Computational Linguistics and Chinese Language Processing
Vol. 20, No. 2, December 2015, pp. 87-106
© The Association for Computational Linguistics and Chinese Language Processing

調變頻譜分解技術於強健語音辨識之研究

Investigating Modulation Spectrum Factorization Techniques for Robust Speech Recognition

張庭豪*、洪孝宗*、陳冠宇+、王新民+、陳柏琳*

Ting-Hao Chang, Hsiao-Tsung Hung, Kuan-Yu Chen,

Hsin-Min Wang and Berlin Chen

摘要

自動語音辨識(Automatic Speech Recognition, ASR)系統常因環境變異而導致效能嚴重地受影響；所以長久以來語音強健(Robustness)技術的發展是一個極為重要且熱門的研究領域。本論文旨在探究語音強健性技術，希望能透過有效的語音特徵調變頻譜處理來求取較具強健性的語音特徵。為此，我們使用非負矩陣分解(Nonnegative Matrix Factorization, NMF)以及一些改進方法來正規化調變頻譜強度成分，藉以獲得較具強健性的語音特徵。本論文有下列幾項貢獻。首先，結合稀疏性的概念，期望能夠求取到具調變頻譜局部性的資訊以及重疊較少的 NMF 基底向量表示。其次，基於局部不變性的概念，希望發音內容相似的語句之調變頻譜強度成分，在 NMF 空間有越相近的向量表示以維持語句間的關聯程度。再者，在測試階段經由正規化 NMF 之編碼向量，更進一步提升語音特徵之強健性。最後，我們也結合上述三種 NMF 的改進方法。本論文的所有實驗皆於國際通用的標竿語料——Aurora-2 連續數字資料庫進行；實驗結果顯示相較於僅使用梅爾倒頻譜特徵之基礎實驗，我們所提出的改進方法皆

*國立臺灣師範大學資訊工程學系
Department of Computer Science & Information Engineering, National Taiwan Normal University
E-mail: {60247029S, 60047064S, berlin}@ntnu.edu.tw
+中央研究院資訊科學所
Institute of Information Science, Academia Sinica.
E-mail: {kychen, whm}@iis.sinica.edu.tw
The author for correspondence is Berlin Chen.

能顯著地降低語音辨識錯誤率。此外，我們也嘗試將所提出的改進方法與一些知名的特徵強健技術做比較和結合，以驗證這些改進方法之實用性。

關鍵詞：語音辨識、雜訊、強健性、調變頻譜、非負矩陣分解

Abstract

The performance of an automatic speech recognition (ASR) system often deteriorates sharply due to the interference from varying environmental noise. As such, the development of effective and efficient robustness techniques has long been a challenging research subject in the ASR community. In this article, we attempt to obtain noise-robust speech features through modulation spectrum processing of the original speech features. To this end, we explore the use of nonnegative matrix factorization (NMF) and its extensions on the magnitude modulation spectra of speech features so as to distill the most important and noise-resistant information cues that can benefit the ASR performance. The main contributions include three aspects: 1) we leverage the notion of sparseness to obtain more localized and parts-based representations of the magnitude modulation spectra with fewer basis vectors; 2) the prior knowledge of the similarities among training utterances is taken into account as an additional constraint during the NMF derivation; and 3) the resulting encoding vectors of NMF are further normalized so as to further enhance their robustness of representation. A series of experiments conducted on the Aurora-2 benchmark task demonstrate that our methods can deliver remarkable improvements over the baseline NMF method and achieve performance on par with or better than several widely-used robustness methods.

Keywords: Speech Recognition, Language Model, Concept Information, Model Adaptation

1. 研究動機

大多數的自動語音辨識系統，在不受雜訊干擾的理想實驗室環境下，皆可獲得良好的辨識效果；但是在真實的日常環境中，往往因為環境中諸多複雜因素的影響，造成訓練環境與測試環境存在不匹配問題，使得此系統之辨識精確率大幅度降低。造成環境不匹配問題的因素有語者變異、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。本研究探討語音辨識之強健性技術，希望降低上述因素所帶來的負面影響，進而使語音辨識系統在實際應用時仍能保有一定的效能表現。

當前所發展出的各種語音強健技術大致可分為三種類型(Lin *et al.*, 2009; Chu *et al.*, 2011)：第一種類型為以聲學模型(Acoustic Model)為基礎之強健性技術(Model-Based Techniques)，此類方法大多是期望透過少量在測試環境所錄製的調適語料來對聲學模型

進行調整，使聲學模型可以近似於輸入含雜訊語音的機率分布參數，達到降低環境不匹配所造成影響的目的。第二類是以語音特徵為基礎之強健性技術(Feature-Based Techniques)。此類方法期望經過適當的正規化處理後，能使含雜訊語音與其原始乾淨語音趨於一致。最後第三類型為綜合式強健性技術，即同時在特徵處理和模型訓練兩階段做改善。

本論文將探討以語音特徵為基礎之強健性技術。其研究的議題主要圍繞在對何種空間正規化？以及在該空間應如何正規化？典型方法是將時間序列域(Temporal Domain)上的語音特徵視為是隨機變數(Random Variable)的樣本(Samples)，利用觀測到樣本去估測隨機變數之統計特性，進而對語音特徵時間序列做線性或非線性的轉換，使其在部分或整體之統計特性能經過正規化的處理。常見的方法有統計圖等化法(Histogram Equalization, HEQ)(Torre *et al.*, 2005)、倒頻譜平均值減去法(Cepstral Mean Subtraction, CMS)(Furui, 1981)以及倒頻譜平均數與變異數正規化法(Cepstral Mean and Variance Normalization, CMVN)(Vikki & Laurila, 1998)。上述方法所利用的統計資訊仍有所不足，無法觀察出明確的時序結構(Temporal Structure)改變。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，相對於時間序列域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。例如，人們發出的聲音大多集中在調變頻譜的低頻處，因為發聲器官限制了語速。加成性的噪音則可能會反應在每個頻率，那些在高頻或與人聲不同的頻帶的資訊就可以被區分出來。近年來在調變頻譜域的語音強健性研究相當熱門，學者們致力於正規化特徵參數之時空結構，藉由強化語音特徵之調變頻譜來提升語音特徵的強健性。相關的技術包括了調變頻譜統計圖等化法(Spectral Histogram Equalization, SHE)(Sun *et al.*, 2007)、分頻式調變頻譜統計正規化法(Sub-Band Modulation Spectrum Compensation)(Huang *et al.*, 2009)與其它一系列資料導向(Data-Driven)之時間序列濾波器法(Xiao *et al.*, 2008; Hermansky & Morgan, 1994)等。

