機率式調變頻譜分解於強健性語音辨識
Probabilistic Modulation Spectrum Factorization for Robust Speech Recognition

朱紋儀 高予真 陳柏琳
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
{698470075, 699470424, berlin}@ntnu.edu.tw

洪志偉
國立暨南國際大學電機工程學系
Department of Electrical Engineering
National Chi Nan University
jwhung@ncnu.edu.tw

摘要

在自動語音辨識技術的發展上，語音強健性一直都是一門重要的研究議題。在眾多的強健性技術中，針對語音特徵參數進行強化與補償為其中之一大主要派別。其中，近年來已有為數不少的新方法，藉由更新語音特徵時間序列及其調變頻譜來提升語音特徵的強健性。本論文即是從語音特徵時間序列的調變頻譜域著手，採用機率式潛藏語意分析之概念，對調變頻譜施以機率式分解並進行成分分析、進而擷取出較重要的成分以求得更具強健性的語音特徵。本方法之所有實驗皆於國際通用的 Aurora-2 連續數字資料庫進行，相較於使用梅爾倒頻譜特徵之基礎實驗，本方法能達到 62.84%的相對錯誤降低率。此外，我們也嘗試將所提方法跟一些知名的特徵強健技術做結合；實驗顯示，相對於單一方法而言，此結合法可進一步提升辨識精確率，代表所提之新方法與許多特徵強健技術有良好的加成性。

關鍵詞： 雜訊強健性、語音特徵參數強化、調變頻譜、機率式潛藏語意分析

一、緒論

大部份的自動語音辨識(automatic speech recognition, ASR)系統，在不受雜訊干擾的理想實驗室發展環境下，皆可獲得良好的辨識效果；但若應用至真實的日常環境中，卻往往因為環境中諸多複雜因素的影響，造成系統之訓練環境與測試環境存在不匹配(mismatch)的問題，使得此系統之辨識精確率大幅度降低。而以上所述造成環境不匹配問題的種種因素包含了：語者發音結構差異、語者腔調變異、加成性背景雜訊、摺積性通道雜訊及其他語者發音的干擾等。所謂的語音辨識之強健性技術，即是致力於降低上述因素所帶來之影響，進而使語音辨識系統在不匹配問題存在的環境下，仍能保有一定的辨識能力。

目前而言，針對雜訊干擾的各種語音強健技術大致可分為三種類型：第一種類型為以聲學模型為基礎之強健性技術(model-based techniques)，其概念為以不變動語音特徵為原則，主要作用於聲學模型空間，期望藉由調整聲學模型之參數而能更正確地代表含環境雜訊之語音特徵；第二種類型為以語音特徵為基礎之強健性技術(feature-based techniques)，它主要作用於語音特徵空間，期望雜訊語音與其原始乾淨語音在此特徵表

示(speech feature representation)域上能趨於一致，藉此降低環境雜訊在語音特徵上所造成的不匹配效應；最後第三個類型為綜合式強健性技術(joint technique)，它同時考慮到上述兩種類型的技術，以達到結合特徵空間與模型空間之資訊為目的。

以語音特徵為基礎之強健性技術的其中之一個研究方向，是對於語音特徵參數之統計特性加以正規化；此方向涵蓋了著名的倒頻譜平均值減去法(cepstral mean subtraction, CMS)[1]、倒頻譜平均數與變異數正規化法(cepstral mean and variance normalization, MVN)[2]與統計圖等化法(histogram equalization, HEQ)[3]等，這些方法皆是直接將時間序列域(temporal domain)上的語音特徵視作為隨機變數(random variable)的樣本(samples)，利用這些樣本估測隨機變數的各樣統計值(statistics)，進而對語音特徵時間序列做線性或非線性的轉換，使其在部分或全部統計特性上能達到正規化的目標。

值得注意的是，環境中的干擾因素不僅會改變語音特徵之統計特性，同時也會引發語音特徵之時空結構(temporal structure)扭曲；而特徵參數時間序列之調變頻譜(modulation spectrum)為一有效描繪時空結構之媒介，相對於上述之語音特徵正規化法的觀念而言，可能具有更廣泛的分析面向，因其同時考慮到了語音特徵隨時間變化的性質(即各調變頻率之成分)。特別一提的是，過去的研究[4]顯示，不同調變頻率成分對語音辨識有著不同的重要性，位於 1 Hz 至 16 Hz 之調變頻率成分包藏了最有用的語意資訊，其中又以 4 Hz 附近的頻率成分特別突出。因此，近年來已有為數不少的學者致力於正規化特徵參數之時空結構，藉此直接或間接地強化語音特徵之調變頻譜，藉此提升語音特徵的雜訊強健性；相關的技術包括了調變頻譜統計圖等化法(spectral histogram equalization, SHE)[5]、分頻式調變頻譜統計正規化法(sub-band modulation spectrum compensation)[6]與一系列資料導向(data-driven)之時間序列濾波器法[7-10]等。

綜觀上述之技術，絕大多皆是藉由正規化時間序列或調變頻譜之統計特性，以降低語句間不匹配的程度，進而提昇語音辨識系統之強健性。本論文嘗試更進一步、以一個嶄新的觀點切入，利用機率式潛藏語意分析(probabilistic latent semantic analysis, PLSA)[11]賦予調變頻譜機率的意義，其透過一組潛藏的主題機率分布，描述調變頻率與調變頻譜強度成分之間的關係。因此，此利用機率式潛藏語意分析來觀察語音特徵的時空結構，可視為一種對於調變頻譜施以機率式分解並同時進行成分分析的方法。實作上，我們藉由機率式潛藏語意分析，從乾淨訓練語音特徵中萃取出一組潛藏的主題機率分布，以利而後任一句乾淨或雜訊語句更新其強度調變頻譜時使用，進而達到強化語音特徵之調變頻譜之目的。

在一系列的語音辨識實驗上，我們發現上述的新方法可以顯著地提升原始語音特徵在雜訊環境下的精確率，其效能等同甚至超越現行許多強健性技術，足見此新方法不僅在理論上具有嶄新的意義、在應用上也有其實際顯著的價值。

