Computational Linguistics and Chinese Language Processing
Vol. 21, No. 2, December 2016, pp. 35-54

使用字典學習法於強健性語音辨識

The Use of Dictionary Learning Approach for Robustness Speech Recognition

顏必成*、石敬弘*、劉士弘+、陳柏琳*

Bi-Cheng Yan, Chin-Hong Shih, Shih-Hung Liu and Berlin Chen

摘要

在有雜訊的環境下，自動語音辨識系統(Automatic Speech Recognition, ASR)的效能往往會有明顯衰退的現象。本論文旨在研究語音強健性技術，希望能夠透過語音特徵的調變頻譜(Modulation Spectrum)正規化以萃取出較具有強健性的語音特徵。為此，我們使用 K-奇異值分解(K-SVD)的字典學習法(Dictionary Learning)於分解調變頻譜的強度(Magnitude)成分，在最小化還原訊號誤差且在其權重矩陣稀疏性的限制下，希望能獲取較具強健性的語音特徵。此外，因調變頻譜強度成分皆為正值，所以我們提出非負K-SVD的方法來解決這個議題，希望能增進自動語音辨識系統在抗噪上的效能。本論文的所有實驗皆於國際通用的 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係數(Mel-Frequency Cepstral Coefficient, MFCC)之基礎實驗和其它常見的調變頻譜分解方法，我們所提出的字典學習法與其改進方法皆能顯著地降低語音辨識錯誤率。最後，我們也嘗試將所提出的字典學習方法與一些經典的強健性技術結合，如：進階前端標準法(Advanced Front-End, AFE)、變異數正規化法(Cepstral Mean and Variance Normalization, CMVN)、統計圖等化法(Histogram Equalization, HEQ)，以驗證其實用性。

關鍵字: 強健性、自動語音辨識、調變頻譜、稀疏編碼、字典學習法。

*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
E-mail: {60447055S, 60447003S, berlin}@ntnu.edu.tw
+中央研究院資訊科學研究所
Institute of Information Science, Academia Sinica
E-mail: journey@iis.sinica.edu.tw

Abstract

The performance of automatic speech recognition (ASR) often degrades dramatically in noisy environments. In this paper, we present a novel use of dictionary learning approach to normalizing the magnitude modulation spectra of speech features so as to retain more noise-resistant and important acoustic characteristics. To this end, we employ the K-SVD method to create sparse representations for a common set of basis vectors that span the intrinsic temporal structure inherent in the modulation spectra of clean training speech features. In addition, taking into account the non-negativity property of amplitude modulation spectrum, we utilize the nonnegative K-SVD method, paired with the nonnegative sparse coding method, to capture more noise-robust features. All experiments were conducted on the Aurora-2 corpus and task. The empirical evidence shows that our methods can offer substantial improvements over the baseline NMF method. Finally, we also integrate the proposed variants of the K-SVD method with other well-known robustness methods like Advanced Front-End (AFE), Cepstral Mean and Variance Normalization (CMVN) and Histogram Equalization (HEQ) to further confirm their utility.

Keywords: Robustness, Automatic Speech Recognition, Modulation Spectrum, Sparse Coding, Dictionary Learning.

1. 緒論

語音是人類最常使用的一種訊息表達方式。在日常生活中，語音往往帶有大量且重要的訊息，因此我們對語音訊號進行處理、分析，毫無疑慮的非常具有發展性，而語音辨識藉由將語音訊號轉換成文字為目標，無論是在語意情感分析、語言輔助學習、智慧機器語音識別上有著相當廣泛的應用。

然而大多數的自動語音辨識系統，在不被干擾的情況下，皆能獲得良好的語音辨識效果，但是在現實環境中，自動語音辨識卻往往因為測試環境與訓練環境不匹配(Mismatch) (Tabrikian, Fostck & Messer, 1999)，使得此系統之效能衰退之現象。上述所造成環境不匹配問題的種種因素包含了：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(Li, Deng, Gong & Haeb-Umbach, 2014)，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。

近年來，字典學習(Dictionary Learning)方法被廣泛地應用在圖像(Lu, Shi & Jia, 2013)、語音處理之領域(Gemmeke, Viratnen & Hurmalainen, 2011) (He, Sun & Han, 2015)，其核心概念是利用字典來線性地表示訊號並獲得其稀疏表示(Sparse Representation)，字典學習是從字典(Dictionary)中選取少量的原子(Atoms)來表示訊號，其中每一個原子都可以當作

是一個基礎訊號的表達，而所有原子組成的集合稱為字典。在字典學習方法中，我們可以直接挑選經過處理過後的訊號成為範本字典(Gemmeke *et al*., 2011)，或者是由其他自動學習字典的方法來求得字典，一般常見的字典學習方法有最優方向法(Method of Optimal Directions) (Engan, Aase & Husoy, 1999)、K-奇異值分解法(K-SVD) (Aharon, Elad & Bruckstein, 2006)、隨機梯度下降法(Stochastic Gradient Descent) (Bottou, 1998)及線上字典學習法(Online Dictionary Learning) (Mairal, Bach, Ponce & Sapiro, 2010)。另外在字典學習法中，其原子相對應的權重也須要一併更新，關於權重的更新方式可以由範數的限制做區分，0-式範數(0-Norm)常見的方法為匹配追蹤演算法(Matching Pursuit, MP) (Mallat & Zhang, 1993)、正交匹配追蹤演算法(Orthogonal Matching Pursuit, OMP) (Pati, Rezaiifar & Krishnaprased, 1993)，上述兩種方式都是透過計算殘差與原子的關聯程度來求取權重。1-式範數(1-Norm)常見的方法為基礎追求法(Basis Pursuit, BP) (Chen, Donoho & Saunders, 2001)以及最小絕對壓縮選擇法(LASSO) (Tibshirani, 1996)，此兩種方法將目標函數視為最佳畫圖函數，並透過迭代更新求得其解。