本論文旨在探究使用非負矩陣分解(Nonnegative Matrix Factorization, NMF)以及一些改進方法來正規化調變頻譜強度成分，以獲得較具強健性的語音特徵。首先，結合稀疏性的概念，期望能夠求取到具調變頻譜局部性的資訊以及重疊較少的 NMF 基底向量表示。其次，基於局部不變性的概念，希望發音內容相似的語句之調變頻譜強度成分，在 NMF 空間有越相近的向量表示以維持語句間的關聯程度。再者，在測試階段經由正規化 NMF 之編碼向量，更進一步提升語音特徵之強健性。最後，我們也結合上述三種 NMF 的改進方法。此外，也嘗試將我們所提出的改進方法與一些現有的特徵強健技術做比較和結合，以驗證這些改進方法之實用性。

2. 調變頻譜正規化法

2.1 調變頻譜之簡介

對於任一特定維度語音頻譜特徵所成的時間序列 $x[n]$而言，其調變頻譜定義如下：

$$X[k] = DFT(x[n]) = \sum_{t=0}^{N-1} x[t]e^{-j\frac{2\pi tk}{N}} \quad ,0 \le k \le \frac{N}{2} \tag{1}$$

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, DFT)，$X[k]$代表語音特徵時間序列 $x[n]$的調變頻譜。由式(1)可看出調變頻譜可以被用來廣泛地分析語句中語音特徵隨時間變化的資訊。而 $X[k]$頻譜序列可視為一種對於原始語音訊號作降低取樣(Down-Sampling)後的調變訊號(由訊號取樣率轉至音框取樣率)，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 $x[n]$之取樣頻率(音框取樣率)有關。例如，在一般設定下，若音框取樣率為 100 Hz，則最高調變頻率為 50 Hz。

過去已有不少學者研究語音特徵之調變頻譜的特性，發現了調變頻譜中的低頻成分是比高頻成分還要重要的特性(Kanedera *et al.*, 1997)。同時，調變頻譜之低頻成分(約 1Hz 至 16Hz)對於語音辨識正確率也有密切的關係，潛藏有重要的語意資訊。其中，最重要的是位於 4 Hz 附近，有學者指出，4 Hz 是人耳聽覺最為敏感之調變頻率(Hermansky, 1998)；另有學者也認為，4 Hz 為人類大腦皮層感知之重要調變頻率(Greenberg, 1997)。當語音訊號受到雜訊影響時，其語音特徵時間序列會受到影響而失真，及其調變頻譜也會跟著受到牽連。很多學者提出作用在調變頻譜的正規化法，以改善調變頻譜受到雜訊干擾的影響。因此，我們可將許多發展在語音特徵時間序列的正規化法應用在調變頻譜使其正規化；而正規化的對象是對其調變頻譜強度(Magnitude)成分$|X[k]|$來進行處理，並保持其相位角不變 $\theta[k]=\angle X[k]$的部分。接著，經處理後被更新的強度成分會與原始相位成分結合，再藉由反傅立葉轉換(Inverse Discrete Fourier Transform, IDFT)來求得新的語音特徵時間序列。若調變頻譜的強度能夠被有效的正規化，便能夠有效解決雜訊產生的環境不匹配問題，使自動語音辨識系統在使用新的語音特徵的情況下能夠獲得較佳的辨識率。以下將會簡單回顧一些常見的調變頻譜正規化法。

2.2 調變頻譜平均正規化法(Spectral Mean Normalization, SMN)

假設當各種音素在一般環境中分布的比例接近一致時，每一維度語音特徵的調變頻譜之平均值應該為一個定值(Huang *et al.*, 2009)：

$$\left|\tilde{X}[k]\right| = |X[k]| - \mu_s + \mu_a \tag{2}$$

在式(2)中，$|X[k]|$為原始的調變頻譜強度成分，μ_s為單一語句的調變頻譜強度成分之平均值，μ_a為所有訓練語句的調變頻譜強度成分之平均值，而$\left|\tilde{X}[k]\right|$便是更新過後的調變頻譜強度成分。

2.3 調變頻譜平均與變異數正規化法(Spectral Mean and Variance Normalization, SMVN)

除了要正規化調變頻譜強度成分之平均值外，也可同時正規化其標準差(Huang *et al.*, 2009)。假設特徵向量參數之平均值與變異數在一般環境中分布的比例接近一致時，我們

可以同時對其平均值和標準差來進行正規化：

$$\left|\tilde{X}[k]\right| = \frac{|X[k]| - \mu_s}{\sigma_s}\sigma_a + \mu_a \tag{3}$$

在式(3)中，μ_s與σ_s為單一語句的調變頻譜強度成分之平均值與標準差；μ_a與σ_a為所有訓練語句的調變頻譜強度成分之平均值與標準差，$\left|\tilde{X}[k]\right|$便是更新過後的調變頻譜強度成分。

2.4 調變頻譜統計圖等化法(Spectral Histogram Equalization, SHE)

利用非線性的轉換(Nonlinear Transformation)，不僅將調變頻譜強度成分之平均值與標準差(或變異數)作正規化，而是整體上使得訓練語句與測試語句的調變頻譜強度成分趨於擁有同一個機率分布函數，正規化全部階層的動差(Sun *et al.*, 2007)：

$$\left|\tilde{X}[k]\right| = F_{ref}^{-1}\left(F_X(|X[k]|)\right) \tag{4}$$

在式(4)中，$F_X(\cdot)$為單一語句某一特徵維度的調變頻譜強度之累積分布函數(Cumulative Distribution Function, CDF)，F_{ref}則是利用所有訓練語句之調變頻譜強度所求得的對應之參考累積分布函數，$\left|\tilde{X}[k]\right|$便是更新過後的調變頻譜強度成分。

2.5 分頻段調變頻譜統計正規化法

此方法的概念是想要改進原始調變頻譜統計正規化法；原始調變頻譜統計正規化法是將全部調變頻帶的頻譜強度值視為是屬於同一隨機變數的樣本(Samples)，且將之一併進行正規化的動作。但是前面提到在語音辨識中，不同調變頻率的成分有不同的重要性，低頻成分是比高頻成分還要相對重要的，因為語言的重要資訊較集中於低頻成分。因此，有學者提出將調變頻帶分成許多子頻段，再分別對每一個子頻段的頻譜強度作上述所提的調變頻譜正規化的方法，而不是單純直接對整個全部調變頻帶做處理(Huang *et al.*, 2009)。因為要強調低調變頻率的重要性，所以在低頻部分的子頻段擁有較窄的頻寬，子頻段的數量也比較多，而高調變頻率便持有相反的特性。由於能更細緻地分析與處理低頻成分的資訊，過去的一些實驗數據顯示出將調變頻率分頻段來正規化的做法，能比全頻帶正規化的方式獲得較好的效能。

3. 三種新穎NMF改進方法用於調變頻譜分解

3.1 傳統非負矩陣分解法(NMF)