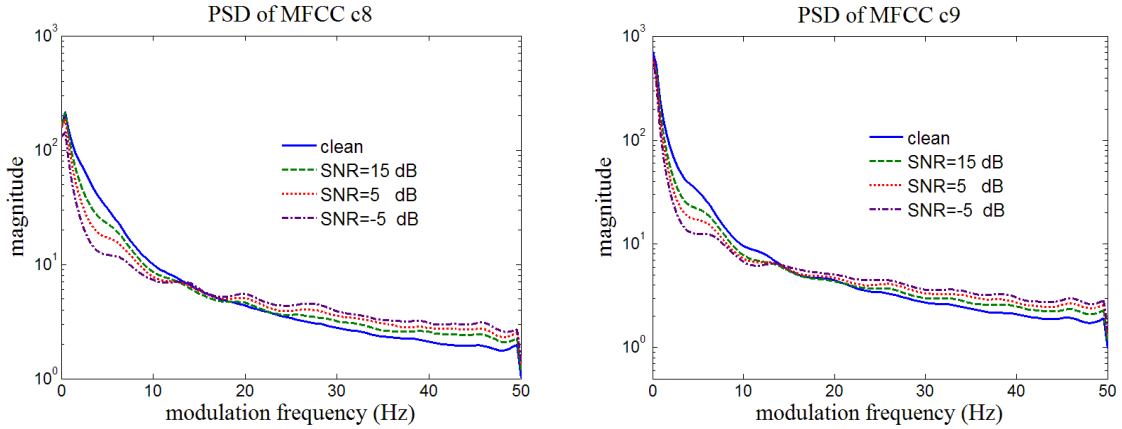
二、正規化時間序列結構特性之方法

（一）語音特徵之調變頻譜受雜訊干擾之影響情形

對於一語音特徵時間序列 $\{x[n]\}$ 而言，其調變頻譜定義如下：

$$X[k] = DFT(x[n]), \tag{1}$$

其中，n 與 k 依序為音框索引與調變頻率索引，DFT 為離散傅立葉轉換(discrete Fourier transform)。式(1)之頻譜序列可視為一種對於原始語音訊號作降低取樣(down-sampled)後的調變訊號(由訊號取樣頻率轉至音框取樣率)，此序列即為所屬語音特徵時間序列之調變頻譜(modulation spectrum)。由式(1)可知，調變頻譜 $X[k]$ 之最高頻率與特徵序列 $x[n]$



(a)c8　　(b)c9

圖一、梅爾倒頻譜參數於乾淨與三種訊噪比情況之功率頻譜密度圖

之取樣頻率(音框取樣率)相關。例如，在一般設定下，音框取樣率為 100 Hz，則最高調變頻率為 50 Hz。

過去已有不少學者投注心力於語音特徵之調變頻譜特性之研究，且大多研究不約而同地顯示，調變頻譜之低頻成分(約 1 Hz 至 16 Hz)對於語音辨識精確度有顯著的關連，而其中尤以 4 Hz 的成分最為重要。有趣的是，4 Hz 也是人耳聽覺最為敏感之調變頻率[12]。此外有一假說，4 Hz 為人類大腦皮層感知之重要調變頻率[13]。當語音訊號受到噪音干擾時，不僅會使其時間特徵時間序列產生失真，同時其調變頻譜也會因而改變。在此，我們用一簡例來說明雜訊對語音調變頻譜之強度(magnitude part)產生的失真。首先，我們求取語音之梅爾倒頻譜參數(Mel-frequency cepstral coefficients, MFCC)於乾淨與不同訊噪比(signal-to-noise ratio, SNR)情況下之調變頻譜，其次，值得注意的是，因為不僅有雜訊會影響調變頻譜之值，尚有其它干擾因素，如語句之說話內容及語者特性等。因此，為降低雜訊以外的其他因素，在此我們要觀察的調變頻譜強度，是經由 1,688 句語句(出自 Aurora-2 語音資料庫[16])之倒頻譜特徵的調變頻譜強度平均而得。圖一中的曲線為梅爾倒頻譜參數於乾淨與三種訊噪比情況之平均調變頻譜強度；其中，圖一(a)對應至第八維梅爾倒頻譜參數 c8，圖一(b)則對應至第九維梅爾倒頻譜參數 c9。從圖一(a)(b)，我們首先觀察到調變頻譜之強度皆較集中於低頻，呼應了前人之發現，即語音特徵的調變頻譜強度主要都集中於低頻成分。其次，若將乾淨特徵與含雜訊特徵之調變頻譜強度加以比較，可明顯看出雜訊對整個調變頻帶都造成失真，其中低頻之強度會因此下降，而高頻之強度則反而上升，兩者之臨界約在 15 Hz 至 20 Hz 之間。最後，若比較不同訊噪比(SNR)之對應曲線，我們發現隨著雜訊的比例升高，調變頻譜強度於低頻下降與高頻上升之幅度也隨之加劇；意謂著雜訊對於調變頻譜之影響為使之整體分布趨於平坦，這與過去一些相似研究之觀點大致相同[7, 8, 17]。

（二）特徵參數之調變頻譜之強健化的相關研究介紹

目前對於調變頻譜改進其雜訊強健性(noise robustness)之技術，大多是對式(1)之 $X[k]$ 其強度成分 $|X[k]|$ 作更新，並保留其相位成分 $\theta[k]=\angle X[k]$ 不變。更新後的強度成分與原始相位成分相結合後，經由反傅立葉轉換(inverse discrete Fourier transform, IDFT)以求得新語音特徵時間序列。若上所述之調變頻譜的強度能夠被適當地更新，則將可有效降低雜訊產生的失真，進而讓使用新的語音特徵的語音辨識系統獲得較佳的辨識率。以下，我們

將簡述幾種更新調變頻譜的演算法，這些方法皆被初步驗證能有效提升語音特徵的雜訊強健性。

1、調變頻譜統計圖法(spectral histogram equalization, SHE)

此項技術[5]是將圖像辨識(pattern recognition)常用的統計圖等化演算法(histogram equalization, HEQ)應用於語音特徵調變頻譜強度的更新上，利用一非線性的轉換(nonlinear transform)，使訓練語句與測試語句的調變頻譜強度趨於同一個機率分布函數(probability distribution function, PDF)。在此方法中，新調變頻譜強度$\left|\tilde{X}[k]\right|$與原始強度$|X[k]|$之關係為

$$\left|\tilde{X}[k]\right| = F_{ref}^{-1}\left(F_X\left(|X[k]|\right)\right) \tag{2}$$

