進階式調變頻譜補償法於強健性語音辨識之研究

Advanced Modulation Spectrum Compensation Techniques for Robust Speech Recognition

葉威志 Wei-Jeih Yeh
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University
s97323524@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
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
jwhung@ncnu.edu.tw

摘要

自動語音辨識是一門很值得研究開發的課題，現今多數的語音辨識系統若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， 則會受到環境雜訊的影響， 導致辨識效能明顯地下降，因此發展多年的環境強健性技術，即是針對此項缺點作改進。

在各種環境強健性技術中，有一類技術爲對語音特徵的調變頻譜作統計上的正規化， 而在先前這一類技術的研究裡，若對分頻段的頻譜做正規化處理，相對於全頻帶正規化的處理法有較好的強健性效能，但其中由於不等切的切割方式，將調變頻譜中低頻部份分的比較細，導致低頻範圍的子頻段，會有頻譜點數不足的問題，影響到我們計算其頻譜特徵統計值的精確度，因此這些方法應有改進的空間。基於此觀察，本論文提出一系列重疊式分頻段調變頻譜統計正規化法，此類方法可以有效提升子頻段中用以計算統計值的頻譜點數，提升統計值的精確度，進而改善分頻段統計正規化法的效能，可以使所得特徵在環境強健性上的效能更爲優越。

本論文採用國際通用的 AURORA-2 連續數字語料庫作一系列的語音辨識實驗，由實驗結果可明確驗證，我們提出的重疊式分頻段方法比起傳統非重疊式分頻段的方法更能有效地提升各種雜訊環境下的辨識精確率。此外，我們也將這些新方法結合傳統之時間序列域特徵正規化法，實驗結果皆顯示這樣的組合皆能比單一方法更有效地提升辨識率，足見它們有良好的加成性。

Abstract

In this paper, we propose a novel scheme in performing feature statistics normalization

Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010), Pages 236-250, Puli, Nantou,Taiwan, September 2010.

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 overlapped non-uniform sub-band segments, and then each sub-band segment is individually processed by the well-known normalization methods, like mean normalization (MN) and mean and variance normalization (MVN). 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 and more spectral samples within each band give rise to more accurate statistic estimates due to overlapping the adjacent segments. For the Aurora-2 clean-condition training task, the new proposed overlapping sub-band spectral MN and MVN provide further error rate reductions over the conventional non-overlapping ones.

關鍵詞：語音辨識、調變頻譜、正規化、強健性語音特徵參數

Keywords: speech recognition, modulation spectrum, statistics normalization, robust speech features

一、簡介

目前多數的語音辨識系統若在不受干擾的安靜環境下，且辨識的字彙量不是很大時，一般而言皆能得到相當滿意的辨識效果，但應用於許多真實的生活環境中，辨識效能一定會有所衰減，主要是實際生活環境與系統發展之環境彼此之間不匹配的情況發生，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。如何抑制環境不匹配的背景雜訊干擾問題，使其所帶來之語音辨識下降的影響達到最小，進而提升語音系統強健性，是本論文主要研究的方向。

爲了降低以上各種變異性所發展的各種技術，一般而言統稱爲強健性技術(robustness techniques)，本論文主要著重於發展降低環境之雜訊干擾的強健性演算法。 而在多種的強健性演算法中，有一類方法是將訓練與測試環境下的語音特徵 (通常爲"倒頻譜"特徵) 其時間序列統計特性做正規化，以降低訓練與測試環境之間的不匹配，進而達到提升辨識率的目的。 著名的語音統計正規化法包括了倒頻譜平均值正規化法(cepstral mean normalization, CMN)[1]、倒頻譜平均值與變異數正規化法(cepstral mean and variance normalization, CMVN)[2]與統計圖等化法(histogram equalization, HEQ)[3]等。以上各種方法主要是執行在語音倒頻譜特徵的時間序列域上，但在其效能的分析上，我們通常進一步探討雜訊及通道效應對於原始特徵之調變頻譜造成的失真，分析這些方法對於調變頻譜失真的改善效果，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統計正規化技術，如調變頻譜統計圖等化法(spectrum histogram equalization, SHE)[4]，此方法是針對語音特徵之調變頻譜的強度成分之機率分佈(probability distribution)作正規化處理；而根據許多的研究[5][6]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。學者 N. Kanedera 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[7]，且主要集中在 4 Hz 附近。藉由以上之各觀點，因此過去本實驗室的研究中[8]，嘗試將調變頻譜中的強度頻譜(magnitude

spectrum)切割成許多子頻段，且將低頻的部份切的比較細，再分別對各自子頻段的統計值作正規化處理，這種方法可以強調出低頻成份中語音的資訊，進而提升辨識率。 但是我們發現，當我們將頻段數目分的越多，低頻的部份會因爲頻段過窄，導致包含的頻譜點數變少，進而影響我們求取統計量的精確度。 也因爲如此，在本論文中，我們提出了基於強度頻譜之重疊式分頻段調變頻譜統計正規化法，其目的是藉由子頻段範圍的延伸，相對的增加各子頻段的頻譜點數，使我們能更精確的計算出統計值。而也因爲各子頻段彼此重疊，能增加彼此的相關性，並且提升辨識的效能。