在很多領域中如何尋找重要的潛藏資訊成分是個重要的議題，而基於非負矩陣分解法(Nonnegative Matrix Factorization, NMF)(Lee & Seung, 1999)的技術可以被用於處理此議題。顧名思義，此方法就是將非負的原始資料所成的矩陣進行分解，表示成兩個也是非負的矩陣乘積，接著利用線性組合的特性來表示原始資料中各個樣本之目的。而其它常見的線性表示法有主成分分析(Principal Component Analysis, PCA)與獨立成分分析(Independent Component Analysis, ICA)。非負矩陣分解法與這兩種線性表示法之差異就是

能夠提供非負的基底向量(Nonnegative Basis Vectors)，且也能夠擁有保證由基底向量組合而成之資料也為非負的特性。非負矩陣分解法的另一個重要特性是想要學習以部分為基礎(Parts-Based)之線性表示法來表示原始的資料，且此線性表示法是一個加法的組合模式。這種以部分為基礎的概念方法擁有直觀的性質，而且對於一個特定任務來說，在與其它分解方法相比下可以得到比較高的解釋性。過去有學者應用非負矩陣分解法在影像處理的領域，例如人臉影樣可以用為五官等局部影像做為非負基底向量經由線性組合(線性編碼)而產生。若是使用上述所提到的，例如 PCA，在分解舉證產生基底向量的過程中可能會產生負值，這些負值在影像處理當中會難以解釋。而在語音領域方面，語音的特徵值有正有負，所以較難以直接地使用非負矩陣分解法；直到近期有學者將非負矩陣分解法用在分析調變頻譜強度以擷取重要語音特徵(Chu *et al.*, 2011)，而可以得到了不錯的強健性效果。NMF 的數學式表示如下：

$$\mathrm{V} \approx \mathrm{WH} = \sum_{k} \mathrm{W}_{ik}\mathrm{H}_{kj} \tag{5}$$

其中$\mathrm{V} \in R^{I\times J}$為一個非負矩陣，而兩個被分解出來的非負矩陣分別為$\mathrm{W} \in R^{I\times K}$和 $\mathrm{H} \in R^{K\times J}$，如圖 1 所示。

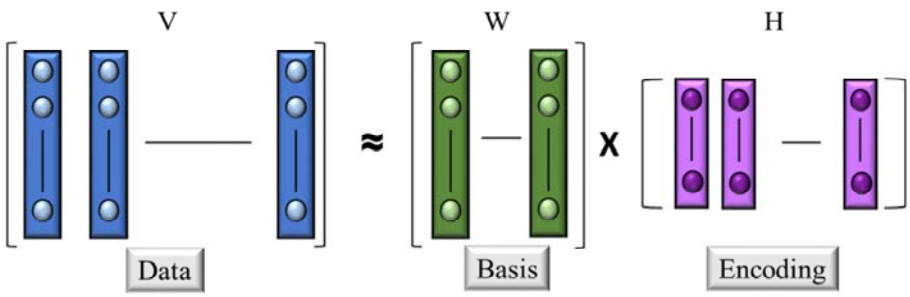

圖 1. 非負矩陣分解法(NMF)示意圖

其中矩陣 W 所包的 K 行即為基底向量，矩陣 H 中的每一行則通常被稱為編碼向量(Encoding)，有著權重的概念，與基底向量進行線性組合去近似資料矩陣 V。I 是每筆資料向量的維度大小；J 為所有資料向量的個數；K 為基底向量的數量。參數 K 是可以自行決定的，通常會選擇小於 I 與 J，但還是會有選擇的限制：

$$(I + J) \times K < I \times J \tag{6}$$

式(6)是學者過去所提出更確切的基底向量個數選擇限制關係式。在非負矩陣分解法的方法中，有著資料壓縮的概念，若是 K 的數目選擇得越少，代表壓縮的比率越高。因為我們對資料進行了壓縮的動作，所以壓縮後的資料跟原始的資料來比較必定會有一些資料是在壓縮過程中被遺失了。我們希望遺失的部分資料越少越好，所以可以定義減損函數(Loss Function)來測量資料前後的相似度。測量由兩個因子矩陣 W 與 H 所重建的訊號 Λ 與原始訊號 V 之間的距離，對分解結果與原始資料的近似程度作量化(Quantification)。

非負矩陣分解法常見的減損函數為歐氏距離(Euclidian Distance 或 Frobenius Norm)：

$$D_F(V||WH) = ||V - WH||_F^2 = \sum_{i,j}(V_{ij} - (WH)_{ij})^2 \tag{7}$$

$D_F(V||WH)$是藉由歐氏距離所提出的減損函數。當重建訊號Λ與原始信號 V 相等時，則 $D_F(V||WH) = 0$。另一個減損函數則是基於 KL 散度(Kullback-Leibler Divergence)：

$$D_{KL}(V||WH) = \sum_{i,j}\left(V_{ij}\ln\frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}\right) \tag{8}$$

當原始信號 V 與重建訊號Λ相等時，$D_{KL}(V||\Lambda) = 0$。因為 KL 散度不具對稱性(Symmetric)，因此減損函數值不能稱為兩個訊號之間的距離值(Distance)，而是兩訊號之間的差異值(Divergence)，而 KL 散度也稱為相對熵(Relative Entropy)。

由於要將資料矩陣 V 分解成 W 與 H，而使誤差最小化。所以使用迭代更新規則將 W 與 H 更新去求得局部最小值(Local Minimum)。最起初提出的方法是使用梯度下降演算法(Gradient Descent Algorithm)與加法迭代(Iteration)規則。後來又有學者提出乘法迭代規則；乘性迭代規則能夠直接地賦予非負矩陣分解法之非負限制的特性。以下是乘法迭代更新規則(Lee & Seung, 2000)：

Euclidian Distance 的乘法更新規則：

$$H_{kj} \leftarrow H_{kj}\frac{(W^TV)_{kj}}{(W^TWH)_{kj}}$$

$$W_{ik} \leftarrow W_{ik}\frac{(VH^T)_{ik}}{(WHH^T)_{ik}} \tag{9}$$

Kullback-Leibler Divergence 的乘法更新規則：

$$H_{kj} \leftarrow H_{kj}\frac{\sum_i W_{ik}V_{ij}/(WH)_{ij}}{\sum_i W_{ik}}$$

$$W_{ik}\frac{\sum_j H_{kj}V_{ij}/(WH)_{ij}}{\sum_j H_{kj}} \tag{10}$$

3.2 非平滑非負矩陣分解法(NSNMF)