本論文旨在探究使用字典學習法以及一些改進方法來分解調變頻譜強度成分，以獲得較具強健性的語音特徵。字典重建是字典學習方法的主要問題，其目標在於如何從原始訊號中學習出具有代表性的原子來組成字典，且在字典學習方法裡通常被隨著使用稀疏編碼來求取原子的權重，而稀疏編碼的目的在於將訊號表示為各個原子的稀疏線性組合，期望能夠求取具調變頻譜局部性的重要資訊。在本論文中，我們分別使用了 K-SVD 演算法搭配匹配追蹤演算法以及正交匹配追蹤演算法，來得到乾淨的語音調變頻譜強度成分。另一方面，因調變頻譜強度成分皆為正值，所以我們提出使用非負 K-SVD 搭配非複數稀疏編碼(Hoyer, 2004)來解決這個議題，以增進自動語音辨識系統在抗噪上的效能。此外，我們嘗試將字典學習方法與一些經典的特徵強健性技術結合，如：進階前端標準法(Advanced Front-End, AFE)、變異數正規化法(Cepstral Mean and Variance Normalization, CMVN)、統計圖等化法(Histogram Equalization, HEQ)，以驗證這些改進方法之實用性。

2. 相關文獻

在語音辨識中，強健性語音特徵技術主要有兩個方法，第一是以模型為基礎的強健性技術(Model-based Technique)，第二是以語音特徵為基礎的強健性技術(Feature-based Technique)，分別介紹如下：

第一，以模型為基礎的強健性技術是使用少量的測試環境之調適語料來對聲學模型進行調整，使聲學模型可以去近似於輸入雜訊語音的機率分布參數，達到降低環境不匹配的情形。常見的技術有最大相似度線性回歸法(Maximum Likelihood Linear Regression, MLLR) (Leggetter & Woodland, 1995)、最大事後機率法則(Maximum a Posteriori, MAP) (Gauvain & Lee, 1994)、平行模型結合法(Parallel Model Combination, PMC) (Gales & Young, 1996)、向量泰勒級數(Vector Taylor Series, VTS) (Kim, Un & Kim, 1998)、遺失特徵理論(Missing Feature Theory, MFT) (Van Segbroeck & Van Hamme, 2011)。

第二，以語音特徵為基礎的強健性技術是在不更改聲學模型的情況下，利用乾淨的

語音特徵去訓練，期望將帶有雜訊的語音特徵還原成乾淨的語音特徵，本文使用語音參數正規化法(Feature Normalization)，此方法目的在正規化語音特徵本身的特徵值及統計分布，再利用測試語音特徵的特徵值來消除雜訊干擾所帶來的影響，此方法常見的技術有倒頻譜平均值減去法(Cepstral Mean Subtraction, CMS) (Viikki & Laurila, 1998)、變異數正規化法(Cepstral Mean and Variance Normalization, CMVN) (Viikki, Bye & Laurila, 1998)、統計圖等化法(Histogram Equalization, HEQ) (de la Torre *et al*., 2005)、倒頻譜平均值與變異數正規化結合自回歸動態平均濾波法(Cepstral Mean and Variance Normalization plus Auto-regressive-moving Average Filtering, MVA) (Chen & Bilmes, 2007)。

語音參數正規化法的特色是有效且容易實現於大部分的自動語音辨識系統，但是只適用於作用於噪音緩慢變化的情形，並且其改進效果往往是有限的。而語音模型調適法，在辨識上可以有較好的效果，由於此方法會根據噪音環境來更新聲學模型，所以此方法在實作上非常的耗時。

此外，與本論文最相關的研究是語音特徵的調變頻譜強健性技術(張庭豪，2015)，其論文實做了多種非負矩陣分解法處理於調變頻譜的強健性技術。並在稀疏非負矩陣分解法(Sparse Nonnegative Matrix Factorization, SNMF) (Hoyer, 2004)針對基底矩陣做稀疏化的限制，其結果顯示對基底矩陣稀疏化對語音強健性的辨識結果是有相當的助益。因此我們利用稀疏編碼的方法，並且延伸到字典學習法。利用帶有稀疏性的方法，更精確地表示語音訊號。我們希望在訓練階段時，讓字典學得不受噪音干擾的乾淨語音特徵，並在測試時讓帶有噪音的語料透過乾淨語音特徵字典在不超過誤差以及範數的限制下求取其對應的稀疏權重，以此還原噪音語料，期望能降低環境不匹配性並獲得較優良的辨識效果。

3. 調變頻譜正規化法

3.1 調變頻譜之簡介

對於一語音特徵時間序列$x[n]$而言，其調變頻譜定義如下：

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

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)，$x[n]$代表某一維度語音特徵時間序列，$X[k]$代表語音特徵時間序列$x[n]$的調變頻譜。式(1)可看出調變頻譜可以廣泛的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-Sampled)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)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(Chen & Bilmes, 2007)。而調變頻譜之低頻成分(約 1Hz 至 16Hz)對於語音辨識精確度也有密切的關係，潛藏了最重要的語意資訊。其中，最重要的是位於 4 Hz，有學者指出，4 Hz 是人耳聽覺最為敏感之調變頻率(Gales & Young, 1996)；也有學者認為，4 Hz 為人類大腦皮層感知之重要調變頻率(Kim *et al.*, 1998)。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。接著處理更新後的強度成分會與原始相位成分結合，再經由反傅立葉轉換(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之問題，使自動語音辨識系統使用新的語音特徵時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。

