而特別一提的是，2009 年時有學者提出了最小變異數調變濾波器設計法 (minimum variance modulation filter, MVMF)[11]，其主要是根據最小化雜訊變異數的最佳化目標，進而推得特徵之時間序列域上的濾波器脈衝響應(impulse response)，藉由對語音特徵濾波處理，而改善語音特徵的雜訊強健性。

以上各種技術主要是直接或間接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， 因此在本篇論文中，我們參考了 MVMF 法[11]的構想，使用變異數最小化之最佳準則來處理語音特徵，但我們所發展的方法與原始 MVMF 法不同點在於，它們是求得最佳的濾波器之頻率響應(frequency response)，即調變頻域上的最佳化，再經由前述之 LSSF 與 MSI 法，求得濾波器處理後的語音特徵，而並非如 MVMF 法直接在特徵之時間序列域的最佳化求得濾波器的脈衝響應。在實驗結果發現，我們所新提出的 MV-LSSF 法與 MV-MSI 法，所對應之辨識效能優於原始 LSSF 法與 MSI

法，且 MV-LSSF 法優於 MVMF 法，而 MV-MSI 法效果則與 MVMF 法十分相近。

本論文其他章節概要如下：在第二章中介紹本論文所提出之新方法，即 MV-LSSF 與 MV-MSI 法；第三章將呈現本論文所提出的新方法之辨識實驗結果與討論，第四章爲結論與未來展望。

二、基於最小變異數之調變頻譜正規化法

在本章中，我們首先簡略介紹前學者所提出之最小變異數調變濾波器法[11]，接著，我們介紹本論文所提出之兩個新方法，即是將最小變異數調變濾波器設計的目標函數，應用至求取濾波器之頻率響應上，進而延伸出兩種特徵時間序列處理演算法，分別爲基於最小變異數之最小平方濾波器法(MV-LSSF)與及基於最小變異數之強度頻譜內插法(MV-MSI)，這兩種新方法的詳細步驟將於本章詳述。

(1)最小變異數調變濾波器法(minimum variance modulation filter, MVMF)

一設計得當的特徵時間序列濾波器，可以凸顯特徵中的語音成分並抑制雜訊成分，進而提升語音特徵的強健性。而最小變異數濾波器(MVMF)設計法[11]，主要是根據三個方向來設計特徵的時間序列濾波器：

1. 濾波器本身可以隨著不同語句（可能對應不同的雜訊干擾環境）而作動態調整。
2. 定義一個『環境失真』的目標函數，經由調變濾波器的設計，使處理後的環境失真目標函數值能趨於最小。
3. 濾波器的設計，除了考量到降低雜訊成分外，也同時考慮到原始語音成分儘量不受到影響與更動。

附帶一提的是，在文獻[11]中的最小變異數調變濾波器爲第一類線性相位濾波器(type I linear-phase filter)，即濾波器長度爲奇數且前後對稱，如我們所知，線性相位濾波器只改變輸入訊號的頻譜強度及造成固定的時間延遲，並不會造成訊號時間延遲上的失真，而第一類線性相位濾波器額外優點則是可以不受限地近似各種型態的濾波器（如低通、高通、帶通與帶拒等濾波器等）。

以下爲 MVMF 法中設計濾波器係數的步驟：

1. 對於任一語句的某一維特徵時間序列，定義其『環境失真』(environmental mismatch)如下式：

$$\alpha = \lambda \int_{-\pi}^{\pi} \left| H(\omega) \right|^2 P_N(\omega) d\omega + \int_{-\pi}^{\pi} \left| 1 - H(\omega) \right|^2 P_S(\omega) d\omega \qquad \text{式(1)}$$

其中，$H(\omega)$ 爲濾波器之頻率響應，$P_N(\omega)$ 爲雜訊的功率頻譜密度，$P_S(\omega)$ 爲乾淨語音的功率頻譜密度，λ 爲比例參數(爲一常數)，代表環境失真中雜訊所佔的比例。從式(1)的等號右邊可看出，前項代表濾波器處理後之雜訊的失真，後項爲濾波器本身對乾淨語音的失真。