其中，$F_X(\cdot)$為單一語句$\{x[n]\}$之調變頻譜強度的機率分布，而$F_{ref}(\cdot)$則為集合所有訓練語句之調變頻譜強度所求的參考機率分布。

上述之 SHE 法可將特徵的調變頻譜強度作非線性的轉換，進而使其機率分布正規化，與其相關連的方法包括了(調變)頻譜平均正規化法(spectral mean normalization, SMN)[6]及頻譜平均與變異數正規化法(spectral mean and variance normalization, SMVN)[6]等。這兩種方法利用了線性轉換(linear transform)，分別對調變頻譜強度之平均值、或平均值與變異數加以正規化，類似 SHE 的觀念，SMN 與 SMVN 同樣可將乾淨語音特徵與雜訊語音特徵之調變頻譜強度之間的不匹配降低，進而提升特徵的雜訊強健性。

2、分頻段調變頻譜統計正規化法

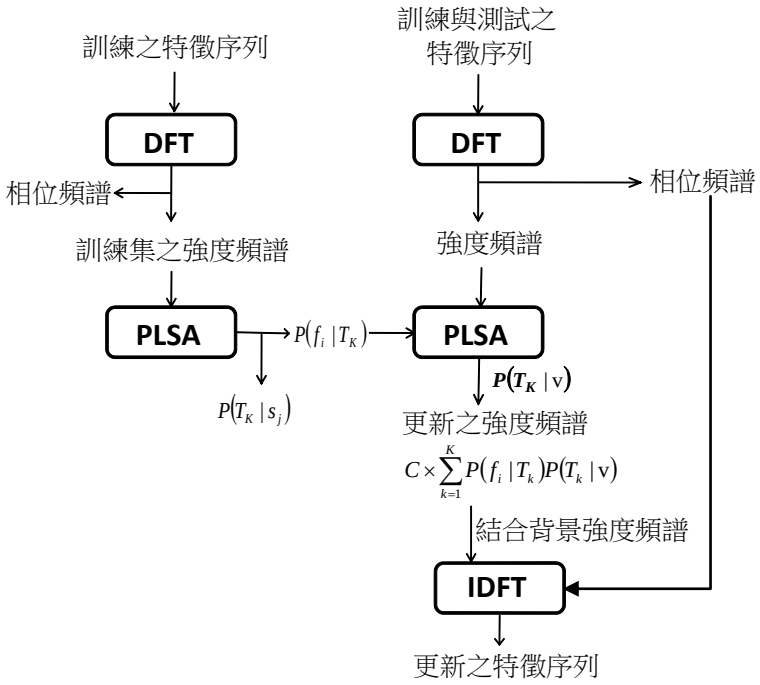
前面所提到的 SHE, SMN 與 SMVN 三種方法，是將全部調變頻帶之頻譜強度值視為同一隨機變數(random variable)的樣本(samples)，進而一齊作正規化。然而，如前所述，不同調變頻率的成分在語音辨識中存在不等價的重要性，低頻成分比高頻成分相對重要。因此，文獻[6]提出將調變頻帶切割成多段的子頻段，再分別對每一個子頻段的頻譜強度作統計值(如先前所提的平均值、變異數或統計圖)正規化處理，而為了強調較低調變頻率的重要性，在低頻部分，子頻段的頻寬較細、子頻段的數目較多，高頻部分則是相反。根據文獻[6]的實驗數據顯示，分頻段正規化相對於全頻帶正規化而言，可以得到更佳的辨識率，然而，其計算複雜度與所需記憶體空間也較大。

3、時間序列結構正規化法(TSN)

時間序列結構正規化法(temporal structure normalization, TSN)[7] 是屬於一種時間序列濾波器(temporal filter)設計之技術，其藉由語音特徵序列通過一事先設計之濾波器，以達到正規化調變頻譜之目的。茲將TSN法所使用的濾波器設計步驟簡述如下：

STEP 1: 將訓練語料庫中，所有乾淨語音特徵序列(對單一種類之特徵而言)對應之功率頻譜密度(power spectral density, PSD)作平均，此平均視為參考功率頻譜密度，以$\{\bar{P}_{SS}[k]\}$表示，其中k為頻率索引。

STEP 2: 對訓練與測試語料庫中，求取個別語音特徵序列之功率頻譜密度，以$\{P_{XX}[k]\}$表示，則濾波器的頻率響應(frequency response)定為：



圖二、以機率式潛藏語意分析為基礎之調變頻譜正規化法之程序

$$H[k] = \sqrt{\frac{\overline{P}_{SS}[k]}{P_{XX}[k]}} \tag{3}$$

STEP 3:將式(3)做反離散傅立葉轉換(IDFT)，所得的序列先後經過窗化(windowing)與直流增益(DC gain)正規化後，最後所得的序列即為TSN所用的濾波器之脈衝響應(impulse response)，以 $\{h[n]\}$ 表示。

值得注意的是，上述的濾波器頻率響應 $\{h[n]\}$ 是隨不同特徵序列而改變(因為式(3)裡的 $\{P_{XX}[k]\}$ 是個別特徵序列的PSD)，個別特徵序列通過其對應的TSN濾波器後，新特徵序列的PSD會逼近於參考PSD，由於PSD可視為平緩化後(smoothed)的調變頻譜強度平方，故TSN的目標相當於將語音特徵序列的調變頻譜強度一致化，藉以降低因雜訊干擾在調變頻譜強度造成的變異。

三、以機率式潛藏語意分析為基礎之調變頻譜正規化法

本論文嘗試機率式潛藏語意分析(probabilistic latent semantic analysis, PLSA)[11]應用於調變頻譜處理，其是一種使用機率模型的方式，找出調變頻譜強度與不同語音特徵序列之間的主題資訊。PLSA 可被視為是一種觀點模型(aspect model)的分析，其透過一組隱藏變數的機率分布，以共同預測一事件發生的可能性，而此組隱藏變數，即可被喻為一組潛藏主題。當我們使用 PLSA 來更新語音特徵時間序列的調變頻譜強度時，其流程圖如圖二所示，而詳細步驟陳述如下：