本論文其他章節概要如下：在第二章將介紹本論文所提出之重疊式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重疊式分頻段調變頻譜統計正規化法的實驗結果，並且分析結合其他強健性特徵正規化法的效能。第四章則是結論與未來展望。

二、重疊式分頻帶調變頻譜統計正規化法

如前章所述，若以分頻段的形式分別對頻譜正規化，會得到較佳的強健性效果。然而，以不等切的方式將調變頻譜切割成大小不一的頻段（如圖一(a)所示），進而正規化，雖然效果相較於未切割的全頻式處理效果較好，然而其潛在問題爲：

1. 我們不能將頻段切得太細，否則將會使較窄細之頻段內所包含的頻譜點數過少甚至等於零，進而使欲估測的統計值因點數過少而精確度大打折扣，甚至無法計算統計值，而使正規化演算法無法執行。

2. 由於這些頻譜點數較少的頻段恰好都在較低頻的位置，如前所述，較低頻的成分對於語音辨識相對重要，若在這些低頻頻段處理不當，會使原先分頻段正規化處理的優勢降低。

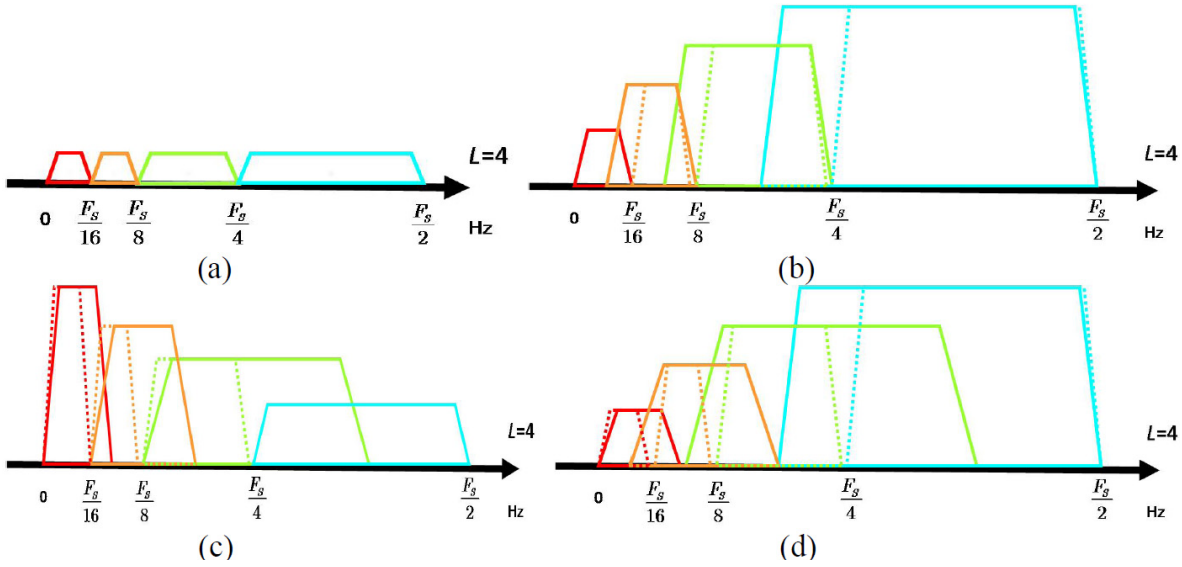
基於以上的觀察，在本論文中，提出了一系列的改進方法來補償上述之低頻段頻譜點數不足的缺點，這一系列的改進方法，基本精神就是將原先定義的子頻段作某程度上的延伸，使其涵蓋的頻率範圍變大，進而得以包括較多的頻譜點數，目的在於求取較精確的統計值，來使原先所構想的分頻段統計值正規化法能有更佳的表現。以上所述之延伸子頻段長度的方式，是將原先方法中互不重疊(non-overlapping)的子頻段，修正爲相鄰子頻段可以互相重疊(overlapping)，且進而衍生出三種重疊的方式（目前初始的研究中，我們暫定其被重疊的頻段寬度爲原始寬度的一半），分別爲

1. 向左重疊式(left-sided overlapping)子頻段切割法，即新的子頻段爲原始子頻段向左（較低頻）延伸，額外包含了其左方第一個子頻段的一半寬度（最低子頻段除外），此類型所對應的各種正規化法，將附帶以上標「*lo*」表示，其中的字母 *l* 與 *o* 分別代表了"left-sided"與"overlapping"。示意圖如圖一(b)所示。

2. 向右重疊式(right-sided overlapping)子頻段切割法，即新的子頻段爲原始子頻段向右（較高頻）延伸，額外包含了其右方第一個子頻段的一半寬度（最高子頻段除外），此類型所對應的各種正規化法，將附帶以上標「*ro*」表示，其中的字母 *r* 與 *o* 分別代表了"right-sided"與"overlapping"。示意圖如圖一(c)所示。

3. 雙邊重疊式(two-sided overlapping)子頻段切割法，即新的子頻段爲原始子頻段向右

(較高頻)延伸，額外包含了其右方第一個子頻段的一半寬度（最低子頻段僅向右延伸，而最高子頻段則僅向左延伸），此類型所對應的各種正規化法，將附帶以上標「*to*」表示，其中的字母 *t* 與 *o* 分別代表了"two-sided"與"overlapping"。示意圖如圖一(d)所示。