2. 假設我們使用第一類有限長度脈衝響應(finite impulse response, FIR)濾波器，則其頻率響應 $H(\omega)$ 如下式表示:

$$H(\omega)=\sum_{l=-(L-1)/2}^{(L-1)/2} h(l)e^{-j\omega l}$$

其中，L 爲濾波器係數$h(l)$的點數，爲一奇數，且$h(l)$滿足前後對稱的性質， 即$h(l)=h(-l)$。將式(2)代入式(1)，可得:

$$\alpha=\lambda\sum_{k=-\frac{L-1}{2}}^{\frac{L-1}{2}}\sum_{l=-\frac{L-1}{2}}^{\frac{L-1}{2}} h(k)h(l)^*\int_{-\pi}^{\pi}P_N(\omega)e^{j\omega(l-k)}d\omega-\sum_{k=-\frac{L-1}{2}}^{\frac{L-1}{2}} h(k)\int_{-\pi}^{\pi}P_S(\omega)e^{-j\omega k}dw$$

$$-\sum_{l=-\frac{L-1}{2}}^{\frac{L-1}{2}} h(l)^*\int_{-\pi}^{\pi}P_S(\omega)e^{j\omega l}dw+\sum_{k=-\frac{L-1}{2}}^{\frac{L-1}{2}}\sum_{l=-\frac{L-1}{2}}^{\frac{L-1}{2}} h(k)h(l)^*\int_{-\pi}^{\pi}P_S(\omega)e^{j\omega(l-k)}d\omega+\int_{-\pi}^{\pi}P_S(\omega)dw \qquad \text{式(3)}$$

藉由矩陣與向量表示法，式(3)可改寫爲:

$$\alpha=\lambda\mathbf{h}^T\mathbf{R}_N\mathbf{h}-2\mathbf{h}^T\mathbf{r}_S+\mathbf{h}^T\mathbf{R}_S\mathbf{h}+\int_{-\pi}^{\pi}P_S(\omega)d\omega \qquad \text{式(4)}$$

其中， $\mathbf{R}_N$爲雜訊倒頻譜特徵的自相關矩陣(autocorrelation matrix)， $\mathbf{r}_S$爲乾淨語音倒頻譜特徵的自相關係數向量， $\mathbf{R}_S$爲乾淨語音倒頻譜特徵的自相關矩陣，$\mathbf{h}$爲調變濾波器係數向量，即$\mathbf{h}=\left[h\left(-(L-1)/2\right)\ \cdots\ h\left((L-1)/2\right)\right]$。

我們知道，一含有加成性雜訊的語音倒頻譜特徵，其雜訊成分與乾淨語音成分成非線性關係的組合，但在這裡，爲了降低計算的複雜度，假設乾淨語音與雜訊二成分在相關係數上呈現線性相加：

$$\mathbf{R}_{N+S}=\mathbf{R}_N+\mathbf{R}_S, \qquad \text{式(5)}$$

其中， $\mathbf{R}_{N+S}$爲含有雜訊的語音倒頻譜特徵之自相關矩陣。

將式(5)代入式(4)，整理後可得:

$$\alpha=\lambda\mathbf{h}^T\mathbf{R}_{N+S}\mathbf{h}+(1-\lambda)\mathbf{h}^T\mathbf{R}_S\mathbf{h}-2\mathbf{h}^T\mathbf{r}_S+\int_{-\pi}^{\pi}P_S(\omega)d\omega \qquad \text{式(6)}$$

3. 爲求得最佳濾波器係數$\mathbf{h}$使環境失真達到極小值，這裡對於上式(6)作針對$\mathbf{h}$的偏微分並令恆等式爲 0，可得最佳之調變濾波器係數爲

$$\mathbf{h}=\left(\lambda\mathbf{R}_{N+S}+(1-\lambda)\mathbf{R}_S\right)^{-1}\mathbf{r}_S \qquad \text{式(7)}$$

(2)根據最小變異數準則所得之調變濾波器的最佳頻率響應

在上一小節中，我們介紹環境失真極小值所對應的調變濾波器最佳係數$\mathbf{h}$，相當於在時間序列域上求取調變濾波器脈衝響應。在這一節中，我們試著在調變頻譜域上，根據相同的最佳化準則，求取調變濾波器之最佳頻率響應。以下爲其推導步驟：

1. 從式(1)所定義之環境失真量，可看出每一個頻率值ω對應的環境失真密度爲

$$\alpha(\omega) = \lambda\left|H(\omega)\right|^2 P_N(\omega) + \left|1 - H(\omega)\right|^2 P_S(\omega), \qquad \text{式(8)}$$