非平滑非負矩陣分解法(Pascual-Montano *et al.*, 2006)直接修改傳統非負矩陣分解法的模型，利用模型的乘法性質，達到矩陣全面的稀疏，以能擷取更局部的資訊(如圖 2 所意示)。非負矩陣分解法將資料矩陣分成兩個矩陣相乘，也就是基底矩陣乘以編碼矩陣。若在一個矩陣中，其元素是非稀疏或平滑的，為了要補償最後兩個矩陣相乘之後能盡可能地近似原始資料矩陣，這將會迫使另一個矩陣面臨稀疏或非平滑的情況。非平滑非負矩陣分解法可以定義如下：

$$V = WSH \tag{11}$$

在式(11)中，矩陣$V \in R^{I \times J}$為資料矩陣；矩陣$W \in R^{I \times K}$為基底矩陣；矩陣$H \in R^{K \times J}$為編碼矩陣；而矩陣$S \in R^{K \times K}$稱為平滑矩陣，其定義如下：

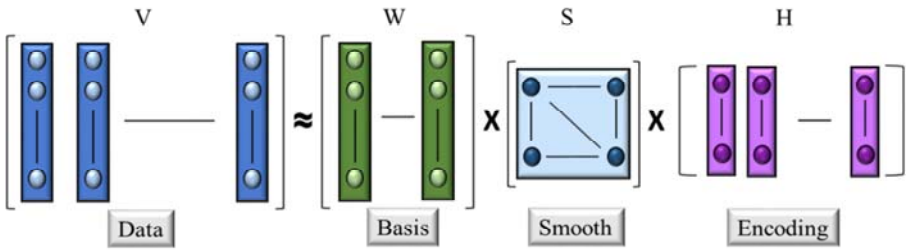

圖 2. 非平滑非負矩陣分解法(NSNMF)示意圖

$$S = (1 - \theta)I + \frac{\theta}{K}11^T \tag{12}$$

式(12)中 1 是一個元素都是 1 的向量，I 是單位矩陣，以及 θ 是一個用來控制整體稀疏程度的參數，此參數 θ 滿足$0 \le \theta \le 1$的範圍中。對平滑矩陣 S 可以解釋為：假設 X 為一個正的非零值向量，而 Y=SX 為轉換後的向量。如果 θ =0，Y=X，意謂著向量 X 中沒有平滑發生；如果 θ =1，向量 Y 中所有的元素會變成一致的數值，此數值會等於向量 X 所有元素的平均，這就是最平滑的向量。由上述可知參數 θ 用來控制平滑矩陣 S 的平滑程度。由於模型的乘法性質，平滑矩陣 S 中若有強烈的平滑情況，將會迫使在基底向量與編碼向量中造成強烈的稀疏，因此也可以說參數 θ 是用來控制整個非負矩陣分解法模型的稀疏程度。特別的是，當參數 θ =0 時，平滑矩陣 S 會等同於一個單位矩陣 I，此時模型會回歸到傳統的非負矩陣分解法的模型。在此，我們將更進一步地說明整個非平滑非負矩陣分解法的流程與乘法更新規則。首先，式(11)中非平滑非負矩陣分解法的模型可以等價地寫成：

$$V = (WS)H = W(SH) \tag{13}$$

用括號來表示平滑矩陣S是先與哪個矩陣做相乘。若是平滑矩陣S先與基底矩陣W做相乘，代表說基底矩陣W會變得平滑，這將會迫使編碼矩陣 H 變得稀疏；同樣地，若是平滑矩陣S先與編碼矩陣H做相乘，代表說編碼矩陣H會變得平滑，這將會迫使基底矩陣W變得稀疏。在非負矩陣分解與更新過程中，上述兩種情況將都會發生的，所以基底矩陣W與編碼矩陣H都會被強制變為具稀疏性。相較於傳統非負矩陣分解法，非平滑非負矩陣分解法的乘法迭代更新規則為：在更新編碼矩陣 H 時，將W換成(WS)；更新基底矩陣W時，將H換成(SH)。

Euclidian Distance 的乘法更新規則：

$$\mathrm{H}_{kj} \leftarrow \mathrm{H}_{kj} \frac{((\mathrm{WS})^{\mathrm{T}}\mathrm{V})_{kj}}{((\mathrm{WS})^{\mathrm{T}}(\mathrm{WS})\mathrm{H})_{kj}}$$

$$\mathrm{W}_{ik} \leftarrow \mathrm{W}_{ik} \frac{(\mathrm{V}(\mathrm{SH})^{\mathrm{T}})_{ik}}{(\mathrm{W}(\mathrm{SH})(\mathrm{SH})^{\mathrm{T}})_{ik}} \tag{14}$$

Kullback-Leibler Divergence 的乘法更新規則：

$$\mathrm{H}_{kj} \leftarrow \mathrm{H}_{kj} \frac{\sum_i (\mathrm{WS})_{ik}\mathrm{V}_{ij}/((\mathrm{WS})\mathrm{H})_{ij}}{\sum_i (WS)_{ik}}$$

$$\mathrm{W}_{ik} \leftarrow \mathrm{W}_{ik} \frac{\sum_j (\mathrm{SH})_{kj}\mathrm{V}_{ij}/(\mathrm{W}(\mathrm{SH}))_{ij}}{\sum_j (\mathrm{SH})_{kj}} \tag{15}$$

而其它部分的演算法流程同傳統非負矩陣分解法。

3.3 基於圖正則化非負矩陣分解法(GNMF)

基於圖正則化非負矩陣分解法(Graph Regularized Non-negative Matrix Factorization, GNMF)(Cai *et al*., 2011)的主要目的在於保留資料的局部不變性(Locally Invariant)(Hadsell *et al*., 2006)，意指原本相鄰的資料向量經過降維或投影後仍然維持相鄰近。資料向量間的遠近關係，或幾何結構資訊可以用一權重矩陣 E 表示，其維度是等於資料向量數量所形成的方陣。最後將權重矩陣 E 納入減損函式中，做為編碼矩陣的正則項(Regularization Term)。

令$\boldsymbol{h_j} = [h_{j1}, \dots, h_{jk}]^T$為編碼矩陣H的第 j 行，$\boldsymbol{h_j}$可被視為是第$\boldsymbol{v_j}$個資料向量相對於新的基底矩陣W之新表示。在此我們討論較常見的歐式距離：

$$d(\boldsymbol{h_j}, \boldsymbol{h_l}) = \left\|\boldsymbol{h_j} - \boldsymbol{h_l}\right\|^2 \tag{16}$$