圖一：各種分頻段方法之示意圖 (a)非重疊式(b)向左重疊式(c)向右重疊式(d)雙邊重疊式

另外，以上所提到之原始方法裡，低頻帶之子頻段頻譜點數較少、導致求取此子頻段之頻譜統計量精確度不足問題，主要是發生在處理單一語句特徵的情形下，當我們欲得到單一（非重疊）子頻段之目標統計值(target statistics)時，由於是把所有訓練語句之此一子頻段頻譜點集合起來再作運算，因此較無上述之點數不足的問題，因此，在本論文中，我們新提出的重疊式方法又可分成兩種，分別爲：

第 I 型：利用全部訓練語句求取單一子頻段的目標統計值時，各子頻段是可以相互重疊的，但重疊方式必須與執行單一語句之子頻段正規化所使用的重疊方式相同。

第 II 型：利用全部訓練語句求取單一子頻段的目標統計值時，各子頻段並不相互重疊，即求取目標統計值時，各子頻段的切割方式與原始方法相同。

根據以上的敘述，在固定子頻段數目的前提下，子頻段的目標統計值上的計算法有兩種選擇，子頻段重疊的樣式上有三種選擇，排列組合之下，可推演出新的分頻段正規化方法共有六類。爲了英文縮寫的表示上可以不至混淆，我們定義以下的表示式：

$$\text{SB正規化法名稱}_{(\text{分頻個數})}^{(\text{子頻段重疊方式})} - X$$

其中 SB 表示分頻段(sub-band)，正規化名稱可爲 SMN（頻譜平均值正規化法）、SMVN（頻譜平均值與變異數正規化法）與 SHE（頻譜統計圖等化法），分頻個數介於 4 到 6 之間，子頻段重疊方式爲 *lo*, *ro* 與 *to* 三種，而 X 可爲 I（第 I 型：目標統計值以重疊式子頻段計算）或 II（第二型：目標統計值以非重疊式子頻段計算）。舉例而言，$SBMVN_{(6)}^{(to)} - I$ 表示了分頻式平均值與變異數正規化法，其中分了 6 個子頻段，子頻段是左右重疊，且目標統計值是以重疊式子頻段方式計算。

以下爲我們所提出之重疊式分頻段調變頻譜統計正規化的步驟，同時，圖二爲此方法的流程圖：

1. 假設一段語音的倒頻譜特徵參數序列如下式(1)表示：

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

其中 M 爲一語音特徵向量中特徵的總數，N 代表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，才能讓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。將特徵序列 $\{x[n];1 \le n \le N\}$ 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(2)表示。

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

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 $\left[0, \frac{F_s}{2}\right]$；因 $X[k]$ 通常爲一複數，我們以極座標(polar form)表示 $X[k]$：

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

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 $\left\{A[k]\right\}$ 去做處理，而保留相位成份 $\left\{\theta[k]\right\}$ 不變。

2. 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)且重疊(overlapping)的方式，切割成 L 個頻段，依照我們所提出的三種重疊方法，每個頻段的範圍如下所示：

（1）向左重疊式的方法，每個頻段的範圍表示爲式(4)：

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

（2）向右重疊式的方法，每個頻段的範圍表示爲式(5)：

$$\begin{cases}\left[0,\left(\frac{1}{2^{L-1}}+\frac{1}{2^{L}}\right)\left(\frac{F_s}{2}\right)\right], & \text{if } \ell=1,\\ \left[\frac{2^{\ell-2}}{2^{L-1}}\left(\frac{F_s}{2}\right),\left(\frac{2^{\ell-1}}{2^{L-1}}+\frac{2^{\ell-1}}{2^{L}}\right)\left(\frac{F_s}{2}\right)\right], & \text{if } \ell=2,3,...,L-1,\\ \left[\frac{2^{\ell-2}}{2^{L-1}}\left(\frac{F_s}{2}\right),\frac{F_s}{2}\right], & \text{if } \ell=L.\end{cases} \quad \text{式(5)}$$

（3）雙邊重疊式的方法，每個頻段的範圍表示爲式(6)：

$$\begin{cases}\left[0,\left(\frac{1}{2^{L-1}}+\frac{1}{2^{L}}\right)\left(\frac{F_s}{2}\right)\right], & \text{if } \ell=1,\\ \left[\left(\frac{2^{\ell-2}}{2^{L}}+\frac{2^{\ell-2}}{2^{L+1}}\right)\left(\frac{F_s}{2}\right),\left(\frac{2^{\ell-1}}{2^{L-1}}+\frac{2^{\ell-1}}{2^{L}}\right)\left(\frac{F_s}{2}\right)\right], & \text{if } \ell=2,3,...,L-1,\\ \left[\frac{2^{\ell-2}}{2^{L-1}}\left(\frac{F_s}{2}\right),\frac{F_s}{2}\right], & \text{if } \ell=L.\end{cases} \quad \text{式(6)}$$

由以上數式可以得知，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；高頻的部分則被切割成較少的頻段，且每個頻段的長度較長，藉此可以強調出低頻的特性。在將$\{A[k]\}$作上述的頻段切割後，我們以$\{A_\ell[k']\}$表示其中的第ℓ個頻段。