（一）藉由乾淨語音特徵序列之調變頻譜強度，求取其對應的 PLSA 生成模型

我們使用 PLSA 的觀念，為每一句乾淨訓練語句之特徵序列的調變頻譜強度建立生成模型，其透過一組共享的潛藏主題機率分布，以描繪每一語音特徵序列與其調變頻譜強度

的對應關係，首先，我們建立一關係矩陣 V，其每一行(column)是個別訓練語句特徵序列之調變頻譜強度，長度為 L，若共有 M 句訓練語句特徵序列，則 V 為 $L \times M$ 的矩陣，通常 $M \gg L$，接下來，我們將矩陣 V 近似為兩個矩陣之乘積[14]：

$$\mathrm{V} \approx \mathrm{G}\mathrm{H}^{\mathrm{T}} \tag{4}$$

其中 G 與 H^{T} 分別為 $L \times K$ 與 $K \times M$ 的矩陣，而 K 即為 PLSA 中預設的潛藏主題個數。在這兩個矩陣裡：

(1) 矩陣 G 的第 i 列(row)的向量，表示為第 i 個調變頻率之強度成分(以 f_i 表示)在不同潛藏主題中生成的機率值，

(2) 矩陣 H^{T} 的第 j 行(以 s_j 表示)則是表示第 j 個語句產生不同潛藏主題的主題機率分布。更明確地說明，關係矩陣 V 中的每一個元素 $a_{i,j}$ 被近似為：

$$a_{i,j} = P(f_i \mid s_j) \approx \sum_{k=1}^{K} P(f_i \mid T_k) P(T_k \mid s_j) \tag{5}$$

即為個別序列 s_j 透過潛藏主題分布估算產生調變頻譜強度 f_i 的機率值。觀察上式可知，機率式潛藏語意分析有兩大類的模型參數需要估算，分別為每一個調變頻譜的主題機率分布 $P(T_K \mid s_j)$ 與各主題生成調變頻譜強度的機率分布 $P(f_i \mid T_K)$；而這些參數則可經由最大化訓練語句中每一個調變頻譜之對數相似度(log-likelihood)，並以期望值最大化法(expectation-maximization, EM)求得。值得一提的是，已有研究證實在適當的設定與推導下，機率式潛藏語意分析與非負矩陣分解(nonnegative matrix factorization, NMF)為等價的概念[15]，而非負矩陣分解則是一項已被廣泛運用於影像處理的演算法。藉由文獻[15]所介紹的演算法，我們可以求得上述 PLSA 的兩大類參數 $\{P(T_K \mid s_j)\}$ 與 $\{P(f_i \mid T_K)\}$。

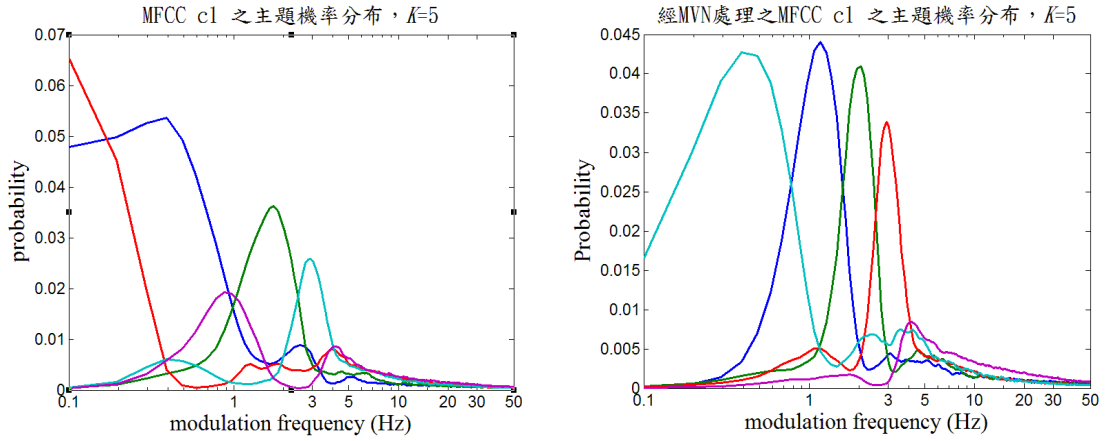
（二）利用 PLSA 生成模型參數，重建語音特徵序列之調變頻譜強度

在這一步驟中，無論是測試語句或是訓練語句，其原始調變頻譜強度(以 v 表示)，皆經由上一步驟所得之機率式潛藏語意分析的兩個機率分布 $P(f_i \mid T_K)$ 與 $P(T_K \mid s_j)$，進行重新估算(新的調變頻譜強度，在此以 $\tilde{\mathrm{v}}$ 表示)。其中必須注意的是，上述之 $P(T_K \mid s_j)$ 是乾淨訓練特徵序列其調變頻譜的主題機率分布，因此對於測試特徵序列之調變頻譜 v 而言，其主題機率分布是未知的。在這裡，為了使訓練語句與測試語句皆通過相同的處理、降低可能的失真，我們對於任一調變頻譜強度 v，利用以下公式進行估算其主題機率分布 $P(T_K \mid \mathrm{v})$：

$$P(T_K \mid \mathrm{v}) = \frac{\sum_{i=1}^{L+1} f_i \cdot h(T_k \mid f_i, \mathrm{v})}{\sum_{j=1}^{L+1} f_j} \tag{6}$$

其中

$$h(T_k \mid f_i, \mathrm{v}) = \frac{P(f_i \mid T_k) P(T_k \mid \mathrm{v})}{\sum_{l=1}^{K} P(f_i \mid T_l) P(T_l \mid \mathrm{v})} \tag{7}$$



圖三、PLSA 對於(a)原始 MFCC 之 c1 (b)MVN 處理後 MFCC 之 c1 所求得的五個主題機率分布頻譜強度

在得到調變頻譜 v 的機率分布 $P(T_K \mid \mathrm{v})$，並配合原有的 $P(f_i \mid T_K)$，我們即可估算初步更新之每一維調變頻譜 $\tilde{\mathrm{v}}_i$，如下式所示：

$$\tilde{\mathrm{v}}_i = C \times \sum_{k=1}^{K} P(f_i \mid T_k) P(T_k \mid \mathrm{v}) \tag{8}$$