此距離用來測量相對於新的基底矩陣W，而兩個資料向量$\boldsymbol{h_j}$與$\boldsymbol{h_l}$在低維度空間中表示之間的差異(Dissimilarity)，距離函式值越大代表此兩個資料向量$\boldsymbol{h_j}$與$\boldsymbol{h_l}$彼此差異越大。

$$\begin{aligned} R_1 &= \frac{1}{2}\sum_{j,l=1}^{v} \left\|\boldsymbol{h_j} - \boldsymbol{h_l}\right\|^2 \mathrm{E}_{jl} \\ &= \sum_{j=1}^{v} \boldsymbol{h_j}^T \boldsymbol{h_j} \mathrm{D}_{jj} - \sum_{j,l=1}^{v} \boldsymbol{h_j}^T \boldsymbol{h_l} \mathrm{E}_{jl} \\ &= Tr(\mathrm{H}^{\mathrm{T}}\mathrm{DH}) - Tr(\mathrm{H}^{\mathrm{T}}\mathrm{EH}) \\ &= Tr(\mathrm{H}^{\mathrm{T}}\mathrm{LH}) \end{aligned} \tag{17}$$

其中R_1是編碼矩陣的正則項，$Tr(\cdot)$為矩陣的跡數(Trace)，$\mathrm{D}_{jj} = \sum_l \mathrm{E}_{jl}$ ，$\mathrm{L} = \mathrm{D} - \mathrm{E}$，L稱作圖拉普拉斯算子(Graph Laplacian)。在此希望使R_1最小化，達到保留資料局部不變性的

目的。將上述所求出的R_1當作懲罰項，加入到傳統 NMF 之歐式距離減損函式中可以得到新的減損函數：

$$O_{Euclidean} = \|\mathrm{V} - \mathrm{WH}\|^2 + \lambda Tr(\mathrm{HLH}^{\mathrm{T}}) \tag{18}$$

同樣地，可利用梯度下降演算法去求出基於圖正則化非負矩陣分解法的乘法更新規則：

$$\mathrm{H}_{kj} \leftarrow \mathrm{H}_{kj} \frac{(\mathrm{W}^{\mathrm{T}}\mathrm{V} + \lambda \mathrm{HE})_{kj}}{(\mathrm{W}^{\mathrm{T}}\mathrm{WH} + \lambda \mathrm{HD})_{kj}}$$

$$\mathrm{W}_{ik} \leftarrow \mathrm{W}_{ik} \frac{(\mathrm{VH}^{\mathrm{T}})_{ik}}{(\mathrm{WHH}^{\mathrm{T}})_{ik}} \tag{19}$$

其中$\lambda \geq 0$，為正則化參數，去控制新的表示之平滑性。

有別於傳統 NMF 方法僅在歐氏空間中求解，GNMF 方法可以視不同應用問題而設計合適的權重矩陣 E。在語音辨識的任務中，聲學模型通常被建立在音素層次，而且主宰語音辨識的表現。因此，本論文也提出利用音素錯誤率建立權重矩陣 E，詳細的描述與實驗稍後將被呈現在第 4.3 節。

3.4 非負編碼矩陣統計圖等化法(HNMF)

傳統的 NMF 方法將訓練資料分解成非負基底矩陣 $\mathrm{W}_{\mathrm{clean}}$ 和編碼矩陣 $\mathrm{H}_{\mathrm{clean}}$ 兩部分，在測試階段時只保留基底矩陣，而丟棄了編碼矩陣的資訊。在應用中，受噪音干擾的語音可能會得到與乾淨語料不相似的編碼向量，此時我們不能確定這樣的資料表示是否已經排除大部分雜訊？再者，即便是乾淨語料中也存在著許多變異性。為了克服上述問題，本論文提出利用統計圖等化法將編碼矩陣做正規化處理。在訓練階段時，我們利用統計圖等化法(HEQ)將乾淨訓練語料的編碼矩陣 $\mathrm{H}_{\mathrm{clean}}$ 的資訊儲存建表，統計編碼矩陣 $\mathrm{H}_{\mathrm{clean}}$ 的參考分布，如圖 3。而在測試階段時求出編碼向量 $\boldsymbol{h}$，再將 $\boldsymbol{h}$ 每一個元素執行統計圖等化法之查表的動作，試圖將含有雜訊的 $\boldsymbol{h}$ 還原回其到對應的乾淨的編碼向量。以能夠去對應由乾淨訓練語料所估測出來的參考分布，使語句的編碼向量在訓練環境與測試環境之機率分布一致，如圖 4 所示。我們認為乾淨的基底向量矩陣乘上正規化後的編碼矩陣應較能夠還原回乾淨的語音特徵。

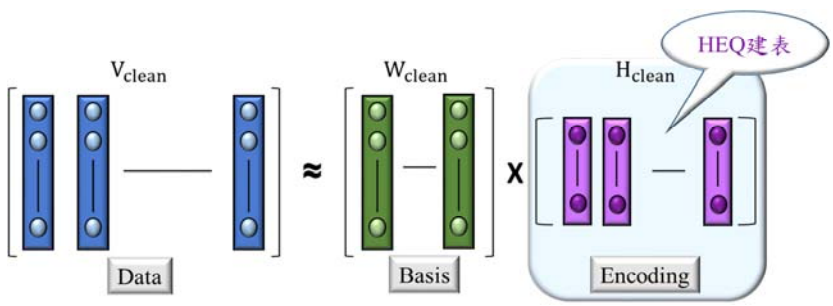

圖 3. 非負編碼矩陣統計圖等化法(HNMF)訓練階段示意圖

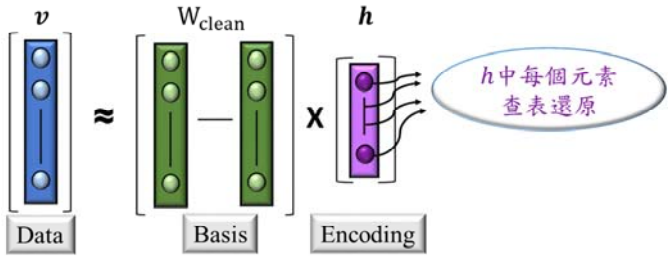

圖 4. 非負編碼矩陣統計圖等化法(HNMF)還原示意圖

4. 實驗結果與分析

4.1 實驗語料庫

本論文實驗所採用的語料庫是 Aurora-2，它是由歐洲電信標準協會(European Telecommunications Standards Institute, ETSI)所發行的語料庫(Hirsch & Pearce, 2000)，以美國成年人的聲音作為錄音來源，內容是連續的英文數字由 0(Zero)到 9(Nine)跟 Oh 等發音字詞。語料庫內有乾淨及含有雜訊的語音，雜訊中有八種不同的加成性雜訊與兩種不同的通道效應，而通道效應是使用國際電信聯合會(ITU)標準中的 G.712 和 MIRS。根據不同的雜訊干擾，分成三個測試集：Set A、Set B 及 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。而其中的訊噪比(SNR)則有七種，為 Clean、20dB、15dB、10dB、5dB、0dB 和-5dB，並且提供二種訓練模式：乾淨情境訓練模式(Clean-Condition Training)與複合情境訓練模式(Multi-Condition Training)。本研究的基礎實驗皆使用乾淨情境訓練模式，故在聲學模型訓練時並沒有使用到任何加成性雜訊的資訊或內涵。