3. 我們將上一步驟所得之不同頻段的強度頻譜$\{A_\ell[k']\}$作統計正規化處理。我們使用的正規化法分別爲：頻譜平均值正規化法(spectral mean normalization, SMN)、頻譜平均值與變異數正規化法(spectral mean and variance normalization, SMVN)與頻譜統計圖等化法(spectral histogram equalization, SHE)，處理後的特徵即以$\{\tilde{A}_\ell[k']\}$表示。詳細地說，SMN 在此的計算方式以下式(7)表示：

$$\tilde{A}_\ell[k'] = A_\ell[k'] - \mu_{\ell,s} + \mu_{\ell,a}, \quad \text{式(7)}$$

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。

SMVN 在此的計算方式以式(8)表示：

$$\tilde{A}_\ell[k'] = \left(\frac{A_\ell[k'] - \mu_{\ell,s}}{\sigma_{\ell,s}}\right)\cdot\sigma_{\ell,a} + \mu_{\ell,a}, \quad \text{式(8)}$$

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變異數，$\mu_{\ell,a}$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。

SHE 的計算方式以式(9)表示：

$$\tilde{A}_\ell\left[k'\right] = F_{\ell,a}^{-1}\left(F_{\ell,s}\left(A_\ell\left[k'\right]\right)\right), \qquad \text{式(9)}$$

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

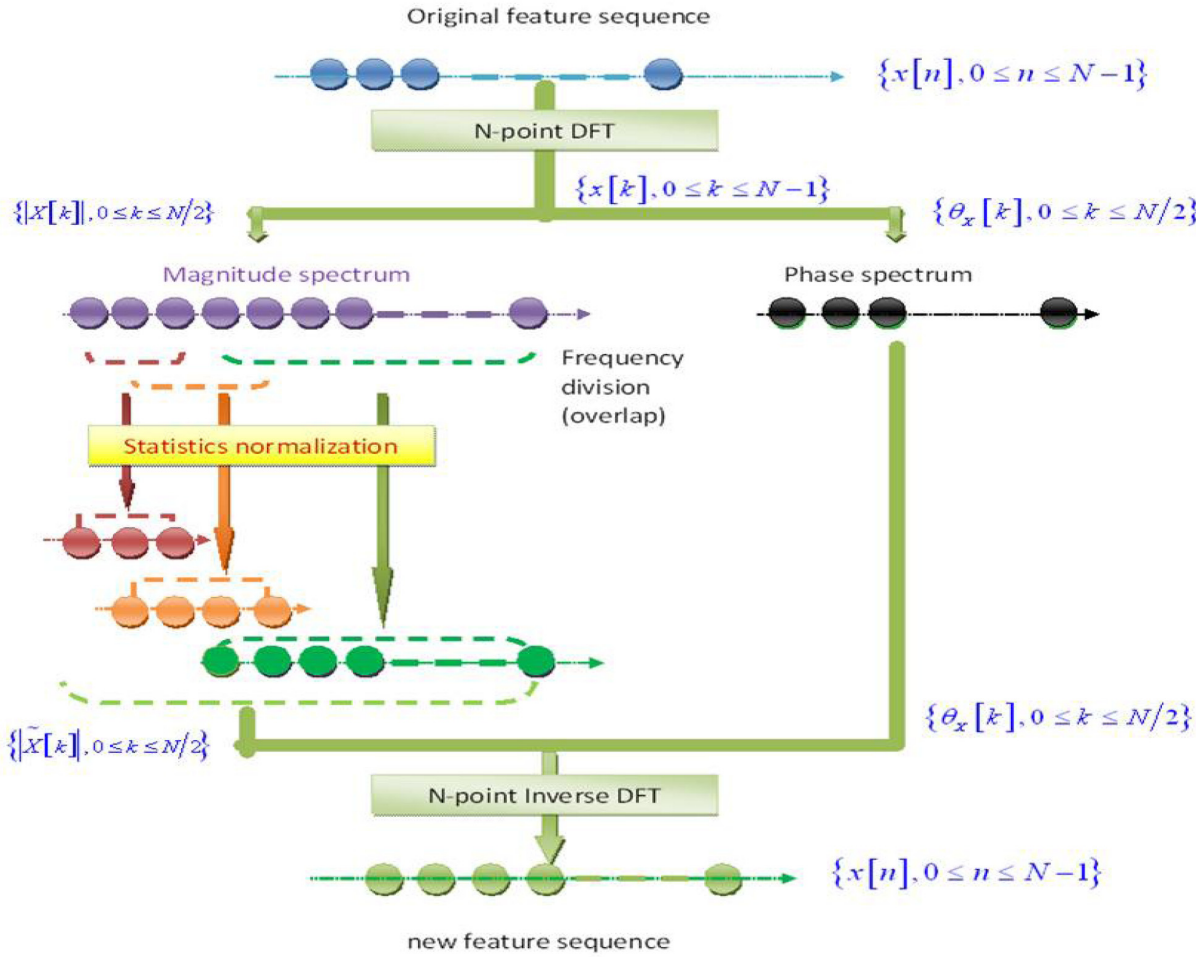
在此要附帶一提的是，式(7)-(9)中，$\left\{\tilde{A}_\ell\left[k'\right]\right\}$ 爲強度頻譜，所以其值必須爲非負的實數，但式(7)與(8)中的運算並無法保證 $\left\{\tilde{A}_\ell\left[k'\right]\right\}$ 必然滿足此條件，因此在我們實際操作時，若發 $\left\{\tilde{A}_\ell\left[k'\right]\right\}$ 之值小於零，則令其值爲零。