其中 C 為原始調變頻譜強度 v 每一維 v_i 的和，即 $C = \sum_{i=1}^{L} \mathrm{v}_i$ 。

此外，在實際操作上，由於原始機率式潛藏語意分析運用於語言模型時，皆會使用模型插補法，將其與背景模型相結合；在此我們採用相同的概念，將乾淨語音特徵之調變頻譜強度之平均作為背景調變頻譜強度(在此以 u 表示)，再利用插補法將其與式(8)之 $\tilde{\mathrm{v}}$ 做線性組合，得到最終之新的每一維調變頻譜強度 $\hat{\mathrm{v}}_i$，如下式所示：

$$\hat{\mathrm{v}}_i = \alpha \mathrm{u}_i + (1-\alpha)\tilde{\mathrm{v}}_i \tag{9}$$

其中 α 為加權值。

最後，我們將更新後之調變頻譜強度與原始調變頻譜相位做組合，並經由反傅立葉轉換(inverse DFT, IDFT)，將其轉換成新的特徵序列。

關於上述以 PLSA 為基礎之更新調變頻譜強度的演算法，有下列二項實作層面上的細節需注意，其描述如下：

(1) 儘管特徵時間序列之長度因語句而異，但是在此我們將其調變頻譜之長度(即其取離散傅立葉轉換的點數)設為定值，因此所有語句之調變頻譜長度皆相同，此外需注意的是，此定值需大於或等於任一待處理之特徵時間序列的長度，以避免時間混疊(time aliasing)的不良效應。

(2) 對更新後的調變頻譜進行反傅立葉轉換後，所得之序列長度會大於或等於原始特徵序列的長度(假設為 N)，因此我們只保留此新序列的前 N 點，作為最終的新特徵序列，此作法是根據最小化平方差(minimum mean squared error, MMSE)的最佳準則而得。

在圖三(a)與三(b)中，我們繪製了由以上 PLSA 法所得到的五個隱藏主題對應之調變頻譜強度(即等式(4)中矩陣 G 的行向量)，圖三(a)是對應原始 MFCC 之 c1 特徵，圖三(b)則是對應經 MVN 處理後 MFCC 之 c1 特徵。這兩圖都顯示了，PLSA 所得之潛藏主

題調變頻譜強度都集中在低頻成分(大約 10 Hz 以下)，如前所述，這區域正是重要語音資訊匯集之處，顯示了 PLSA 可以有效將語音特徵序列重要的調變頻譜成分擷取出、並抑制不重要或容易受干擾的中高頻成分。而圖三(a)與三(b)的主要差別，在於後者的多數主題頻譜強度其極低頻之近直流成分(DC)很小，這是因為 MVN 處理後的 MFCC 特徵，其直流成分為零，PLSA 法所得的主題頻譜強度忠實地反映了這個前提。

四、實驗結果與分析

（一）實驗語料庫

本論文所使用的實驗語料庫為 Aurora-2 英文連續數字語料庫[16]，參與錄音計畫的語者皆是美國成年人。為了評估雜訊或通道對於語音的影響，測試部分的語音分別摻有八種不同來源的加成性雜訊(additive noise)和兩種不同特性的通道效應。根據不同種類的干擾，分成三個測試集：Set A, Set B 與 Set C。Set A 的語音分別含有地下鐵(subway)、人聲(babble)、汽車(car)和展覽會館(exhibition)等四種加成性雜訊與 G.712 通道效應； Set B 的語音則分別含有餐廳(restaurant)、街道(street)、機場(airport)和火車站(train station)等四種加成性雜訊與 G.712 的通道效應；Set C 分別加入了地下鐵(subway)與街道(street)兩種雜訊與 MIRS 通道效應。其中，而其中的訊噪比則有七種，分別為 clean (∞dB)、20 dB、15 dB、10 dB、5 dB、0 dB 和-5 dB。Aurora-2 資料庫提供兩種訓練聲學模型的模式：乾淨情境訓練模式(clean-condition training)與複合情境訓練模式(multi-condition training)，本論文統一使用乾淨語料訓練模式來進行實驗，訓練集的乾淨語音共有 8,440 句，其中並無加成性雜訊，卻包含了 G.712 的通道效應，因此在三個測試集中，訓練集只與測試集 Set C 有通道上的不匹配。

（二）實驗設定

在前端處理方面，本論文的基礎實驗是採用梅爾倒頻譜係數做為語音特徵參數，其中預強調(pre-emphasis)參數設為 0.97，視窗函數為漢明窗(Hamming window)，取樣音框長度(frame length)為 25 毫秒，音框間距(frame shift)為 10 毫秒，每個音框是以 39 維特徵向量表示，其中包含 12 維的梅爾倒頻譜係數(c_1~c_{12})與第零維倒頻譜係數(c_0)，附加上其第一階差量係數(delta coefficient)和第二階差量係數(acceleration coefficient)。

在聲學模型的設定上，每個數字模型(one, two, …, nine, zero 和 oh)皆由一個由左到右(left-to-right)形式的連續密度隱藏式馬可夫模型(continuous density hidden Markov model, CDHMM)表示，其中包含 16 個狀態(state)，每個狀態則有 20 個高斯混合(Gaussian mixtures)。靜音模型則為 1 個狀態，內含 36 個高斯混合的模型。上述所有聲學特徵的建立、聲學模型的訓練與各種辨識實驗都是使用 HTK 工具套件[18]完成。

（三）辨識效能評估方式

辨識效能評估的方式是採用美國標準與科技組織(the national institute of standards and technology，NIST)[19]所訂立的評估標準，進行正確轉譯文句字串與辨識字串的比較。評估單位是以詞精確率(word accuracy)為單位，計算正確轉譯文句字串與辨識字串間的詞取代個數(substitutions)和詞插入個數(insertions)；計算公式如下所示：

$$\text{詞精確率}(\%) = \frac{\text{詞正確辨識個數} - \text{詞插入個數}}{\text{輸入詞總數}} \times 100\% \qquad (10)$$