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

假設當各種音素在理想環境中占的比例接近一致時，每一維度特徵的調變頻譜之平均值應該為一個定值(Huang, Tu & Hung, 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。

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

除了要正規調變頻譜強度成分之平均值，也要正規其變異數(張庭豪，2015)。假設特徵向量參數之平均值在理想環境中比例接近一致時，平均值應為零，且特徵向量參數之分布可以利用變異數來進行檢測：

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

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變異數；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變異數，$\left|\tilde{X}[k]\right|$便是更新過後的調變頻譜強度成分。

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

利用非線性的轉換(Nonlinear Transform)，不只將調變頻譜強度成分之平均值與變異數作正規化，也使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(Viikki & Laurila, 1998)：

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

在式(4)中，$F_X(\cdot)$為單一語句的調變頻譜強度的機率分布(Probability Distribution Function, PDF)，F_{ref}則是利用所有訓練語句之調變頻譜強度所求的參考機率分布，$\left|\tilde{X}[k]\right|$便是更新過後的調變頻譜強度成分。

3.5 分頻段調變頻譜統計正規化法

此方法的概念是想要改進調變頻譜統計正規化法；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同一隨機變數(Random Variable)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(Viikki *et al.*, 1998)。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細的頻寬，子頻段的數量也比較多，而高調變頻率便持有相反的特性。由於掌握住低頻成分的資訊，根據學者的實驗數據，顯示出將調變頻率分頻段且正規化的做法，能比全頻帶正規化的方式獲得較好的效能。

4. 使用字典學習法於調變頻譜分解

4.1 字典學習法介紹

字典學習法是一種利用字典來表示資料的方法。其精髓在於透過學習而來的字典(Dictionary)，配合稀疏編碼(Sparse Coding)挑選出字典中重要的原子(Atoms)；並且使得一些多餘的資訊(Redundancy)稀疏，最後以線性組合的方式近似還原出原始訊號(Tosic & Frossard, 2011)。相對於其他降維方法的技術如：主成分分析法(Principal Component Analysis)、線性鑑別分析(Linear Discriminant Analysis)，字典學習法沒有拘泥於減少資料維度的特性；反之，字典學習法追求的是如何習得資料中重要的特徵，以及潛在的資料意義。所以字典通常都是透過遠大於輸入資料的維度，鉅細靡遺表示資料。就計算複雜度而言，相比之下較花時間且並非每種資料都適合升維的字典學習法，我們也可以依照資料的屬性字典設計成降維的字典。一般而言，字典學習法可以依照資料選取的概念分成三種常見的方法，分別是：機率學習法(Tosic & Frossard, 2011) (Wipf & Rao, 2004) (Probabilistic Learning Methods)、向量編碼與分群法(Gersho & Gray, 1991) (Vector Quantization or Clustering Methods)、特殊結構法(Tosic & Frossard, 2011) (Particular Construction Methods)。關於機率字典學習法(Wipf & Rao, 2004)，採用稀疏貝氏學習法(Sparse Bayesian Learning)去解決字典的目標函數以及字典更新問題。而特殊結構字典學習法(Yaghoobi, Daudet & Davies, 2009)，他們探討著針對不同輸入信號時，可以在目標函數加入不同的參數矩陣函數(Parametric Function)使得字典最佳化。而向量編碼與分群

法除了著名的 K-SVD 以外(Aharon *et al*., 2006)，近年來有些研究學者基於 K-SVD 資料分布是高斯分布的假設而提出拉式分布的說法，改善 K-SVD 重建項(Reconstruction Term)2-式範數的限制為 1-式範數，其提出的方法為 l_1-K-SVD(Mukherjee, Basu & Seelamantula, 2016)，其方法能有效的改善 2-式範數過於平滑(Over-smoothing)的問題，其方法除了提升了辨識率，以及加快了收斂速度，並拉近了原始字典與習得字典的尤拉距離(Euclidean Distance)。

4.2 K-SVD字典學習法

K-SVD 字典學習法是由 Aharon 等提出(Aharon *et al*., 2006)，其根本想法源自於向量編碼問題(Vector Quantization)以及 K-means 演算法(Gersho & Gray, 1991)，屬於廣義的 K-means (Generalized K-means)演算法。Aharon 等認為編碼問題可以針對字典原子加入 0 式範數(0-Norm)的限制使得字典在近似還原輸入資料時，能刪減掉不必要的資訊，並重新表示為式(5)：

$$min_{D,X}\{\|Y - DX\|_F^2\} \quad subject\ to\ \forall i\ ,\ \|X_i\|_0 \le T_0 \tag{5}$$

其中 Y 為原始訊號、D 為欲學習的字典、X 為字典相對應的權重矩陣、T_0是一個非零的實數。迭代更新的關係式(5)也可以等價表示為式(6)：

$$min_{D,X} \sum_i \|X_i\|_0 \quad subject\ to\ \|Y - DX\|_F^2 \le \epsilon \tag{6}$$