亦即式(1)爲式(8)對ω的積分。與 MVMF 法不同之處，在於我們這裡直接求取最佳頻率響應$H(\omega)$，使式(8)中的環境失真$\alpha(\omega)$達到極小值，進而使其積分值α達到極小值。將式(8)作針對$H(\omega)$的偏微分，並令其爲 0，我們得到:

$$\lambda H^*(\omega)P_N(\omega) - P_S(\omega) + H^*(\omega)P_S(\omega) = 0, \qquad \text{式(9)}$$

2. 假設乾淨語音與雜訊二成分在功率頻譜密度上呈現線性相加，如下所示:

$$P_{N+S}(\omega) = P_N(\omega) + P_S(\omega) \qquad \text{式(10)}$$

其中，$P_{N+S}(\omega)$爲含有雜訊語音之特徵的功率頻譜密度。代入式(9)，可得最佳濾波器的頻率響應如下所示:

$$H^*(\omega) = \frac{P_S(\omega)}{\lambda P_{N+S}(\omega) + (1-\lambda)P_S(\omega)} \qquad \text{式(11)}$$

由上式可看出$H^*(\omega)$其實是一正實數，故可去掉複數共軛符號而得:

$$H(\omega) = \frac{P_S(\omega)}{\lambda P_{N+S}(\omega) + (1-\lambda)P_S(\omega)} \qquad \text{式(12)}$$

式(12)顯示最佳濾波器的頻率響應爲零相位，但其會對應到一個非因果(non-causal)的濾波器。我們可以使用沒有相位失真的因果性(causal)線性相位濾波器，使其頻率響應的強度成分等同於式(11)所示，則其效果等同於原非因果的濾波器。

(3) 根據最小變異數準則所得之調變頻譜正規化法

在原始的 LSSF 法與 MSI 法[10]中，我們會定義一參考調變頻譜，做爲各語音特徵序列的調變頻譜正規化的目標，目的希望原始的特徵序列，經由這些方法處理後所得的特徵序列，其調變頻譜的強度成份能趨於一致，其中，參考調變頻譜其實是由用以訓練之乾淨語音特徵的調變頻譜平均而得，並可因此得知，在原始的 LSSF 法與 MSI 法中間接使用的時間序列濾波器其頻率響應之強度平方爲：

$$\left|H(\omega)\right|^2 = \frac{P_S(\omega)}{P_{N+S}(\omega)}, \qquad \text{式(13)}$$

其中$P_{N+S}(\omega)$爲含有雜訊語音之特徵的功率頻譜密度，$P_S(\omega)$爲乾淨語音的功率頻譜密度。若將式(13)與式(12)中相較，發現原始 LSSF 與 MSI 之濾波器頻率響應（強度），等於最小變異數準則推得的濾波器頻率響應（強度）之正平方根，且其中參數λ設爲 1。

根據以上的觀察，我們試圖對於 LSSF 法與 MSI 法中所間接使用的濾波器作改變，換言之，我們將使用經過(12)式所推出、符合最小變異數之最佳準則之濾波器頻率響應，取代式(13)，來得到 LSSF 與 MSI 法中欲逼近的目標功率頻譜密度，冀望能進一步提

升這兩種調變頻譜正規化的效能。根據這樣的改變所對應的 LSSF 與 MSI 法，我們分別命名爲『基於最小變異數準則之 LSSF 法』（簡稱 MV-LSSF）與『基於最小變異數準則之 MSI 法』（簡稱 MV-MSI）。

- **基於最小變異數準則之 LSSF 法（MV-LSSF）**

在 MV-LSSF 法中，我們將每一個待正規化的 N 點特徵序列$\{x[n];1\le n\le N\}$先定義一 $2P$ 點的參考調變頻譜，做爲此特徵序列的調變頻譜正規化的目標，如下所示：

$$Y(\omega_k)=\left|Y(\omega_k)\right|\exp\left(j\theta_X(\omega_k)\right),\qquad 0\le k\le 2P-1 \qquad \text{式(14)}$$

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

$$\left|Y(\omega_k)\right|=H(\omega_k)\left|X(\omega_k)\right|=\left|\frac{P_S(\omega_k)}{\lambda P_X(\omega_k)+(1-\lambda)P_S(\omega_k)}\right|\left|X(\omega_k)\right|,\qquad 0\le k\le 2P-1 \qquad \text{式(15)}$$