4.2 實驗設定

在本論文中的基礎實驗是採用梅爾倒頻譜係數(Mel-frequency Cepstral Coefficients, MFCC)做為語音特徵參數，取樣頻率(Sampling Rate)為 8,000Hz，預強調(Pre-emphasis)參數設為 0.97；使用的窗函數為漢明窗(Hamming Window)，音框長度(Frame Length)是 25 毫秒，音框間距(Frame Shift)為 10 毫秒。每一個音框的語音特徵是使用 13 維梅爾倒頻譜係數(第 1 維至第 12 維還有第 0 維)，加上其一階差量和二階差量，共 39 維之特徵參數。本論在對語音特徵進行強健性(正規化)處理時，只針對 13 維的靜態特徵參數進行處理，待處理完成後才額外將語音特徵的一階差量和二階差量加入形成最後每一個音框的語音特徵。

4.3 辨識效能評估方式

辨識效能的評估方式是依照美國國家標準與科技局(National Institute of Standards and Technology, NIST)所訂立的評估標準，進行每一句測試語句之正確轉寫詞串與語音辨識詞串的比較。評估方式是以詞正確率(Word Accuracy Rate)為主，計算正確轉寫詞串與語音辨識詞串彼此間的詞取代個數(Substitutions)、詞插入個數(Insertions)和詞刪除個數(Deletions)：

$$\text{詞正確率}(\%) = \frac{\text{詞正確辨識個數} - \text{詞插入個數}}{\text{輸入詞總數}} \times 100\% \tag{20}$$

最後在評估整體語音辨識效能時，我們參照國際學者之設定，對測試語句在每一種噪音的訊噪比的詞正確率結果做加總與取平均的動作(去掉極端的訊噪比 Clean 跟-5，只計算範圍 20dB 到 0dB 中的平均詞正確率)；本論文以下的全部實驗皆是利用平均詞正確率來評估語音辨識的效能。

4.4 非平滑非負矩陣分解法(NSNMF)之實驗結果

由 3.2 節的敘述可知 NSNMF 因為乘法的性質，若是平滑程度高的矩陣S與W或H矩陣其中一個相乘，為了要補償能儘可能的近似重建原始資料，會迫使另一個矩陣達到稀疏的效果。如此經遞迴地使用乘法更新規則，最終可達到稀疏化矩陣W與H的效果。實驗數據如表 1 所示；由實驗結果可見，隨著θ的增加語音辨識之詞正確率的確能夠逐漸地被提高。雖然在$\theta = 0.3$以前較無明顯效果，可能因為迫使矩陣稀疏的程度並不高；而$\theta = 1$時，表示迫使矩陣稀疏程度最高，而數據顯示的確能夠表現得最好。

表 1. 非平滑之非負矩陣分解法(NSNMF)在使用不同θ值下之詞正確率(%)

	Set A	Set B	Set C	Average
NMF	67.09	70.98	68.22	68.87
$\theta = 0.1$	67.54	71.62	66.01	68.87
$\theta = 0.2$	66.91	71.35	64.72	68.25
$\theta = 0.3$	66.89	71.63	67.85	68.98
$\theta = 0.4$	70.19	73.64	68.72	71.28
$\theta = 0.5$	69.46	73.60	66.22	70.47
$\theta = 0.6$	70.82	74.18	71.20	72.24
$\theta = 0.7$	72.07	75.29	69.73	72.89
$\theta = 0.8$	72.99	76.25	72.75	74.25
$\theta = 0.9$	74.12	76.98	74.67	75.37
$\theta = 1$	77.05	79.75	77.47	78.21

4.5 基於圖正則化非負矩陣分解法(GNMF)之實驗結果

在探討 GNMF 的效能時，我們首先在求取權重矩陣 E 時使用 0-1 權重的方式；為此，我們先算出所有訓練語句(8,440 句)彼此間的關聯度。關聯度的估測是透過計算兩兩訓練語句彼此間的音素錯誤率(Phone Error Rate, PER)而得；我們會先求得每一句訓練語句經人工轉寫(Transcription)之音素序列(Phone Sequence)。本論文音素錯誤率的算法是使用編輯距離(Edit Distance)的算法，計算每一句訓練語句(當成目標語句)的音素序列與其它訓練語句的音素序列彼此間的音素取代個數、音素插入個數、音素刪除個數，並依下式計算音素錯誤率：

$$\text{音素錯誤率}(\%) = \frac{\text{音素取代數} + \text{音素插入數} - \text{音素刪除數}}{\text{目標訓練語句音素總數}} \times 100\% \tag{21}$$

所以最後求得權重矩陣 E 是個維度大小為 8,400*8,400 的矩陣，其中每個元素都紀錄著每一句訓練語句與其它語句彼此間的音素錯誤率，對角線上的為一個位置為某一句訓練語句自己本身所以差異是 0。值得一提的是，當使用編輯距離算差異度時，E_{jl}與E_{lj}的值可能不會一樣，是因為不同的目標語句(任兩句訓練語句，彼此的音素序列長度可能會是不同的)的假定，所以權重矩陣 E 是不對稱的。但我們認為兩個訓練語句間彼此的關聯度應該是對稱的(一樣的)；因此我們採折衷方式，將E_{jl}與E_{lj}的值都改為兩者的相加取平均，使權重矩陣 E 變成一個對稱矩陣。再者，我們設了一個門檻值(Threshold) α：

$$\begin{cases} E_{jl} \le \alpha, & E_{jl} = 1 \\ E_{jl} > \alpha, & E_{jl} = 0 \end{cases} \tag{22}$$

當E_{jl}或E_{lj}大於門檻值時，代表說這兩句訓練語句彼此間的音素錯誤率較大(訓練語句差異大)，因此關聯度設為 0，希望兩個訓練語句彼此間沒有關聯；而當E_{jl}或E_{lj}小於等於門檻值時代表這兩句訓練語句彼此間的音素錯誤率較小，應有較大的關聯性，因此設為 1。因為設定了一個門檻值α，若門檻值α設定的較嚴格(值較小時)，權重矩陣 E 就會顯得零值越多而越稀疏化。另外，在本論文中對於式(18)中的λ值設定為 100。GNMF 的實驗數據如表 2 所示。當α值越高時，代表門檻越寬鬆，權重矩陣E中非零值的元素也會越多；

表 2. GNMF 使用不同門檻值的之詞正確率(%)

	Set A	Set B	Set C	Average
NMF	67.09	70.98	68.22	68.87
$\alpha = 0.3$	68.00	72.09	67.92	69.62
$\alpha = 0.5$	67.86	72.22	68.02	69.64
$\alpha = 0.7$	67.07	71.18	66.77	68.65
$\alpha = 0.8$	67.97	72.48	67.79	69.74
$\alpha = 0.9$	68.49	72.64	68.00	70.05