值得注意的是，根據原 Aurora-2 資料庫的設定，每一種雜訊的平均詞精確率計算方式

表一、 PLSA 法作用於原始 MFCC 特徵的辨識結果，其中 Avg(%)與 RR(%)分別為總平均辨識精確率與相對錯誤降低率。

平均詞精確率(%)		Clean	Set A	Set B	Set C	Avg.	RR
MFCC baseline		99.79	72.46	68.31	78.82	72.07	－－
PLSA	*K*=5	99.56	89.20	90.20	89.41	89.62	62.84
	K=10	99.59	89.05	90.25	89.25	89.57	62.66
	K=15	99.61	88.81	90.15	88.87	89.36	61.90
	K=20	99.59	88.78	90.18	88.69	89.32	61.76

是對於 20 dB 至 0 dB 的五種訊噪比(SNR)辨識率取平均，而排除掉乾淨情況和-5dB 二種極端的訊噪比的辨識率；本論文後續的所有平均辨識率皆是遵循此種呈現方式。

（四）實驗結果呈現與討論

1、PLSA 法作用於原始 MFCC 所得之辨識率

我們將所提出的 PLSA 法對於原始 39 維 MFCC 語音特徵參數時間序列做處理，其對應的平均辨識精確率詳列於表一之中；在 PLSA 法的參數設定上，我們令潛藏主題個數 K 分別為 5,10, 15 與 20，而式(9)中的加權值 α 則預設為 0.85。從表一的數據中，我們有以下幾點發現：

(1) 在匹配的乾淨環境下，相較於基礎實驗結果，PLSA 法對應的辨識精確率略為下降，但其下降程度並不顯著(最大下降率為 0.23%)，且跟選擇隱藏主題個數並無明顯關係。此現象驗證了，以 PLSA 為基礎的模型足以充分呈現語音特徵之調變頻譜強度的特性。對於調變頻譜強度而言，PLSA 為一種高效能編碼(encoding)的方式，其中只需使用少量之隱藏主題，就足以保有語音特徵之調變頻譜強度內含的辨識資訊。

(2) 在不匹配的雜訊干擾環境下，PLSA 處理後之語音特徵其表現明顯優於原始語音特徵。跟基礎實驗結果比較，在使用主題數為 5(K =5)的 PLSA 法時，辨識精確率被提升了 17.55%，相對錯誤降低率高達 62.84%，而其它主題數的 PLSA 法也有十分類似的效能。因此，我們所提出的 PLSA 法確實能有效提升原始梅爾倒頻譜特徵之雜訊強健性。此外，回顧圖三所顯示之主題機率分布對應之頻譜強度，可推知藉由 PLSA 的處理，對應至非語音成分失真之高頻調變頻譜成分會被大幅縮減，因而得到辨識精確率的進步。

(3) 相較於乾淨環境情況下，在雜訊干擾環境中增加主題機率分布數量 K，辨識精確率反而會微幅下降。然而跟乾淨環境之情形類似，不同主題數之 PLSA 其所造成的辨識精確率差距並不明顯，平均而言最大差距僅 0.30%(從 89.62%下降至 89.32%)。

2、PLSA 法結合其他強健性特徵演算法所得之辨識率

其次，我們將原始語音特徵先經過倒頻譜平均消去法(CMS)[1]或倒頻譜平均與變異數正規化法(MVN)[2]處理後，再透過我們所提出的 PLSA 法加以處理，藉此觀察 PLSA 與 CMS 或 MVN 這兩種典型的特徵序列處理技術是否有加成性，其所對應的辨識精確率分別列於表二與表三。觀察這兩個表的數據、並與表一比較，我們可知：

表二、 PLSA 為基礎之方法作用於經 CMS 處理之 MFCC 特徵的辨識結果，其中 RR_1 (%)與 RR_2 (%)分別為對比於基礎實驗與 CMS 法之辨識率之相對錯誤降低率。

平均詞精確率 (%)		Clean	Set A	Set B	Set C	Avg.	RR_1	RR_2
MFCC baseline		99.79	72.46	68.31	78.82	72.07	—	—
CMS		99.82	79.31	82.46	79.90	80.69	30.86	—
PLSA+ CMS	*K*=5	99.66	89.70	91.09	90.00	90.32	65.34	49.93
	K=10	99.69	89.63	91.00	89.89	90.23	65.02	49.87
	K=15	99.71	89.49	90.89	89.78	90.11	64.59	48.78
	K=20	99.73	89.38	90.92	89.70	90.06	64.41	48.52

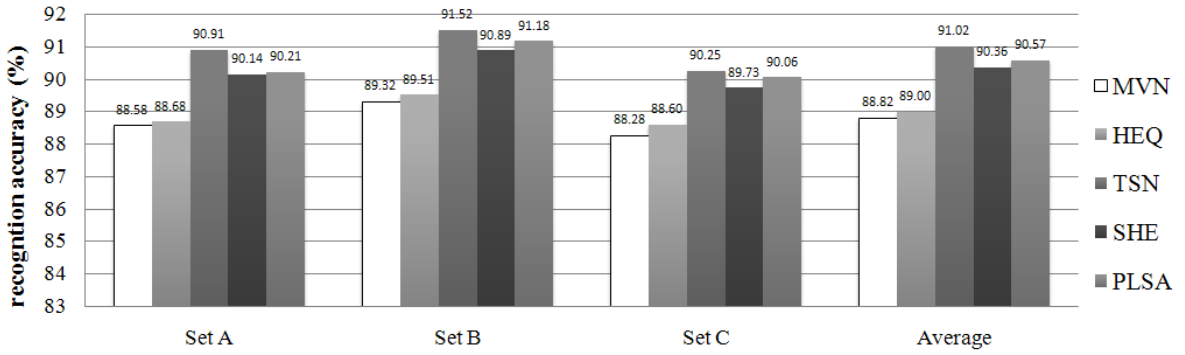
表三、 PLSA 為基礎之方法作用於經 MVN 處理之 MFCC 特徵的辨識結果，其中 RR_1 (%)與 RR_2 (%)分別為對比於基礎實驗與 MVN 法之辨識率之相對錯誤降低率。