4. 在處理完每一頻段之後，我們將各頻段的強度頻譜 $\left\{\tilde{A}_\ell\left[k'\right]\right\}$ 照其頻率由小到大重新加權組合，得到新的全頻段強度頻譜 $\left\{\tilde{A}[k];0 \le k \le \left\lceil\frac{N}{2}\right\rceil\right\}$ (重疊部份的權重爲 0.5，非重疊部分之權重爲 1.0)。得到的結果就是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(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(10)表示：

$$\tilde{x}[n] = \frac{1}{N}\sum_{k=0}^{N-1}\left(\tilde{A}[k]e^{j\theta[k]}\right)e^{j\frac{2\pi nk}{N}}, \qquad 0 \le n \le N-1., \qquad \text{式(10)}$$

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(10)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 4 中，若我們欲利用全部訓練語句求取單一子頻段的目標統計值時，如同前面所述，若是第 I 型重疊式分頻段正規化法，則各子頻段重疊方式與執行單一語句之子頻段正規化所使用的重疊方式相同，相對地，第 II 型重疊式分頻段正規化法中，求取單一子頻段的目標統計值所用之各子頻段彼此並不重疊。



圖二：重疊式分頻段調變頻譜統計正規化法流程圖

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

在本論文中的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 Aurora-2 語音資料庫[9]，內容是以美國成年男女所錄製的一系列連續的英文字字串，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，分別爲：地下鐵(subway)、人聲(babble) 、 汽車(car)、展覽館(exhibition)、餐廳(restaurant)、街道(street)、機場(airport)、火車站(train station)等環境雜訊，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。Aurora-2 語音資料庫裡包含兩種不同的訓練環境和三種不同的測試環境(Set A、 Set B 與 Set C)，本論文所使用的是乾淨狀態訓練環境(clean condition training)； Set A 與 Set B 的語料只包含加成性雜訊，Set C 的語料同時包含加成性雜訊與通道效應。

實驗中所使用的原始特徵參數爲梅爾倒頻譜係數(MFCC)第零維至第十二維，附加上其

一階差量與二階差量，共 39 維；而使用的聲學模型爲由左向右(left-to-right)之隱藏式馬可夫模型(hidden Markov model, HMM)[10]，藉由隱藏式馬可夫模型訓練軟體 HTK[11]訓練所得，包括了 11 個數字模型(oh, zero, ... , nine)以及靜音模型(silence)，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) 實驗結果呈現與討論

首先，我們觀察本論文所提出之重疊式分頻段法所得到的實驗結果，表一、表二與表三分別呈現分頻段法中使用 SMN、SMVN 與 SHE 作爲正規化技術所對應的的平均辨識精確率，其中 $SBSMN_{(5)}$、$SBSMVN_{(5)}$與 $SHE_{(5)}$ 爲原始未重疊之分頻段正規化法，另外，AR 與 RR 分別代表了相對於基礎實驗（baseline，使用原始梅爾倒頻譜特徵與其一階及二階差量）的絕對錯誤降低率與相對錯誤降低率。在此，我們固定分頻段的 L 爲 5。

從表一、二與三中，我們可以觀察到以下幾點：

1. 相對於基礎實驗而言，各種重疊式分頻段正規化法，與原始非重疊式分頻段法類似，在語音辨識精確度上都得到十分明顯的提升。而值得一提的是，本論文所新提出的重疊式分頻段正規化法，無論是第 I 型或第 II 型，相對於原始非重疊式的分頻段法，大都能得到更進步的辨識率，此尤以 SMN（表一）與 SMVN（表二）特別明顯，且三種重疊方式（向左重疊、向右重疊與雙邊重疊）也幾乎都可改善辨識率。

2. 對於重疊式分頻段法而言，從表一與表二很明顯地看出，第 II 型幾乎都明顯優於第 I 型，如 SMN 法中第 II 型的 $SBSMN_{(5)}^{(lo)} - II$ (87.55%)比第 I 型的 $SBSMN_{(5)}^{(to)} - I$ (79.77%)有高達 7.77%的絕對辨識率的提升，此結果暗示了，即使使用最簡單的頻譜平均值正規化法(SMN)，若分頻段的方式得當，其效果仍十分顯著，幾乎不亞於較複雜的頻譜平均值與變異數正規化法(SMVN)，類似情形也發生在 SMVN 法上，其最佳辨識率（ $SBSMVN_{(5)}^{(to)} - II$ ：88.66%）趨近於更複雜的 SHE 法所能達到的辨識率。而第 II 型（目標統計值使用非重疊之子頻段計算）優於第 I 型（目標統計值使用重疊之子頻段計算）的可能解釋在於，我們使用重疊式分頻段的方法，主要是爲了彌補單一語句中分頻段之頻譜點數的不足，當將相鄰的頻段之頻譜也一同取進來求取統計值時，雖然頻譜點數增加了，但相對引進了非當下處理之子頻段的誤差。而目標統計值是利用『所有』訓練語句之分頻段的頻譜點，並沒有點數不足的問題，因此若如重疊式之第 I 型所得到的目標統計值，可能反而比原始非重疊式之第 II 型所得到的目標統計值之精確度差，而得到較差的辨識精確率。