其中$H(\omega_k)$即是採用式(12)之根據最小變異數準則所求取之最佳頻率響應，而強度成份$\left|X(\omega_k)\right|$和相角成份$\theta(\omega_k)$爲$\{x[n]\}$經過 *2P* 點之離散傅立葉轉換(discrete Fourier transform, DFT)所得到。在此，必須注意的是，隨著不同的語句變化，特徵長度 N 會因此改變，但是在此的 DFT 取樣點數 *2P* 則設定爲一固定值，換句話說，每一語句對應的參考調變頻譜的長度都是相等的。

在此，我們利用最小平方化(least-squares)的最佳化準則求得一新的特徵序列，使新的特徵序列$\{y[n]\}$的調變頻譜逼近式(14)的參考調變頻譜，如下所示：

$$y[n]=\min_{\{\bar{y}[n]\}0\le n\le N-1\}}\sum_{k=0}^{2P-1}\left|\sum_{n=0}^{N-1}\widehat{y}[n]e^{-j\frac{2\pi nk}{2P}}-Y(\omega_k)\right|^2,\qquad (2P\ge N) \qquad \text{式(16)}$$

其中 *2P* 爲 DFT 取樣點數，*N* 爲此特徵序列的點數。由此可知，式(16)中，藉由 MV-LSSF 法中所求得之新特徵序列$\{y[n]\}$，其 *2P* 之 DFT 與式(14)的參考調變頻譜之間具有最小平方誤差的性質。

- **基於最小變異數準則之 MSI 法（MV-MSI）**

在 MV-MSI 法的過程中，首先爲每一個待正規化的 N 點特徵序列，$\{x[n];1\le n\le N\}$定義一個 N 點的參考調變頻譜，作爲此徵序列之調變頻譜正規化之目標，如下所示：

$$\hat{Y}(\omega_{k'})=\left|Y(\omega_{k'})\right|\exp\left(j\theta_X(\omega_{k'})\right),\qquad 0\le k'\le N-1 \qquad \text{式(17)}$$

其中相位成份$\theta_X(\omega_{k'})$爲$\{x[n]\}$經過 *N* 點之 DFT 而得，由於 MV-LSSF 法中原始 *2P* 點的參考頻譜(如式(14))所涵蓋的頻率範圍與這裡式(17)所求的$\hat{Y}(\omega_{k'})$頻率範圍相同，因此我們利用線性內插法(linear interpolation)的方式，藉由式(15)當中 *2P* 點之以最小變異數準

則所得之最佳參考頻譜強度 $\left\{\left|Y\left(\omega_k\right)\right| 0 \le k \le 2P-1\right\}$ 來求取 $\left\{\left|\hat{Y}\left(\omega_{k'}\right)\right| 0 \le k' \le N-1\right\}$ 之近似值。而就式(17)的 $\hat{Y}(\omega_{k'})$ 來說，本身爲一實數序列之離散傅立葉轉換，其強度成份 $\left|\hat{Y}(\omega_{k'})\right|$ 必須遵守左右對稱之性質，即爲如下所示:

$$\left|\hat{Y}\left(\omega_{k'}\right)\right| = \left|\hat{Y}\left(\omega_{N-k'}\right)\right| \qquad \text{式(18)}$$

因此我們利用 $\left\{\left|Y\left(\omega_k\right)\right|\right\}$ 的左半部執行內插法，求取 $\left\{\left|\hat{Y}\left(\omega_{k'}\right)\right|\right\}$ 的左半部 $\left\{\left|\hat{Y}\left(\omega_{k'}\right)\right| 0 \le k' \le \left\lceil\frac{N}{2}\right\rceil\right\}$，再利用左右對稱的性質，求取 $\left\{\left|\hat{Y}\left(\omega_{k'}\right)\right|\right\}$ 右半部 $\left\{\left|\hat{Y}\left(\omega_{k'}\right)\right| N-1-\left\lceil\frac{N}{2}\right\rceil \le k' \le N-1\right\}$。最後可得到 $\left\{\left|\hat{Y}\left(\omega_{k'}\right)\right| 0 \le k' \le N-1\right\}$。

接下來，我們直接對式(17)中的 $\left\{\left|\hat{Y}\left(\omega_{k'}\right)\right|\right\}$ 作 N 點的反傅立葉轉換(inverse discrete Fourier trans, IDFT)，求得新的特徵序列 $\left\{y[n]\right\}$，如下所示:

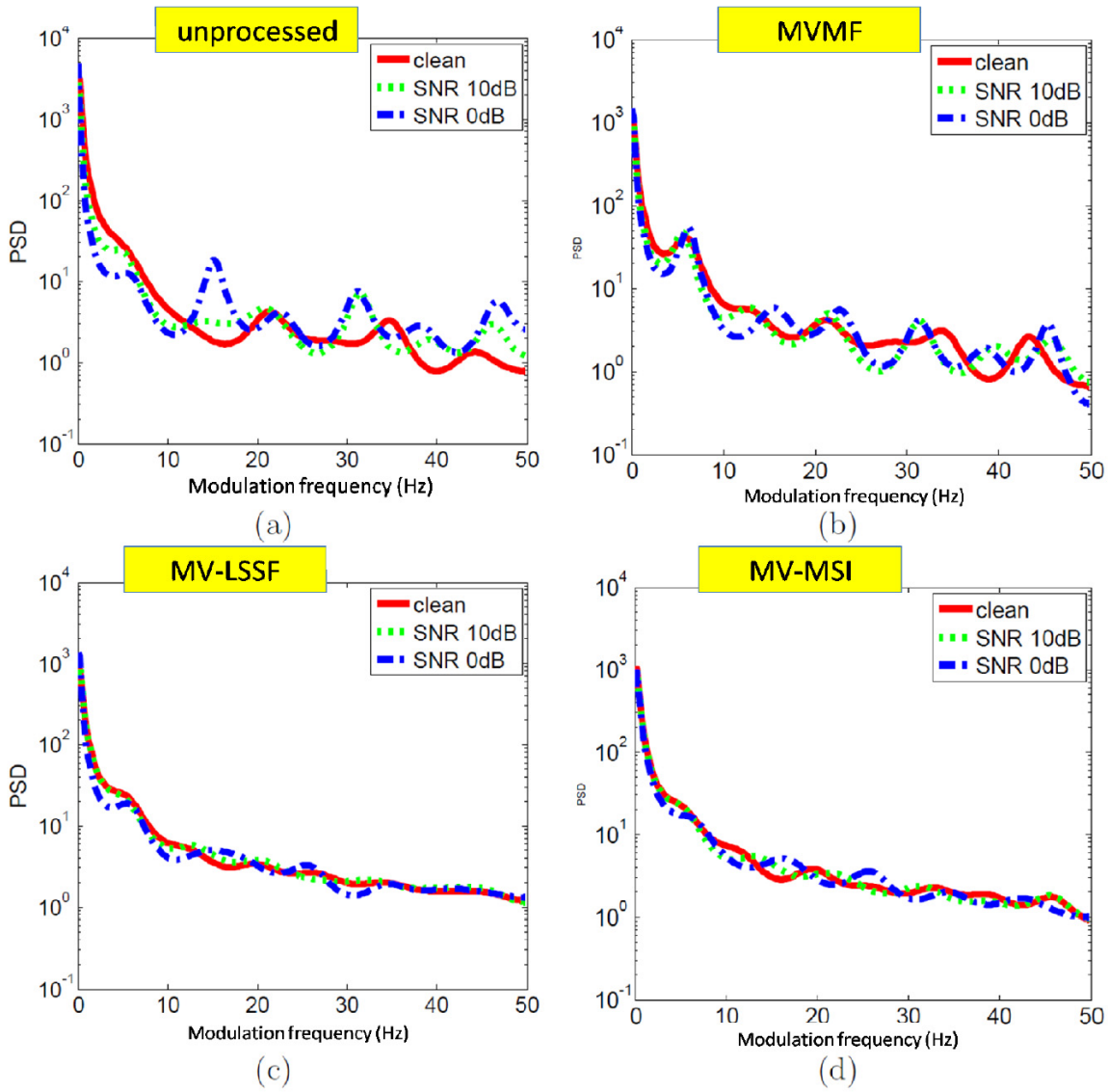
$$y[n] = \frac{1}{N}\sum_{k'=0}^{N-1}\hat{Y}\left(\omega_{k'}\right)\exp\left(j\frac{2\pi n k'}{N}\right), \qquad 0 \le n \le N-1 \qquad \text{式(19)}$$