其中ϵ為一個給定的錯誤容忍值。

在 K-SVD 字典學習法中，我們透過類似 K-means 迭代更新的方式求解算式(6)。在稀疏編碼階段中，我們可以使用任何匹配追蹤的方法找出權重矩陣 X，並且透過 0 式範數的限制，讓權重矩陣中的向量X_i內的元素(Element)個數不會超過T_0，藉此使得矩陣 X 稀疏。接著在字典更新階段，我們是以字典的行(Column)向量D_i為單位，亦記作原子(Atom)逐步迭代更新。透過殘差矩陣(Residual Matrix) E_k，如式(7)，其中x^k為權重矩陣(Weight Matrix)X 的第 k 列(Row)向量、d_k為字典的第 k 行。E_k代表著整個重建信號 DX 少了第 k 組原子(Atom)之權重向量後與輸入信號(Input Signal)Y 的差。

$$E_k = Y - \left(DX - d_k x^k\right) \tag{7}$$

接著我們定義了ω，ω是為了取得E_k中所對應的權重矩陣不為 0 的列向量而存在的數字集合，其定義式如下式(8)。其中x^k表示權重矩陣中第 k 列的向量(Row vector)；而x_i代表權重矩陣中的第 i 行的向量(Column vector)。

$$\omega = \{i \big| 1 \le i \le N,\ x^k(i) \ne 0\} = \{i | 1 \le i \le N, x_i(k) \ne 0\} \tag{8}$$

當E_k加入ω集合後，得到的非零的元素矩陣，我們重新記做E_k^w稱為限制殘差矩陣(Restricted Residual Matrix)。來對E_k^w使用奇異值矩陣分解法取得E_k^w中第一筆重要的資訊，分解後所得到的左奇異值向量矩陣我們記為 U；奇異值矩陣我們記作Δ；右奇異值向量矩陣寫成V^T。那麼新的原子(Atom) d_k等於第一筆左奇異值向量的值u_1；而對應的權重向量(Weight Vector) X^k可以透過第一個奇異值與第一筆右奇異值向量的乘積，$\Delta_{(1,1)} \cdot V_1$來更

演算法 1、K-SVD 字典學習法

初始階段:

$D^0 \in R^{n\times K}$，設 $J = 1$。

1、稀疏編碼階段(Sparse Coding Stage):本論文使用 MP、OMP 的演算法求得x_i：

for $i = 1,2, \ldots\ldots, N$

$$min_x\{\|y_i - Dx\|_2^2\} \quad subject\ to\ \ \|x\|_0 \leq T_0$$

2、字典更新階段(Dictionary Update Stage):

- 定義:字典 D 的行向量:d_k、權重矩陣的列向量:x^k、而有被原子使用到的對應權重行向量記錄成集合$\omega_k = \{i|1 \leq i \leq N, x^k(i) \neq 0\}$。
- 計算

$$E_k = Y - (DX - d_x x^k)，$$

- 殘差矩陣E_k透過ω_k取得E_k中不為零的行向量，得到新組成矩陣E_k^ω

  在利用 SVD 分解法$E_k^\omega = U\Delta V^T$。

  更新: $d_k = u_1$

  更新: $x^k = \Delta(1,1) \cdot v_1$

$J = J + 1$

反覆 1、2 直到收斂

新，更新式(9)如下：

$$\text{update: } d_k = u_1$$
$$X^k = \Delta_{(1,1)} \cdot V_1 \tag{9}$$

完整的 K-SVD 作法如演算法 1 所示。

4.3 非負K-SVD字典學習法

非負 K-SVD 字典學習法是和 K-SVD 字典學習法同時誕生的，源自於同一組研究團隊。為了使 K-SVD 更能夠完整的描述輸入資料為正數時的情況；進而在 K-SVD 求解權重以及更新字典時加入非負數的限制，而新生成的演算法稱為 NN-K-SVD。NN-K-SVD 為了使輸出的字典(Dictionary)以及權重矩陣(Weight)皆為正數，所以在學習的稀疏編碼階段(Sparse Coding Stage)時，求解權重的方法利用非負數稀疏編碼法(Non-negative Sparse Coding) (Hoyer, 2004)。

演算法 2、非負 K-SVD 字典更新規則

初始階段:

令 $d_i = \begin{cases} 0 & u_i<0 \\ u_i & \text{otherwise} \end{cases}$，$\mathrm{x(i)} = \begin{cases} 0 & v_1(i)<0 \\ v_1(i) & otherwise \end{cases}$，

其中u_1、v_1是$E_r^{\omega_k}$接過奇異值分解後，第一筆奇異值左向量與右奇異值向量。

重複 J 次:

(1) 令:$\mathrm{d} = \frac{E_r^{\omega_k} x}{x'x}$，　當 $\mathrm{d(i)} = \begin{cases} 0 & d(i)<0 \\ d(i) & otherwise \end{cases}$

(2) 令:$\mathrm{x} = \frac{d' E_r^{\omega_k}}{d'd}$，　當 $\mathrm{x(i)} = \begin{cases} 0 & x(i)<0 \\ x(i) & otherwise \end{cases}$

在非負 K-SVD 演算法中，學習的方法與 K-SVD 如出一轍。不過為了滿足非負數的限制，我們對目標函數式(5)了一些調整。我們在最小平方法裡加入非負數的限制。也就是說對於字典，我們只保留前 L 大的原子(Atom)來參與字典學習，如式子(10)。

$$min_x \|y - D^L x\| \quad \text{s.t.} \quad x \geq 0 \tag{10}$$

在稀疏階段(Sparse Coding Stage)，我們使用上述提及的非負稀疏編碼法(NNSC)中求取矩陣 S 的方式，求解權重 x。而在字典更新階段(Dictionary update stage)，為了使得原子d_k (Atom)更新後也為正數，則加入原子之值為正的限制，如式子(11)所示：

$$min_{d_k, x^k} \left\| E_r^{\omega_k} - d_k x^k \right\| \quad s.t. \quad d_k, x^k \geq 0 \tag{11}$$