3. 從表三之各式 SHE 法的辨識率來看，雖然重疊式分頻法仍大多數優於非重疊式分頻法，但進步幅度十分有限，最大的改善率僅有 0.21%（ $SBSHE_{(5)}^{(to)} - I$ ），且第 II 型也並無較第 I 型佳，各式 SHE 法大致都達到 90.5%左右的辨識精確率，推敲其可能原因，在於 SHE 本身已可達到相當不錯的效果，因此要有進一步提升，難度較大，另一方面，SMN、SMVN 與 SHE 其分別必須求取平均值、平均值與變異數與機率分佈，而如我們所知，更高階的統計量在估測上欲得到更佳的精確度，必須使用更多的取樣點，換言之，同樣數目的取樣點，對於提升較低階的統計量（如平均值）估測精確度明顯優於較高階的統計量（如變異數或機率分佈），因此，跟表一至三呈現的數據一致，我們在 SMN 能達到的改善程度，優於 SMVN 與 SHE。

表一、各種分頻式 SMN 法之平均辨識率(%)

	Set A	Set B	Set C	Average	AR(%)	RR(%)
baseline	71.92	68.22	77.61	71.58	—	—
$SBSMN_{(5)}$	77.62	76.10	81.19	77.73	6.15	21.63
$SBSMN_{(5)}^{(lo)} - I$	77.83	76.60	81.83	78.14	6.56	23.08
$SBSMN_{(5)}^{(ro)} - I$	77.95	76.55	81.07	78.01	6.44	22.64
$SBSMN_{(5)}^{(to)} - I$	79.48	78.93	82.03	79.77	8.19	28.82
$SBSMN_{(5)}^{(lo)} - II$	87.89	86.99	87.97	87.55	15.97	56.18
$SBSMN_{(5)}^{(ro)} - II$	85.81	86.11	85.44	85.86	14.28	50.24
$SBSMN_{(5)}^{(to)} - II$	86.73	86.04	86.96	86.50	14.92	52.50

表二、各種分頻式 SMVN 法之平均辨識率(%)

Method	Set A	Set B	Set C	Average	AR(%)	RR(%)
baseline	71.92	68.22	77.61	71.58	—	—
$SBSMVN_{(5)}$	87.36	88.53	88.03	87.96	16.38	57.65
$SBSMVN_{(5)}^{(lo)} - I$	87.88	89.09	88.52	88.49	16.91	59.51
$SBSMVN_{(5)}^{(ro)} - I$	87.09	88.26	87.71	87.68*	16.10	56.66
$SBSMVN_{(5)}^{(to)} - I$	87.32	88.52	88.08	87.95*	16.37	57.61
$SBSMVN_{(5)}^{(lo)} - II$	87.94	89.12	88.61	88.55	16.97	59.70
$SBSMVN_{(5)}^{(ro)} - II$	88.01	89.19	88.57	88.59	17.02	59.87
$SBSMVN_{(5)}^{(to)} - II$	88.09	89.17	88.78	88.66	17.08	60.10

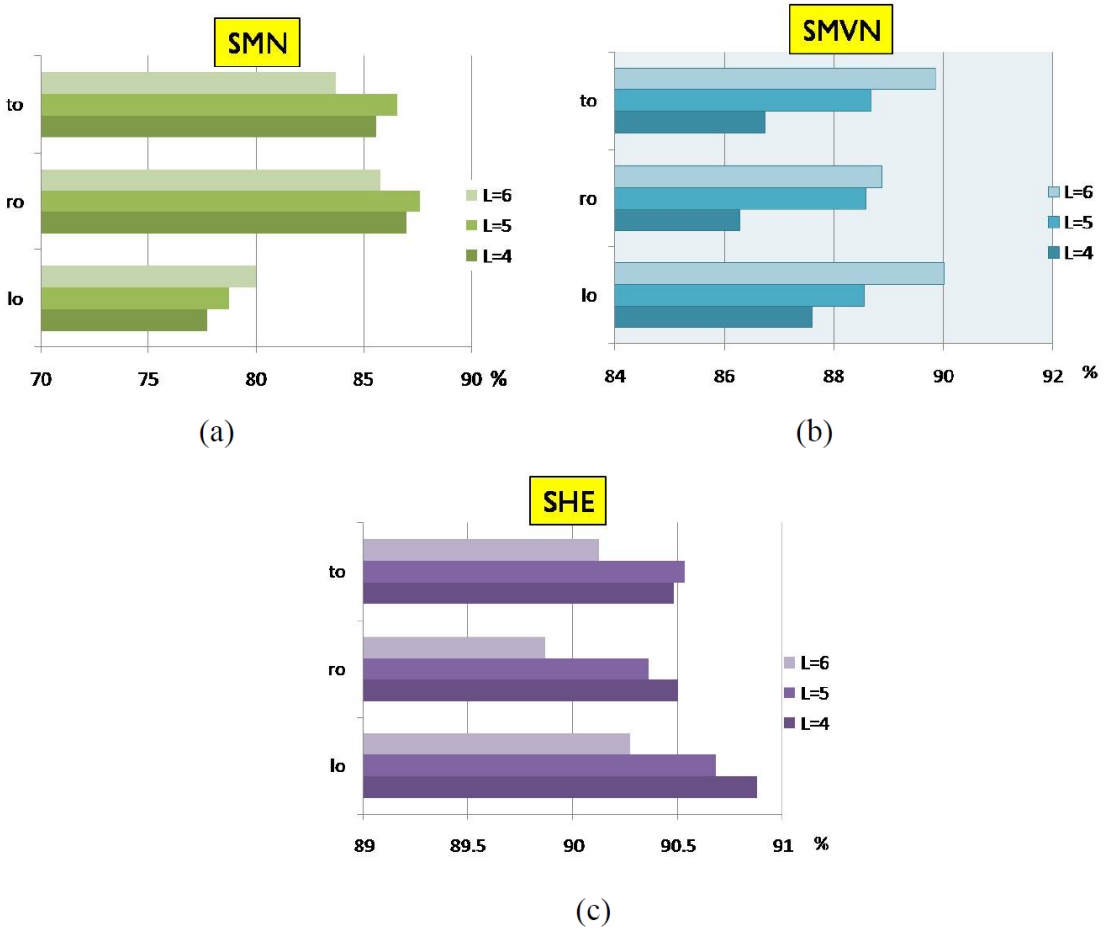
表三、各種分頻式 SHE 法之平均辨識率(%)