數據顯示語音辨識的詞會隨α值變大而逐漸提高。可能是因為若α的值越低，門檻越嚴格，權重矩陣E中非零值的元素就越少，導致任兩語句間的關聯度在求取矩陣W 與 H時較不會被強調。

表 3. GNMF-a 使用權重矩陣全域給值之詞正確率(%)

	Set A	Set B	Set C	Average
NMF	67.09	70.98	68.22	68.87
$\alpha = 0.9$	68.49	72.64	68.00	70.05
GNMF-a	70.63	74.27	70.78	72.12

基於此觀察，本論文嘗試改成讓權重矩陣E的每一個元素都能擁有適當的權重值，而不使用0-1權重(我們認為只設定一個門檻值就將權重值一分為二的作法可能會較粗糙一些)；我們利用式(23)將E中的每個元素之音素錯誤率做轉換權重的動作，可將權重值限制在 0 到 1 之間：

$$\mathrm{E}_{jl} = \frac{1}{1 + \mathrm{PER}_{j.l}} \tag{23}$$

如此做法，可以將各個訓練語句彼此間的關聯程度做比較精細的描述，而不是只有 0 或 1 的權重值。例如：音素錯誤率 0%的轉換後會變成 1；音素錯誤率 40%的轉換後會變成 0.714；音素錯誤率 100%的轉換後會變成 0.5；音素錯誤率 160%的轉換後會變成 0.385。讓越低的音素錯誤率能夠有越高的權重值。特別的是，權重矩陣E之對每一個角線位置的值，也就是代表某一語句本身的關聯程度；因其音素錯誤率為 0%，所以關聯程度會為 1。相關的數據如表 3 所示；當改成使用全域都有值之權重矩陣E (對應方法簡稱為 GNMF-a) 會比使用預設的一個門檻值之權重矩陣E的效果會來的好一些，使詞正確率提高了 2.07%；同時也比傳統非負矩陣分解法(NMF)提高了 3.25%的詞正確率。

值得注意的是，在上述實驗中使用基於音素錯誤率求取權重矩陣E的方式，在此方式中語音特徵的所有維度的調變頻譜強度是使用相同的權重矩陣E。另一方面，我們也嘗試基於語句每一維度的調變頻譜強度，個別地利用歐式距離來計算的語句間的的關聯程度，並且也使用類似式(23)的轉換式，求出不同維度的權重矩陣E，此方式簡稱 GNMF-eu。實驗結果如表 4 所示，基於歐式距離使得不同維度對應著不同的權重矩陣去進行 NMF 中矩陣W 與 H求取，最後反應在語音辨識效能上似乎沒有較使用音素錯誤率的方式來的好。

表 4. GNMF-eu 之詞正確率(%)

	Set A	Set B	Set C	Average
NMF	67.09	70.98	68.22	68.87
GNMF-a	70.63	74.27	70.78	72.12
GNMF-eu	69.34	72.26	69.44	70.53

4.6 非負編碼矩陣統計圖等化法(HNMF)之實驗結果

如第三節所提及，我們進一步保留在訓練階段獲得之乾淨訓練語料編碼矩陣 H 的累積分布函數(CDF)資訊，將其儲存建表以供測試階段使用統計圖等化法來正規化每一測試語句的編碼向量。其數據如表 5 所示，此方法(HNMF)在基底個數等於 5 時，可以達到優於非平滑非負矩陣分解法(NSNMF)的效能。但可發現若基底個數若持續增加時，似乎就無法與 NSNMF 競爭，且效能提昇的程度不明顯。

表 5. HNMF 之不同基底個數之詞正確率(%)

	Set A	Set B	Set C	Average
K=5	77.65	80.16	77.25	78.57
K=10	69.55	74.32	67.77	71.10
K=15	67.73	72.60	65.26	69.18
K=20	66.71	71.74	63.56	68.09
K=30	67.43	72.66	64.05	68.84

4.7 三種非負矩陣分解法改進方法之結合

接著我們結合非平滑非負矩陣分解法(NSNMF)以及基於圖正則化非負矩陣分解法(GNMF)，稱為 NSGNMF(以下所結合之 GNMF 皆使用 GNMF-a)，實驗數據如表 6 所示。雖然非平滑非負矩陣分解法原本就有 78.21%之不錯的詞正確率，不過加上基於圖正則化非負矩陣分解法利用訓練語句間的相關聯度之概念，能夠有 1.24%的正確率提升。最後我們再對編碼矩陣做 HEQ 正規化處理來提升語音辨識效能(表示成 NSHGNMF)；不過再結合 HNMF 之後效果並沒有預料中顯著，只有些許的詞正確率提昇。在表 6 中也列出兩種常見的調變頻譜正規化法(SHE 與 PCA)來作為比較比較(Kao *et al.*, 2014)。SHE 是利用

表 6. NMF 改良方法結合與之詞正確率(%)比較

	Set A	Set B	Set C	Average
NMF	67.09	70.98	68.22	68.87
GNMF	70.63	74.27	70.78	72.12
NSNMF	77.05	79.75	77.47	78.21
HNMF	77.65	80.16	77.25	78.57
NSGNMF	78.22	80.92	78.95	79.45
NSGHNMF	78.28	80.96	78.98	79.49
SHE	74.82	77.44	76.47	76.20
PCA	70.90	73.34	71.39	71.97

HEQ 將調變頻譜強度成分的平均值與標準差正規化，並同時正規化其它階層的動差使訓練語句與測試語句的調變頻譜強度的機率分布趨於一致。PCA 則是對所有訓練語句的調變頻譜強度成分求取共變異數，接著利用前 *r* 個特徵值(Eigenvalues)去找其對應的 *r* 個特徵向量(Eigenvectors)以當作調變頻譜強度成分的 PCA 子空間之基底，使測試語句的調變頻譜強度成分能夠投影到 PCA 子空間以達到正規化的目的。

表 7. 結合 CMVN 與 NMF 之詞正確率(%)

	Set A	Set B	Set C	Average
CMVN	75.93	76.76	76.82	76.44
CMVN+NSNMF	83.56	85.51	83.27	84.28
CMVN+GNMF	83.58	84.78	82.36	83.81
CMVN+HNMF	82.88	84.84	82.37	83.56
CMVN+NSGNMF	83.94	85.76	83.61	84.61
CMVN+NSHGNMF	83.98	85.85	83.71	84.67

表 8. 結合 HEQ 與 NMF 之詞正確率(%)

	Set A	Set B	Set C	Average
HEQ	80.03	82.05	80.10	80.85
HEQ+NSNMF	83.84	85.88	83.70	84.63
HEQ+GNMF	83.71	84.76	82.53	83.89
HEQ+HNMF	82.89	85.52	83.59	84.08
HEQ+NSGNMF	84.02	85.89	83.79	84.72
HEQ+NSHGNMF	84.05	85.93	83.82	84.76