而更新殘差矩陣$E_r^{\omega_k}$利用 SVD 分解時可能產生負數，我們採用上圖的演算法，如果第一次 SVD 分解後的左右奇異值向量皆負數；則同乘(-1)。接下來奇異值向量內的元素(Element)小於零則設為零，其他大於零的數字則保留不變。開始重複 J 次使得$E_r^{\omega_k}$越來越接近$\mathrm{d}x'$。完整非負 K-SVD 更新規則如演算法 2 所示。

4.4 權重的更新方法

4.4.1 匹配追蹤

匹配追蹤(Matching Pursuit, MP)，是字典學習法第一階段稀疏編碼(Sparse Coding)中，求得權重矩陣(Weight) X 的常見解法。演算法概念為貪婪法則。首先將輸入信號(Input signal) x 令成冗餘項量(Residual) r。再利用投影量的方式衡量出與冗餘向量 r 相關程度最大的原子(Atom) d_i後，則冗餘向量 r 與最相關的原子 d_i的內積值式(12)即為對應權重 α 在第 i 維度的分數。

$$\left(d_i^T \cdot r\right) \tag{12}$$

演算法 3: 匹配追蹤演算法(MP)

1. 輸入資料:輸入信號 x、字典 D。
2. 輸出資料:權重向量 α。
3. 目標函數:

$$\min_{\alpha \in R^n} \|x - D\alpha\|_2^2 \qquad s.t \quad \|\alpha\|_0 \leq L$$

4. 初始步驟:

$$\alpha \leftarrow 0 \text{、} r(\text{冗餘向量}) \leftarrow x$$

5. while $\|\alpha\|_0 < L$ do
   - 找尋與冗餘向量(Residual Vector)關聯程度最大的原子。

$$\hat{\imath} \leftarrow argmax_{\hat{\imath}=1,\ldots\ldots,p} |(d_{\hat{\imath}}^T \cdot r)|$$

   - 更新冗餘向量以及對應的權重(Weight)

$$\alpha[\hat{\imath}] \leftarrow \alpha[\hat{\imath}] + (d_{\hat{\imath}}^T \cdot r)$$

$$r \leftarrow r - (d_{\hat{\imath}}^T \cdot r) d_{\hat{\imath}}$$

6. end while

之後再將輸入信號 r 減去信號 r 在最相關的原子d_i上的投影量分量即為式(13)，所得到的冗餘向量 r 則為下一次迭代的輸入信號。

$$r = r - \left(d_i^T \cdot r\right) d_i \tag{13}$$

完整匹配追蹤的算法如演算法 3 如示。

4.4.2 正交匹配追蹤

正交匹配追蹤法(Orthogonal Matching Pursuit, OMP)是匹配追蹤法的改良方法。正交匹配追蹤法(OMP)在衡量冗餘向量與原子的關聯程度時考慮的是正交投影量，而不是投影量。而計算正交投影量的方法為 Gram-Schmidt 正交化法，其中更新冗餘向量 τ 及權重 α 如式(14)及(15)所示：

$$\tau \leftarrow (\mathrm{I} - D_\tau (D_\tau{}^T D_\tau)^{-1} D_\tau{}^T)\mathrm{x} \tag{14}$$

$$\alpha_\tau \leftarrow (D_\tau{}^T D_\tau)^{-1} D_\tau{}^T \mathrm{x} \tag{15}$$

4.4.3 非負數稀疏編碼法

非負數稀疏編碼法(Non-Negative Sparse Coding, NNSC)的更新演算法是參照非負矩陣分解的乘法式更新式(Multiplicative Update Rules) (張庭豪，2015)，其特性為藉著輸入矩陣皆為正數，然而透過乘法更新的關係使得更新後的矩陣也為正數。非負數稀疏編碼是為了求解式(10)而產生的方法。

演算法 4: 非負數稀疏編碼(NNSC)

目標函數:

$$min_{A,S}\frac{1}{2}\|X-AS\|^2+\lambda\sum_{ij}S_{ij}$$

限制條件如下: $\forall_{ij}: X_{ij}$ 、$A_{ij}\geq 0$ 、$S_{ij}\geq 0$、$\lambda\geq 0$ and

$\forall_i: \|a_i\|=1$，a_i *是* A*行向量*$(column)$.

(1) 初始階段:透過隨機變數並設變數為正數初始A^0與S^0，其 中 A 矩陣的行向量須為單位向量，並設時間參數 t=0。

(2) 反覆迭代直到收斂:

- 更新 A 矩陣:

  $A'=A^t-\mu(A^tS^t-X)(S^t)^T$。

  任何負數出現在A'中都設為 0，重新使得A'行向量成完單位向量，並設定$A^{t+1}=A'$。

- 更新 S 矩陣，運用乘法更新式(Multiplicative Update Rule):。

  $$S^{t+1}=S^t.*(A^TX)./(A^TAS^t+\lambda)$$

t=t+1

$$min_{A,S}\frac{1}{2}\|X-AS\|^2+\lambda\sum_{ij}S_{ij}$$

限制條件為: $\forall_{ij}: X_{ij}$ 、$A_{ij}\geq 0$ 、$S_{ij}\geq 0$ and $\forall_i: \|a_i\|=1$ (16)

其中 X 為輸入矩陣。X、S、A 為非負矩陣且 A 矩陣的行向量須為單位向量、λ 為平衡係數且$\lambda\geq 0$。

求解時我們先隨機初始非負矩陣 A 與矩陣 S 來滿足式(10)；其中 A 的行向量須為單位向量，並設時間參數 t=0。透過類似非負矩陣分解的乘法更新式反覆迭代更新 A 與 S 使得式(16)得到最佳解。NNSC 的求解如演算法 4 所示。