Method	Set A	Set B	Set C	Average	AR	RR
baseline	71.92	68.22	77.61	71.58	—	—
$SBSHE_{(5)}$	90.45	91.29	90.19	90.73	19.16	67.40
$SBSHE_{(5)}^{(lo)} - I$	90.66	91.44	90.33	90.91	19.33	68.00
$SBSHE_{(5)}^{(ro)} - I$	90.69	91.40	90.52	90.94	19.36	68.12
$SBSHE_{(5)}^{(to)} - I$	90.72	91.42	90.53	90.96	19.38	68.20
$SBSHE_{(5)}^{(lo)} - II$	90.49	91.28	90.26	90.76	19.18	67.49
$SBSHE_{(5)}^{(ro)} - II$	90.03	90.99	90.06	90.42	18.84	66.29
$SBSHE_{(5)}^{(to)} - II$	90.18	91.10	90.30	90.57*	18.99	66.83

接著，我們觀察當重疊方式與分頻段的數目變化時，對各種重疊式分頻段法之效能的影響，圖三(a)(b)(c)分別顯示了向左重疊(*lo*)、向右重疊(*ro*)及雙邊重疊(*to*)對應到分頻段數 L=4, 5 與 6 之三種正規化法(SMN, SMVN 與 SHE)的平均辨識率，由於之前已討論過，

第 II 型的效果皆優於第 I 型，這裡僅顯示第 II 型的結果。由圖三可看出：

1. 對於 SMN 而言，向右重疊(*ro*)及雙邊重疊(*to*)法似乎是較佳的方式，而分段數以 *L*=5 最好。
2. 對於 SMVN 而言，三種重疊法所得到的效果十分接近，但分段數以 *L*=6 最好，可以達到 90%左右的辨識率。
3. 對於 SHE 而言，各種重疊法與分段數所得之結果都十分相近，彼此辨識率的差距最大僅在 1%左右。



圖三：不同重疊式(*lo*, *ro* 與 *to*)與不同分頻段數目(*L*=4, 5, 6)之統計正規化法的平均辨識率，其中(a)爲 SMN 法(b)爲 SMVN 法(c)爲 SHE 法

最後，我們檢視本論文所提出的各種調變頻譜上的分頻式統計正規化法，與傳統時間序列域上的特徵統計正規化法是否具有加成性， 我們列舉了 $SBSMN_{(4)}^{(lo)}-I$ 法、$SBSMVN_{(4)}^{(lo)}-I$ 法與 $SBSHE_{(4)}^{(lo)}-I$ 法結合時域上特徵正規化法之實驗結果綜合整理成表四，表中四個處理技術：CMN、CMVN、MVA[12]、HEQ 分別爲倒頻譜平均值正規化法、倒頻譜平均值與變異數正規化法、平均值與變異數正規化結合自回歸動態平均濾波器法與統計圖等化法。從表四中，我們觀察到：

1. 重疊式調變頻譜正規化法結合時域上特徵正規化法與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： $SBSMN_{(4)}^{(lo)}-I$ 法結合 CMVN 的辨識率爲 89.47%，比起 CMVN 的辨識率 85.35%與 $SBSMN_{(4)}^{(lo)}-I$ 法的辨識率 78.99%，都有明顯的提升。唯獨在 $SBSHE_{(4)}^{(lo)}-I$ 法結合其他特徵正規化法，皆無法進一步提升辨識率，原因可能如前面所敘述，調變頻譜統計圖等化法本身已經具備很好的效能，因此就算結合時域上特徵正規化法，也無法明顯提升辨識率。

2. $SBSMN_{(4)}^{(lo)}-I$ 分別與 CMN、CMVN、MVA、HEQ 結合，辨識率幾乎達到 90%左右，這代表我們用相對簡單的一階統計正規化法 $SBSMN_{(4)}^{(lo)}-I$ 結合 CMN、CMVN、MVA、HEQ 即可達到十分突出的效果。

表四：時間序列域之統計正規化法與分頻式調變頻譜域之統計正規化法結合後所對應之平均辨識率(%)