上式(19)的 $\left\{y[n]\right\}$ 即爲藉由 MV-MSI 所得之新語音特徵序列。

在這裡，我們對提出的 MV-LSSF 與 MV-MSI 法與 MVMF 法作初步效能的比較，我們根據這些方法在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。這裡使用 AURORA-2 資料庫[12]中的 FAK_3Z82A 乾淨語音檔，經由加入不同訊雜比(signal-to-noise ratio, SNR)的地下鐵(subway)雜訊所產生之雜訊語音檔，轉成特徵函數後，再分別經過上述三種方法作處理。圖一(a)(b)(c)(d)分別代表原始未經處理之第十二維 MFCC 特徵序列(c_{12})、MVMF 法、MV-LSSF 法與 MV-MSI 法處理後之 c_{12} 序列的功率頻譜密度(power spectral density, PSD)曲線圖。根據圖一，我們可以發現：

1. 觀察圖一(a)得知，在不同 SNR 值下(clean, 10dB 與 0dB)，未處理過的 c_{12} 序列，其功率頻譜密度(PSD)曲線因受到加成性雜訊(additive noise)的影響，存在明顯的失真情形。而經由圖一(b)可看出，MVMF 法處理後之 c_{12} 序列，在較低的調變頻率範圍[0,10Hz]，其 PSD 失真的情況已有很明顯的降低，但相對於較高的調變頻率範圍，PSD 失真的情形並沒有太大的改善。

2. 圖一(c)爲 MV-LSSF 法處理後所得到的 c_{12} 序列之 PSD 圖，很明顯可看出在全部的調變頻率範圍，其 PSD 失真情形皆有效地降低，尤其是較高頻的調變頻率範圍[35, 50Hz]，在不同的 SNR 值下之 PSD 曲線趨近於乾淨語音特徵(clean)之 PSD 曲線；而圖一(d)爲 MV-MSI 法處理後所得到的特徵序列之 PSD 圖，可看出不管較低的調變頻率[0,10Hz]之間或較高的調變頻率範圍[10Hz, 50Hz]，其 PSD 失真的情況也有很明顯的降低。



圖一： (a)原始 c_{12} 特徵序列、(b)MVMF 法、(c)MV-LSSF 法與(d)MV-MSI 法作用於不同訊雜比下之 c_{12} 特徵序列之功率頻譜密度曲線圖

三、實驗結果與分析討論

本章將介紹本論文相關辨識實驗的各類設定，第一小節介紹實驗所用的 Aurora-2 語音資料庫與辨識效能評估方式，第二小節介紹語音辨識實驗所使用的語音聲學模型、呈現基本實驗的辨識結果並加以討論。

(一) 實驗環境與架構設定

我們實驗中所採用的語音資料庫爲歐洲電信標準協會(European Telecommunication Standard Institute, ETSI)所發行的 Aurora-2 語音資料庫[12]，它是由美國成年男女以人工方式錄製的一系列連續英文數字字串，其中測試語料庫的每一字串中，加入各種加成性雜訊及通道效應的干擾，這八種加成性的雜訊，分別爲：地下鐵(subway)、人的嘈雜聲(babble)、汽車(car)、展覽館(exhibition)、餐廳(restaurant)、街道(street)、機場(airport)、火車站(train station)等環境的雜訊，並以不同程度的訊雜比(signal-to-noise ratio, SNR)摻雜，分別爲：clean、20 dB、15 dB、10 dB、5 dB、0 dB 與-5 dB；而通道效應分別爲.712

與 MIRS 兩種通道標準，它們是透過國際電信聯盟(International Telecommunication Union,ITU)[13]所訂定而成的。

在本論文當中之所有的語音辨識實驗，皆使用 12 維梅爾倒頻譜係數(c_1~c_{12})與 1 維對能量(log-energy)，附加其一階差量與二階差量，共 39 維，作爲原始特徵參數。而實驗中所使用的聲學模型(acoustic models)是隱藏式馬可夫模型(hidden Markov model, HMM)。在訓練方式上，藉由 HTK[14]這套軟體來訓練隱藏式馬可夫模型，包含了 11 個數字模型(zero, one,. . . nine and oh)以及一個靜音(silence)模型，每個數字的 HMM 內皆包含 16 個狀態，而每個狀態是由 20 個高斯密度函數組成。