表 9. 結合 AFE 與 NMF 之詞正確率(%)

	Set A	Set B	Set C	Average
AFE	87.68	87.10	86.29	87.17
AFE+NSNMF	87.74	87.65	86.32	87.42
AFE+GNMF	87.45	87.72	86.23	87.31
AFE+HNMF	87.81	87.22	86.36	87.28
AFE+NSGNMF	87.85	87.66	86.54	87.51
AFE+NSHGNMF	87.82	87.70	86.55	87.52

4.8 結合不同時間序列正規化法之結果

最後我們探討額外結合不同時間序列正規化法(CMVN 與 HEQ)與本論文所提出三種調變頻譜非負矩陣分解法的實驗結果，如表 7 與 8 所示。本論文所提出三種調變頻譜非負矩陣分解法都能與先經過不同時間序列正規化法處理過後的語音特徵相結合使用而得到效能提昇。值得注意的是，CMVN 與 HEQ 皆是在語句的音框層面(Frame Level)對每個音框分別作正規化，而 NMF 的方法是在整體語句層次(Utterance Level)正規化，因分別處理不同的面向，所以在結合後有加成性的效果。效果提升最顯著的是 CMVN 與 NSHGNMF 的結合；其次是與 HEQ 結合的 NSHGNMF，皆能有不錯的進步。我們也將所提出方法與與進階前端標準(Advanced Front-End Standard, AFE)處理過語音特徵(Macho *et al*., 2002)做結合，其結果如表 9 所示。AFE 是近年來歐洲電信標準協會(ETSI)所推出的特徵向量擷取方法，是一個著名且成效非常好的常見基礎系統設置，在多種任務上被證實能顯著地提升語音辨識系統在雜訊環境中的效能。當 AFE 與 NSHGNMF 做結合時只能有些微提升；我們猜測可能是因為 AFE 本身已具備有很完善的語音特徵正規化處理程序，若再加 NSHGNMF 時，語音特徵可能會被過度地正規化而導致語音辨識效能無法被顯著提升。

5. 結論

本論文探討了非負矩陣分解法的三種改進方法並將之運用在語音特徵的調變頻譜正規化上；希望藉此能夠擷取出更強健性的調變頻譜基底向量，而達到增進語音強健性的目的。第一種是非平滑非負矩陣分解法(NSNMF)，利用添加了一個平滑矩陣 S，變更傳統非負矩陣分解法的模型；利用模型乘法的性質，使一個矩陣平滑，進而迫使另一個矩陣達到稀疏的效果。第二種是基於圖正則化非負矩陣分解法(GNMF)，在減損函式中增加了一個額外的正則項。利用幾何結構與局部不變性的特性，求得訓練語句間的關聯程度並創造一個權重矩陣以供使用，使經正規化的語音特徵能夠增加鑑別力。第三種是非負編碼矩陣統計圖等化法(HNMF)，希望能夠利用在訓練階段時可獲得的編碼矩陣，利用統計圖等化法將其累積分布函數資訊建表儲存，希望在測試階段時能藉此將含雜訊語句的編碼向量進一步正規化。

當將此三種非負矩陣分解法之改進方式運用在 Aurora-2 上時，皆能使語音辨識效能有所進步。整體上來說，NSNMF 使用的矩陣稀疏性而有較顯著且一致的效能提升；GNMF 雖沒有帶來大幅度的效能提升，但是其所利用語句之間的關聯程度資訊也能對語音特徵正規化有所幫助，像是與 NSNMF 結合，也能稍微提昇精確率；另外，HNMF 在少許基底個數時能提供不錯的語音辨識效能提升。

致謝

本論文之研究承蒙教育部－國立臺灣師範大學邁向頂尖大學計畫(102J1A0800)與行政院科技部研究計畫(MOST 104-2221-E-003-018-MY3 和 MOST 103-2221-E-003-016-MY2)之經費支持，謹此致謝。

參考文獻

Cai, D., He, X., Han, J., & Huang, T. S. (2011). Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(8), 1548-1560.

Chu, W.-Y., Hung, J.-W., & Chen, B. (2011). Modulation spectrum factorization for robust speech recognition. In *Proceedings of the APSIPA Annual Summit and Conference*, 18-21.

Furui, S. (1981). Cepstral analysis techniques for automatic speaker verification. *IEEE Transactions on Acoustic, Speech and Signal Processing*, *29*(2), 254-272.

Greenberg, S. (1997). On the origins of speech intelligibility in the real world. In *Proceedings of the ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels.*

Hadsell, R., Chopra, S., & LeCun Y. (2006). Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1735-1742.

Hermansky, H., & Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, *2*(4), 578-589.

Hermansky, H. (1998). Should Recognizers Have Ears? *Speech Communication*, *25*(1-3), 3-27.

Hirsch, H. G., & Pearce, D. (2000). The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In *Proceedings of the ISCA ITRW ASR*.

Huang, S.-Y., Tu, W.-H., & Hung, J.-W. (2009). A study of sub-band modulation spectrum compensation for robust speech recognition. In *Proceedings of the ROCLING XXI: Conference on Computational Linguistics and Speech Processing*.

Kanedera, N., Arai, T., Hermansky, H., & Pavel, M. (1997). On the importance of various modulation frequencies for speech recognition. In *Proceedings of the European Conference on Speech Communication and Technology*.

Kao, Y.-C., Wang, Y.-T., & Chen, B. (2014). Effective modulation spectrum factorization for robust speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association*, 2724-2728.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788-791.

Lee, D. D., & Seung, H. S. (2000). Algorithms for Non-negative Matrix Factorization. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 556-562.

Lin, S.-H., Chen, B., & Yeh, Y.-M. (2009). Exploring the use of speech features and their corresponding distribution characteristics for robust speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *17*(1), 84-94.

Macho, D., Mauuary, L., Noé, B., Cheng, Y. M., Ealey, D., Jouvet, D., Kelleher, H., Pearce, D., & Saadoun, F. (2002). Evaluation of a noise-robust DSR front-end on Aurora databases. In *Proceedings of the Annual Conference of the International Speech Communication Association*.

Pascual-Montano, A., Carazo, J. M., Kochi, K., Lehmann, D., & Pascual-Marqui, R. D. (2006). Nonsmooth nonnegtive matrix facotorization (nsNMF). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(3), 403-415.

Sun, L.-C., Hsu, C.-W., & Lee, L.-S. (2007). Modulation Spectrum Equalization for robust Speech Recognition. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*.

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

Vikki, A., & Laurila, K. (1998), Segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, *25*, 133-147.

Xiao, X., Chng, E. S., & Li, H. (2008). Normalization of the speech modulation spectra for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, *16*(8), 1662-1674.