平均詞精確率 (%)		Clean	Set A	Set B	Set C	Avg.	RR_1	RR_2
MFCC baseline		99.79	72.46	68.31	78.82	72.07	—	—
MVN		99.82	88.58	89.32	88.28	88.82	59.97	—
PLSA+ MVN	*K*=5	99.66	89.98	91.01	89.72	90.34	64.70	13.60
	K=10	99.68	90.06	91.06	89.90	90.43	65.74	14.40
	K=15	99.73	90.07	91.14	89.91	90.47	65.88	14.76
	K=20	99.72	90.21	91.18	90.06	90.57	66.24	15.54

(1) 相對於使用原始 MFCC 特徵之基礎實驗而言，CMS 與 MVN 皆能改善辨識精確率，其中又以 MVN 的改進效果較好，可提供高達 17%左右的精確率提升。而我們所提的 PLSA 法，在四種潛在主題數的選擇下，皆優於 MVN 法。

(2) 當 PLSA 法與 CMS 結合時，相較於單一 PLSA 法或單一 CMS 法而言，都能使辨識精確率更有效的提升，此進步的現象也同樣發生於 PLSA 法與 MVN 的結合上。整體平均辨識率都可超過 90%，另外，在結合 PLSA 法的前提下，CMS 與 MVN 的表現差距很小(辨識精確率的差距僅有 0.5%左右)，這代表了在 PLSA 法前置處理方法的選擇上，我們可使用簡易的 CMS 法，即可趨於較複雜的 MVN 達到的效能。

(3) 跟表一呈現的數據類似，在結合 CMS 或 MVN 後的 PLSA，其潛在主題數目的多寡與辨識精確率並無顯著的關係。而改變潛在主題數目所造成的平均精確率變化皆在 0.3%以下，這也顯示了我們可以使用很少的潛在主題(如 *K*=5)，也就是說在簡化運算複雜度的前提下亦不影響 PLSA 法的優異性。



圖四、PLSA 法與其它調變頻譜更新法的效能比較(經 MVN 前置處理過)

3、PLSA 法與其他調變頻譜更新法的效能比較

在這一節中，我們將所提出的 PLSA 法，與一系列的語音特徵時間序列處理技術進行辨識精確率的比較。這些時間序列處理技術，包括了在第二章中提到的 HEQ、TSN 與 SHE 等，都是直接或間接地更新特徵之調變頻譜，進而強化雜訊強健性。由於在文獻中提及這些技術時，都是直接展示其作用於 MVN 前置處理後之特徵的辨識效果，在這裡，我們同樣先將原始語音特徵 MFCC 先經 MVN 處理後，再分別運作這些技術。

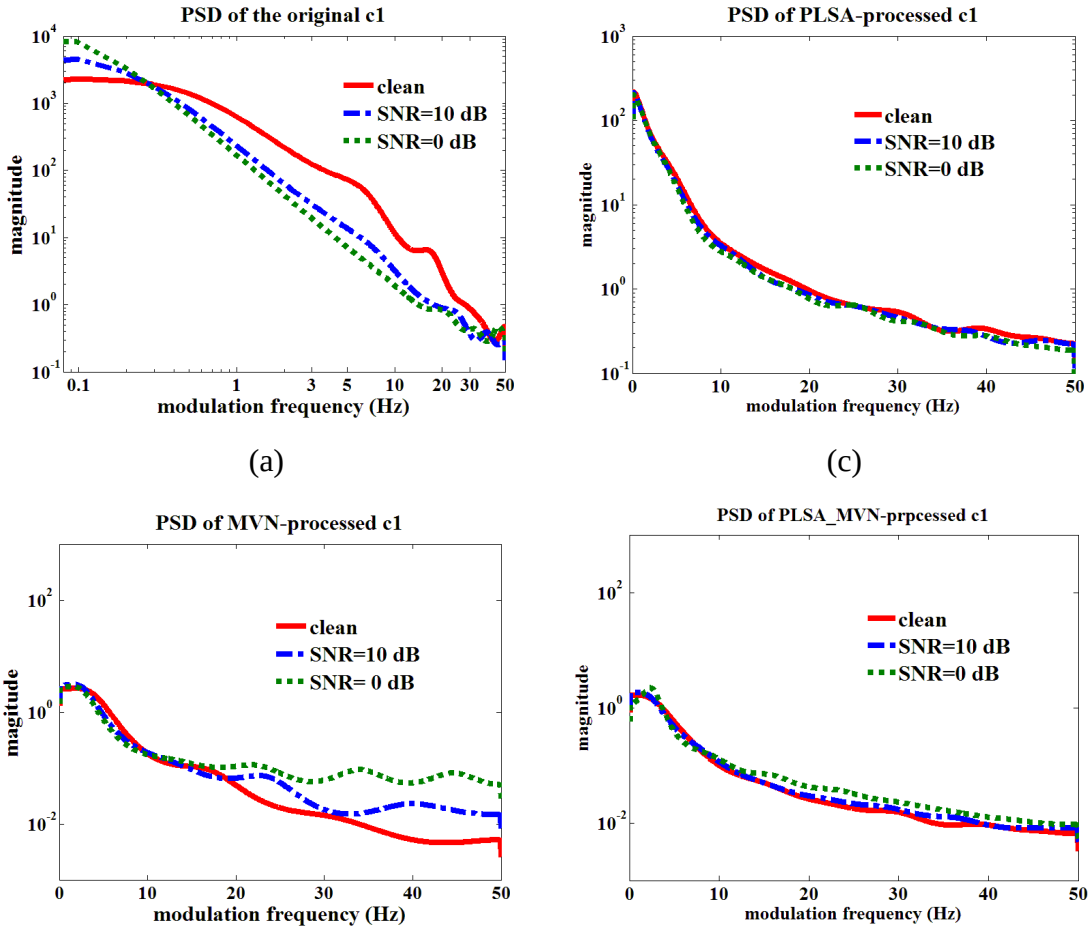
圖四中展示了上述這些技術所得之平均辨識精確率，我們提出的 PLSA 法所採用的隱藏主題數為 20。從此圖中，我們看到這裡所用的所有方法皆能提升 MVN 特徵的辨識精確率，所引用的三種方法中，又以 TSN 的效能最好，能達到 91.02%的總平均辨識率，雖然我們提出的 PLSA 法，辨識效能略低於 TSN，但也可使總平均辨識率提升至 90.57%，此初步顯示了 PLSA 法足以與現今有名的調變頻譜更新技術在效能上並駕齊驅。