Method		Set A	Set B	Set C	average	AR (%)	RR (%)
Baseline		71.98	67.79	78.28	71.56	—	—
CMN		79.37	82.47	79.90	80.72	9.15	32.18
CMN	$SBSMN_{(4)}^{(lo)}$ - I	89.38	90.25	88.76	89.60	18.04	63.44
	$SBSMVN_{(4)}^{(lo)}$ - I	88.15	89.45	88.77	88.79	17.23	60.59
	$SBSHE_{(4)}^{(lo)}$ - I	89.08	90.02	89.05	89.45	17.89	62.90
CMVN		85.03	85.56	85.60	85.36	13.79	48.50
CMVN	$SBSMN_{(4)}^{(lo)}$ - I	89.39	90.27	88.74	89.61	18.05	63.47
	$SBSMVN_{(4)}^{(lo)}$ - I	88.57	89.64	88.90	89.06	17.50	61.54
	$SBSHE_{(4)}^{(lo)}$ - I	89.01	89.97	88.95	89.38	17.82	62.66
MVA		88.12	88.81	88.50	88.47	16.91	59.46
MVA	$SBSMN_{(4)}^{(lo)}$ - I	89.42	90.34	88.91	89.69	18.12	63.73
	$SBSMVN_{(4)}^{(lo)}$ - I	88.65	89.68	89.20	89.17	17.61	61.92
	$SBSHE_{(4)}^{(lo)}$ - I	89.08	90.03	89.12	89.47	17.90	62.96
HEQ		86.91	88.32	87.50	87.59	16.03	56.37
HEQ	$SBSMN_{(4)}^{(lo)}$ - I	89.94	90.95	89.58	90.27	18.71	65.79
	$SBSMVN_{(4)}^{(lo)}$ - I	88.17	89.84	88.94	88.99	17.43	61.29
	$SBSHE_{(4)}^{(lo)}$ - I	89.16	90.35	89.36	89.68	18.11	63.69

在本章中，我們呈現了本論文所提出之各種新技術的實驗結果。在一系列的重疊式分頻段調變頻譜正規化法中，我們實驗了重疊式分頻段調變頻譜平均值正規化法、重疊式分頻段調變頻譜平均值與變異數正規化法及重疊式分頻段調變頻譜統計圖等化法，並且分成向左重疊式(*lo*)、向右重疊式(*ro*)、雙邊重疊式(*to*)等三種方式去做討論，試著要去補

償以不等切的方式切割特徵的調變頻譜，所造成的低頻頻段頻譜點數不足的問題。由實驗結果發現，對 SB-SMN 與 SB-SMVN 二者而言，重疊式相較於傳統的非重疊式的分頻段法幾乎都能明顯提升辨識率（無論第 I 型或第 II 型而言），但是 SB-SHE 法卻未能有類似的結果，可能的原因爲先前所提，SB-SHE 法相對另外兩個方法而言，額外處理了頻譜特徵之高階動差，已經擁有很高的辨識率，以至於即使利用重疊式的分頻法，無法有明顯的進步；另外，我們將所提之重疊式分頻段調變頻譜正規化法分別與時間序列域上的 CMN 法、CMVN 法、MVA 法及 HEQ 法作結合，由實驗結果發現，除了重疊式統計圖等化法之外，無論哪一種組合，皆能比單一方法更有效地提升辨識率，此驗證了特徵時間序列域與調變頻譜域上的正規化，彼此有明顯加成互補的效果。

四、結論與未來展望

在本論文中，我們提出了一系列重疊式分頻段調變頻譜統計正規化的演算法，以重疊的方式去改善不等切的調變頻譜頻段，再分別針對每個重疊式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，相對於傳統的分頻段而言，這些新方法有明顯的改進效果；我們也發現若欲求取每個子頻段的目標統計值時，仍然採用傳統的分頻段方法，所得的辨識效果更爲優越；另外，我們也將各種重疊式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，兩種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，我們所提出的重疊式分頻段之新方法，不僅能有效改善傳統的分頻段方法，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 。

對於未來的展望，雖然我們在論文中已經顯示重疊式分頻段調變頻譜統計正規化法的效能，但仍然有許多地方不夠完善，我們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重疊範圍。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，探討這一系列重疊式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背景雜訊環境(例如回音環境等)，進一步驗證這些新方法的效能與實用性，這些都是未來能夠嘗試研究發展的方向。

參考文獻

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

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

[3] 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.

[4] L. Sun, C. Hsu and L. Lee, "Modulation spectrum equalization for robust speech recognition," *IEEE Workshop on Automatic Speech Recognition Understanding* (*ASRU*), pp. 81-86, 2007.

[5] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. On Speech and Audio Processing*, pp. 578-589, 1994.

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

[7] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the importance of various modulation frequencies for speech recognition," *European Conference on Speech Communication and Technology* (*Eurospeech*), 1997.

[8] W. Tu, S. Huang and J. Hung, "Sub-band modulation spectrum compensation for robust speech recognition," *IEEE Workshop on Automatic Speech Recognition Understanding* (*ASRU* 2009), pp. 261-265, 2009.

[9] D. Pearce and H. Hirsch, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," *Proceedings of ISCA*, 2000.

[10] H. Stark and J. W. Woods, "Probability and random processes with applications to signal processing," *Prentice-Hall*, 2002.

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

[12] Chia-Ping Chen; J.A. Bilmes ”MVA Processing of Speech Features”, *IEEE Trans. On Audio, Speech and Language Processing*, vol. 15, no. 1, Jan 2007