5. 實驗結果與分析

5.1 實驗語料庫

Aurora-2 是歐洲電信標準協會(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。語料庫內有乾淨及附有雜訊的語音，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，故測試集中所有加成性噪音是與訓練語料不同的語句，而只有測試集 C 的通道效應與訓練語料不同。

5.2 實驗設定

在本文中的基礎實驗是採用梅爾倒頻譜係數(MFCC)做為語音特徵參數，取樣頻率(Sampling Rate)為 8000Hz，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。每一個音框的特徵使用 13 維梅爾倒頻譜係數(第 1 維至第 12 維還有第 0 維)，加上其一階差量計算和二階差量計算，共 39 維之特徵參數。在特徵的強健性處理方法，本文在處理特徵時，只針對 13 維的靜態特徵參數(Static Feature)進行處理，處理完成後才額外將一階差量和二階差量加入。

5.3 辨識效能評估方式

辨識效能的評估方式是採用美國標準與科技組織(NIST)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。評估方式是以詞正確率(Word Accuracy Rate)為主，計算正確轉寫文句詞串與辨識詞串彼此間，詞的取代個數(Substitutions)、詞插入個數(Insertions)和詞刪除個數(Deletions)：

$$\text{詞正確率}(\%) = \frac{\text{詞正確辨識個數} - \text{詞插入個數}}{\text{輸入詞總數}} \times 100\% \tag{17}$$

本文參照國際學者之設定，在對每一種噪音的訊噪比的結果作加總的動作時，去掉極端的訊噪比 clean 跟-5 dB，只計算範圍 20dB 到 0dB 中的平均詞精確率或平均詞錯誤率的結果再取其平均值。本論文的全部實驗皆是利用平均詞精確率來評估計算辨識的結果。

5.4 基礎實驗結果

首先，最基本的實驗是在以梅爾倒頻譜係數(MFCC)於乾淨語料訓練下所得到的辨識結果。本論文也比較常見的時間序列域特徵正規化法，包含有倒頻譜平均值與變異數正規化法(CMVN)、統計圖等化法(HEQ)、ETSI 所提供的進階前端標準(Advanced Front-End Standard, AFE)，以及作用在調變頻譜上的調變頻譜平均值與變異數正規化法(SMVN)。另外，在調變頻譜上的矩陣分解方法中我們以常用的非負矩陣分解法(NMF)來當作是一

表 1. 基礎實驗數據結果

[Table 1. Recognition accuracy rates (%) averaged over different noise types and different SNRs for several representative acoustic feature normalization methods.]

更新特徵	Set A	Set B	Set C	Avg.
MFCC	54.87	48.87	63.95	54.29
SMVN	59.02	63.60	58.49	60.75
CMVN	75.93	76.76	76.82	76.44
HEQ	80.03	82.05	80.10	80.85
AFE	87.68	87.10	86.29	87.17
MFCC+NMF	67.09	70.98	68.22	68.87
CMVN+NMF	83.56	85.51	83.27	84.28
HEQ+NMF	83.84	85.88	83.70	84.63
AFE+NMF	87.74	87.65	86.32	87.42

個強基礎實驗(Strong Baseline)，此非負矩陣分方法用可在不同的時間序列域中，如 NMF 結合基本的 MFCC(記做 MFCC+NMF)、結合 CMVN(記做 CMVN+NMF)、結合 HEQ(記做 HEQ+NMF)以及結合 AFE(記做 AFE+NMF)，其實驗結果如表 1 所示。表 1 顯示出幾個實驗現象。第一，正規化法消去了語音特徵的平均值及其變異數，其原理就是消去了穩定的通道效應且減少語音特徵分布的差異，因此能有效地抗噪，由表 1 可看出常見的時間序列域特徵正規化法，如 CMVN，都能有效地大幅提升辨識正確率(比 MFCC 進步約 20%)，而在調變頻譜上的正規化法 SMVN 雖然相較於 MFCC 也能提升辨識率，但提昇的幅度卻不太大(比 MFCC 進步約 6%)，這有待更進一步的深入探究。第二，統計圖等化法利用正規化特徵參數的整體分布，對特徵參數的統計分布之所有動差進行正規化，比傳統正規化法多考量了更多的統計資訊,因此在抗噪效果上又比傳統正規化法來得好，如表 1 所示，HEQ 比 CMVN 多了約 4%的進步。第三，由知名 ETSI 單位所提供的 AFE 特徵跟其他正規化法相比，會有最好的辨識結果，相較於 MFCC 有 33%的進步，且相較於 HEQ 也有約 7%的進步。最後，不同的時間序列域之調變頻譜的非負矩陣分解方法都能有效抗噪進而提升辨識正確率，在 MFCC 的調變頻譜上做非負矩陣分解可以得到最多的進步(約 14%)，次之的是在 CMVN 的調變頻譜上有約 8%的進步，接著是 HEQ 的調變頻譜上約有 4%的進步，最少的進步是在 AFE 的調變頻譜上(約 0.3%的進步)。在本基礎實驗中，NMF 的基底個數設定是 5(最好的結果)(張庭豪，2015)。

5.5 基於K-SVD字典學習法於調變頻譜分解

在本小節中，我們將探討所提出使用的 K-SVD 字典學習法於調變頻譜分解之實驗。在調變頻譜上，考量到輸入資料的維度是 513 維，然而運算相當的耗費時間，再衡量到須與前人比較應用在調變頻譜的抗噪技術，所以我們的字典仍屬於降低資料維度的字典，關於此小節實驗設定部分，本論文所使用的字典維度分別設定為 5、10 及 30。