4、PLSA 降低調變頻譜強度失真的效能

最後，除了辨識精確率的驗證，我們嘗試進一步藉由 PLSA 法於不同訊噪比(SNR)下所得之特徵序列調變頻譜強度，檢視其降低雜訊所產生之失真的能力。圖五(a)-(d)為單一語句其原始與經過各種處理方法後之第一維梅爾倒頻譜特徵參數 c1 於三種不同訊噪比(clean、10dB 與 0dB)之功率頻譜密度(power spectral density, PSD)。首先，觀察圖五(a)可發現，雜訊干擾所引發的不匹配效應，明顯普遍存在整個調變頻率範圍[0, 50Hz]內。圖五(b)則顯示，MVN 可有效降低低調變頻率之 PSD 失真，但是不匹配情形仍舊存在於中高調變頻率成分。圖五(c)與圖五(d)則是前述圖五(a)與(b)兩類特徵參數經過 PLSA 處理過後之 PSD 曲線；從這兩圖皆可發現，藉由 PLSA 法，可大幅降低整個頻率範圍的 PSD 失真。而相較於圖五(c)，圖五(d)中於不同訊噪比之 PSD 曲線則更為一致，顯示 PLSA 與 MVN 結合後，能更有效降低雜訊產生的失真。

五、結論與未來展望

本論文針對語音特徵時間序列之調變頻譜提出嶄新的分析與強化技術，利用機率式潛藏語意分析(PLSA)賦予調變頻譜強度其機率的意義，並透過一組潛藏的主題機率分布，以描繪語句與調變頻譜強度之關係，同時予以機率式分解與成分分析，並藉此更新調變頻譜強度以求取更具強健性的語音特徵序列。辨識實驗結果顯示，所提出的新方法能有效提升雜訊環境下語音辨識精確率，且展示了此新方法與時間序列域之正規化法能有互



圖五、c1 特徵序列於三種訊噪比情況下之 PSD 曲線，其中(a)為原始 MFCC 特徵，(b)為 MVN 處理後之特徵 (c)為 PLSA 處理後之特徵 (d)為 PLSA 結合 MVN 處理後之特徵

補的作用。未來，我們期望能嘗試將其它資料分解(data factorization)的技術運用於調變頻譜的分析上[20]，進而比較其特性與探索優缺點。同時，將 PLSA 法與統計圖等化法及其延伸[21]作結合；並且，應用 PLSA 法探索語音訊號於其它域的特性。

參考文獻

[1] S. Furui, "*Cepstral Analysis Techniques for Automatic Speaker Verification*," IEEE Trans. on Acoustic, Speech and Signal Processing, Vol. 29(2): pp. 254-272, 1981

[2] A. Vikki, and K. Laurila, "*Segmental Feature Vector Normalization for Noise Robust Speech Recognition*," Speech Communication, Vol. 25: pp. 133-147, 1998

[3] A. D. L. Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, and A. J. Rubio, "*Histogram equalization of speech representation for robust speech recognition*," IEEE Trans. on Speech and Audio Processing, Vol. 13(3): pp. 355-366, 2005

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

[5] L. C. Sun, C. W. Hsu, and L. S. Lee, "*Modulation Spectrum Equalization for robust*

Speech Recognition," in Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2007
[6] S-Y. Huang, W. H. Tu, and J-W. Hung, "*A study of sub-band modulation spectrum compensation for robust speech recognition*," in Proc. ROCLING XXI: Conf. on Computational Linguistics and Speech Processing, 2009
[7] X. Xiao, E. S. Chng, and H. Li, "*Normalization of the speech modulation spectra for robust speech recognition*," IEEE Trans. on Speech and Audio Processing, 2008
[8] J-W. Hung and W-Y. Tsai, "*Constructing modulation frequency domain based features for robust speech recognition*," IEEE Trans. Acoustic, Speech, Language Processing, 2008
[9] H. Hermansky and N. Morgan., "RASTA processing of speech," IEEE Trans. on Speech and Audio Processing, 2(4): pp. 578-589, 1994
[10] J. Koehler et al., "*Integrating RASTAPLP into Speech Recognition*," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 421-424, 1994
[11] T. Hofmann, "*Probabilistic latent semantic analysis.*" in Proc. Uncertainty in Arterial Intelligence, UAI, 1999
[12] H. Hermansky, "*Should Recognizers Have Ears?*" Invited Tutorial Paper, in Proc. ESCA-NATO Tutorial and Research Workshop on Robust speech recognition for unknown communication channels, pp. 1-10, 1997
[13] S. Greenberg, "*On the origins of speech intelligibility in the real world*," in Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels,1997
[14] B. Chen, "*Word topic models for spoken document retrieval and transcription,*" ACM Transaction on Asian Language Information Processing, Vol. 8, No. 1, pp. 2:1-2:27, 2009
[15] J. Driesen, H. Van Hamme, "*Modeling vocabulary acquisition, adaptation and generalization in infants using adaptive Bayesian PLSA*," Neurocomputing, 2011
[16] H. G. Hirsch and D. Pearce, "*The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions*," in Proc. ISCA ITRW ASR 2000
[17] C-P. Chen and J. Bilmes, "*MVA processing of speech features*," IEEE Trans. on Audio, Speech and Language Processing, Vol. 15, No. 1, pp. 257-269, 2007
[18] http://htk.eng.cam.ac.uk/
[19] http://www.nist.gov/index.html
[20] W-Y. Chu, J-W. Hung and B. Chen, *"Modulation spectrum factorization for robust speech recognition*," in Proc. APSIPA Annual Summit and Conference (APSIPA ASC), 2011
[21] B. Chen, W-H. Chen, S-H. Lin, and W-Y. Chu, *"Robust speech recognition using spatial–temporal feature distribution characteristics*," Pattern Recognition Letters, Vol. 32, No. 7, pp. 919-926, 2011