(二) 實驗結果呈現與討論

首先，我們在表一中，呈現了幾種在第一章中所提到的語音特徵時間序列處理技術，在上述實驗環境所得之辨識精確度，而表中的 AR 與 RR 分別代表了相對於基礎實驗(baseline)的絕對錯誤降低率與相對錯誤降低率。由此表中可以發現，這幾種時間序列特徵強健性方法，都能有效改善雜訊環境下的辨識結果，在這些方法中，我們依照其辨識效能由高至低依序爲 HEQ，MVA，MVMF，LSSF，MSI，CMVN 與 CMN，因此，HEQ 法是最能有效改善雜訊環境影響下之語音辨識，雖然其他六種方法的辨識效能不及 HEQ 法，但是相較於基礎實驗而言，它們都對語音辨識效能都有顯著的提升，且除了 CMN 與 CMVN 外，其他方法與 HEQ 所得之辨識率的差距皆在 2%以內。

表一：數種語音特徵時間序列處理技術所得之辨識精確率(%)

Method	Set A	Set B	Set C	average	AR (%)	RR (%)
baseline	71.98	67.79	78.28	71.56	—	—
CMN	80.69	83.41	80.09	81.66	10.10	35.51
CMVN	83.55	83.75	81.57	83.23	11.67	41.03
HEQ	86.90	87.73	87.56	87.36	15.80	55.56
MVA	86.69	86.89	84.98	86.43	14.87	52.29
MVMF	85.09	87.47	86.22	86.27	14.71	51.72
LSSF	85.66	86.88	85.87	86.19	14.63	51.44
MSI	84.99	86.55	85.56	85.73	14.17	49.82

接著，在表二中，我們呈現所提出的兩個新方法，MV-LSSF 與 MV-MSI，在上述實驗環境所得之辨識精確率，並在這表中，列出了 MVMF 法與原始 LSSF 與 MSI 法所得之辨識率，以供比較。在 MV-LSSF 與 MV-MSI 中，式(15)中的比例參數λ設定爲 0.5（此接近參考文獻[11]對 MVMF 法中的最佳設定值 0.49），MV-LSSF 法所用的離散傅立葉轉換(DFT)點數 2P，設定爲 1024 點，MV-MSI 法所用的離散傅立葉轉換(DFT)點數 2P，設定爲 256 點，這些設定與原始 LSSF 與 MSI 的設定皆相同，因此可以比較

出更新目標調變頻譜功率密度所帶來的差異性。

表二：數種語音特徵時間序列處理技術所得之辨識精確率(%)

Method	Set A	Set B	Set C	average	AR (%)	RR (%)
baseline	71.98	67.79	78.28	71.56	—	—
MV-LSSF	**86.77**	**88.02**	**87.00**	**87.32**	**15.76**	**55.41**
LSSF	85.66	86.88	85.87	86.19	14.63	51.44
MV-MSI	**85.18**	**87.16**	**85.94**	**86.12**	**14.56**	**51.20**
MSI	84.99	86.55	85.56	85.73	14.17	49.82
MVMF	85.09	87.47	86.22	86.27	14.71	51.72

在表二中，我們有以下的觀察結果，

1. MV-LSSF 法與原始 LSSF 法相較下，整體平均辨識率可提升 1.13%，而 MV-MSI 法相較於原始 MSI 法而言，整體平均辨識率可提升 0.39%。由此可知，MV-LSSF 與 MV-MSI 分別優於原始之 LSSF 與 MSI，可能原因爲，MV-LSSF 與 MV-MSI 使用了最小變異數的最佳準則，其目標功率頻譜之求取上同時考慮了當下處理之單一語句的功率頻譜（式(15)中的 $P_X(\omega_k)$）與平均乾淨功率頻譜（式(15)中的 $P_S(\omega_k)$），對語音特徵可達到較佳的強健化效果。

2. 在調變頻譜域推得的 MV-LSSF 法和 MV-MSI 法，與時間序列域推得的 MVMF 法相比較，在整體的平均辨識率上，MVMF-LSSF 法約有 1%的進步率，而 MVMF-MSI 則約與 MVMF 法相等（些微退步了 0.15%）。然而，由於 MVMF 必須使用到反矩陣的運算（如式(7)所示），相對於 MV-LSSF 與 MV-MSI 而言，執行上複雜度較高，由此看出我們所新提出的兩種方法，既可與 MVMF 法效果並駕齊驅或略佳，且在運算上更有效率，因此更具優勢。同時，我們所提出的 MV-LSSF 法，其達到的辨識率進步程度已經與表一所討論之最佳效果的 HEQ 法幾乎相同，足見其優越性。