我們的實驗目的是希望透過學習而得到的乾淨語音字典近似噪音語音信號，藉此還原噪音訊號。字典學習法的實驗分成兩個階段，分別是訓練階段，以及測試階段。在訓練階段時先運用四個資料集的乾淨語句調變頻譜特徵，當作輸入資料後並運用 K-SVD 字典學習法搭配權重的解法求得乾淨的字典以及乾淨的權重矩陣。而在此階段時，我們捨棄乾淨的權重矩陣，只保留乾淨的語音字典特徵。接下來在測試階段時，輸入的資料為具有噪音調變頻譜，並且以一句話為單位地利用在訓練階段時所創造的乾淨字典配合權重的解法求得該句對應的權重。最後再將測試語句所對應的權重矩陣乘上訓練時的乾淨字典來還原調變頻譜的特徵。

表 2. 使用 K-SVD 搭配 MP 求解權重於 MFCC、CMVN、HEQ、AFE 等特徵
[Table 2. Recognition accuracy rates (%) for MFCC, CMVN, HEQ, and AFE features for use in K-SVD integrated with MP method.]

更新特徵	Set A	Set B	Set C	Avg.
MFCC+ Dict(5)	63.57	69.00	63.61	65.39
MFCC+ Dict (10)	62.82	68.61	61.50	64.31
MFCC+ Dict (30)	63.70	70.14	59.95	64.59
CMVN+ Dict (5)	82.40	84.23	83.10	83.24
CMVN+ Dict (10)	81.97	83.94	82.72	82.87
CMVN+ Dict (30)	80.83	82.31	81.38	81.51
HEQ+ Dict (5)	81.73	84.50	82.58	82.94
HEQ+ Dict (10)	79.96	82.82	80.85	81.21
HEQ+ Dict (30)	78.85	81.89	79.41	80.05
AFE+ Dict (5)	85.98	87.02	84.87	85.96
AFE+ Dict (10)	85.66	86.66	84.35	85.56
AFE+ Dict (30)	86.09	86.94	84.39	85.81

本論文所使用的 K-SVD 搭配兩種不同求解權重的方法，一為 MP，另一為 OMP，並且作用在不同的語音特徵上，如 MFCC、CMVN、HEQ 及 AFE，實驗結果分別列在表 2 及表 3。對於實驗的結果，我們有四點發現。第一，由表 2、表 3 的實驗結果我們可以得知辨識結果並非隨著字典的原子數增加而變好。會導致此現象的原因可能因為字典是

學習乾淨語音訊號，當字典維度升高時，代表著字典把乾淨特徵學得越詳細。而當噪音影響加劇時，訊號中乾淨的語音特徵相對地不明顯，此時字典維度越高反而越不能辨識出噪音訊號是屬於哪一種乾淨語音特徵，所以在噪音影響下反而無法提升辨識效果。第二，觀察表 3 中 AFE 特徵下三個維度(5、10 及 30)的辨識結果。可以發現字典維度 30 比起維度 5 時的辨識率提升了 0.3%。關於此現象的原因，我們覺得是由於 AFE 特徵已經對噪音影響乾淨訊號的不匹配性有了非常有效的處理。故將字典為度提高時，能更仔細的描述噪音語音特徵。第三，我們可以從表 2、表 3 比較ＯＭＰ以及ＭＰ求取權重對於語音辨識的效果。可以發現不同特徵且字典維度 5 的情況下使用 OMP 求取權重的辨識效果除了 HEQ 特徵降低 0.42%的辨識正確率外，其他特徵比起 MP 的辨識效果都比較好。由於 OMP 求取權重的方法是衡量殘差和原子們的垂直分量，此方法可以除了確保每次迭代求取的權重都是最佳解以外，還可以加快權重收斂的速度。

表 3. 使用 K-SVD 搭配 OMP 求解權重於 MFCC、CMVN、HEQ、AFE 等特徵
[Table 3. Recognition accuracy rates (%) for MFCC, CMVN, HEQ, and AFE features for use in K-SVD integrated with OMP method]

更新特徵	Set A	Set B	Set C	Avg.
MFCC+ Dict(5)	65.82	71.00	65.33	67.38
MFCC+ Dict (10)	62.11	68.77	61.32	64.07
MFCC+ Dict (30)	62.02	67.10	59.78	62.97
CMVN+ Dict (5)	83.65	85.74	84.06	84.48
CMVN+ Dict (10)	79.54	83.95	82.67	82.05
CMVN+ Dict (30)	83.28	85.34	83.59	84.07
HEQ+ Dict (5)	81.22	84.18	82.18	82.52
HEQ+ Dict (10)	79.11	82.29	79.99	80.47
HEQ+ Dict (30)	75.73	78.98	77.18	77.30
AFE+ Dict (5)	86.09	86.91	85.02	86.01
AFE+ Dict (10)	85.88	86.71	84.62	85.74
AFE+ Dict (30)	86.56	87.47	84.88	86.31

5.6 基於非負K-SVD字典學習法於調變頻譜分解

由於本實驗是針對調變頻譜強度部分做特徵增強，故經處理後的強度若出現負值將不符合強度的物理意義。然而 K-SVD 字典學習法在訓練階段時，所用到的 SVD 分解法，將會把輸入資料分解出負數元素(Element)連同後續的更新步驟中一併地加入實驗所創造的乾淨字典。且在實驗的最後階段，透過具有負數的字典乘上權重矩陣所還原出來的調變頻譜也可能會有負數。有鑑於前文所提及的狀況，我們透過非負 K-SVD 字典學習法有效