最後，我們嘗試將所提出之 MV-LSSF 與 MV-MSI 法，與特徵統計正規化法之一的 CMVN 法結合，觀察其是否能促成辨識率更進一步的提升，在此，原始語音倒頻譜特徵先經過 CMVN 處理後，再分別經過三種 MV 法，即 MVMF、MV-LSSF 與 MV-MSI，所得之辨識率詳列於表三。從此表中，我們明顯看出，上述三種 MV 法與 CMVN 法皆有明顯的加成性，相對於單一 CMVN 法而言，將 MVMF、 MV-LSSF 與 MV-MSI 處理於 CMVN 後的特徵上，分別有 6.66%、7.24%與 6.50%之平均辨識率的提升，若將表三與表二相較，也可看出，結合 CMVN 法後，三種 MV 法也能有更明顯進步的效能。跟之前結果類似，結合了 CMVN 法後，MV-LSSF 的效果仍然最好，其次爲 MVMF 與 MV-MSI。

表三：數種語音特徵時間序列處理技術所得之辨識精確率(%)

Method		Set A	Set B	Set C	average	AR (%)	RR (%)
baseline		71.98	67.79	78.28	71.56	—	—
CMVN		80.69	83.41	80.09	81.66	10.10	35.51
CMVN	MVMF	88.47	88.98	86.70	88.32	16.76	58.93
	MV-LSSF	**89.78**	**90.39**	**89.14**	**89.90**	**18.33**	**64.47**
	MV-MSI	**88.17**	**88.92**	**86.60**	**88.16**	**16.59**	**58.35**

四、結論與未來展望

本論文中，我們基於最小變異數準則中所定義的環境失真，使其降至極小值進而求得頻譜上調變濾波器之最佳頻率響應，並應用於兩種調變頻譜正規化法：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)，進而發展出了新的兩種新方法，即 MV-LSSF 與 MV-MSI。由實驗結果發現，在語音辨識效能上，這兩種在調變頻譜域所發展的新方法，若與時間序列域的 MVMF 法相比較，MV-LSSF 法辨識率進步約有 1.13%，而 MV-MSI 則約與 MVMF 法相等，但 MV-LSSF 和 MV-LSI 比 MVMF 有較低的運算複雜度。

在未來展望中，我們將進一步研究最小變異數調變濾波器法的理論基礎，並希望能藉由更嚴謹的數學分析與推導，將所求得的最佳頻率響應應用於其他調變頻譜正規化法中，使辨識效能可以更加提升。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，或是應用於其他類型的干擾失真環境，探討這一系列調變頻譜正規化法在不同類型之語音辨識系統的效能，進一步驗證我們提出的改進點與探討其實用性。

參考文獻

[1] 王小川, "語音訊號處理," 全華科技圖書, 2004.

[2] S. Ikbal, H. Hermansky and H. Bourlard, "Phase autocorrelation (PAC) derived robust speech features," *2003 International Conference on Acoustics, Speech and Signal Processing* (*ICASSP 2003*), pp.133-136, 2003.

[3] J. Hung, J. Shen and L. Lee, "New approaches for domain transformation and parameter combination for improved accuracy in parallel model combination (PMC) techniques," *IEEE Trans. on Speech and Audio Processing*, pp.842-855, 2001.

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

[5] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for

noise robust speech recognition," *Proceedings of the 38th Southeastern Symposium on System Theory Speech Communication*, Vol. 25, pp.133-147, 1998.

[6] H. Hermansky and N. Morgan, "RASTA processing of speech features," *IEEE Trans. on Speech and Audio Processing*, 1994.

[7] C. Chen and J. Bilmes, "MVA processing of speech features," *IEEE Trans. on Audio, Speech, and Language Processing*, pp.257-270, 2006.

[8] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust large vocabulary speech recognition," *IEEE Trans. on Audio, Speech and Language Processing*, pp.845-854, 2006.

[9] X. Xiao, E. S. Chug and H. Li, "Normalizing the speech modulation spectrum for robust speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP 2007*), 2007.

[10] C. Pan C. Wang J. Hung, "Improved modulation spectrum normalization techniques for robust speech recognition," *2008 International Conference on Acoustics, Speech and Signal Processing* (*ICASSP 2008*), pp.4089-4092, 2008.

[11] Y. B. Chiu and R. M. Stern, "Minimum variance modulation filter for robust speech recognition," *2009 International Conference on Acoustics, Speech and Signal Processing* (*ICASSP 2009*), pp.3962-3965, 2009.

[12] 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*, pp.181-188, 2000.

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

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