地改善 K-SVD 在更新時產生負數的問題。

非負 K-SVD 與 K-SVD 的差異除了求取權重時採用的方法為使用乘法更新式的非負稀疏編碼外，非負 K-SVD 在求解目標函數也加入非負的限制，並在更新字典以及權重時使用相當嚴格的演算法確保字典以及權重在每次迭代後的結果皆為正數。

表 4. 使用非負 K-SVD 與非負稀疏編碼求解權重於 MFCC、CMVN、HEQ、AFE 等特徵
[Table 4. Recognition accuracy rates (%) for MFCC, CMVN, HEQ, and AFE features for use in NN-KSVD integrated with NNSC method]

更新特徵	Set A	Set B	Set C	Avg.
MFCC + Dict(5)	65.59	71.22	64.56	67.12
CMVN + Dict(5)	83.80	85.83	84.24	84.62
HEQ + Dict(5)	82.24	85.16	83.40	83.60
AFE + Dict(5)	87.50	88.27	86.84	87.54

由於表 2 和表 3 的數據呈現出在維度為 5 時有最好的辨識結果，所以本小節的實驗非負 K-SVD 字典學習法維度設定就只有設定為 5。關於非負 K-SVD 字典的實驗結果，我們可以歸納出三個要點。首先，與表 1 的四種常見時間域序列特徵相比(如：MFCC、CMVN、HEQ 及 AFE)，我們引入的非負 K-SVD 於調變頻譜領域的辨識結果都比基礎實驗的四種常見特徵有效。MFCC 的辨識率提升了 12.83%、CMVN 的辨識率提升了 8.18%、HEQ 的辨識率提升了 2.75%、AFE 的辨識率提升了 0.37%。第二，比較表 4 與表 3 並針對字典維度 5 的情況下，我們比較 K-SVD+OMP 與 NNK-SVD+NSC 的辨識效果。發現除了 MFCC 降低了 0.26%以外，CMVN、HEQ 以及 AFE 三種不同特徵的辨識結果，皆為非負 K-SVD 較為優異，其中 CMVN 提升了 0.14%、HEQ 提升了 1.08%、AFE 提升了 1.53%。由此可知，當考量調變頻譜的特性並加入非負的限制後，對於辨識效果的提升是有幫助的。第三，我們與強基礎實驗相比，透過觀察表 4 以及表 1 的 MFCC+NMF、CMVN+NMF、HEQ+NMF、AFE+CNMF，可以得知 NMF 是一個很有用的抗噪技術，但本論文所提出使用的非負 K-SVD 在輸入特徵為 CMVN 以及 AFE 時，相較於 NMF 方法在 CMVN 與 AFE 分別有 0.34%以及 0.12%的辨識進步率。最後，由第一點分析延伸。我們覺得將字典學習法以及稀疏編碼加入非負的限制後，並實作在調變頻譜域上是相當具有前瞻性的，然而相關實驗仍持續進行中，期望在此非負的限制條件下其辨識結果仍是值得期待的。

6. 結論與未來展望

本論文探討了字典學習法應用在語音特徵的處理，並將之運用在調變頻譜上，希望能夠擷取出更強健性的基底向量，而達到降低訓練語料和測試語料的不匹配性進而達到語音強健性的目的。本論文使用了兩種字典學習的方法，第一種是採用 K-SVD 字典學習法，但我們可以從本論文的實驗觀察到，隨著字典的增大，辨識效果並沒有如期的提升。由

此可知實作在調變頻譜的實驗仍是適合降維的字典。第二種是使用非負的 K-SVD 字典學習方法，該方法在非負的限制下，我們更可以詮釋調變頻譜強度能量皆為正值的概念。其效果相較於使用 K-SVD 字典學習法也有所提昇，抗噪效果在部分特徵的實作上也勝過非負矩陣分解的方法。

在未來展望的方面，我們希望試著使用超完備的字典來表示原始語音訊號，但由於使用 K-SVD 演算法的計算複雜度過於龐大，所以我們期望使用線上字典學習方法來學習字典，希望能夠加快學習字典的收斂速度；再者，我們也希望能嘗試在訓練階段時使用範本字典選取的方法，此方法藉由事先選擇較為重要的語音特徵，來代替字典學習中字典的學習階段。另外，我們也希望能繼續探討在使用非負 K-SVD 演算法時，造成語音辨識效果不升反降的情形。最後在未來，我們希望能透過訊號分離(Source Separation) (Gemmeke *et al*., 2011)的方式解決環境不匹配的問題，如此一來辨識結果應該能得到提升。

參考文獻 Reference

Aharon, M., Elad, M. & Bruckstein, A. M. (2006). The KSVD: An algorithm for designing of overcomplete dictionaries for sparse representations. *IEEE Transactions on Signal Processing*, *54*, 4311-4322.

Bottou, L. (1998). Online algorithms and stochastic approximations. In D. Saad (Eds.), *Online Learning and Neural Networks*. Cambridge, UK: Cambridge University Press.

Chen, C. P. & Bilmes, J. A. (2007). MVA processing of speech features. *IEEE Transactions on Audio Speech and Language Processing*, *15*(1), 257-270.

Chen, S. S., Donoho, D. L. & Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM review*, *43*(1), 129-159.

de la Torre, A., Peinado, A.M., Segura, J. 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.

Engan, K., Aase, S. O. & Husoy, J. H. (1999). Method of optimal directions for frame design. In *Proc. of IEEE International Conference of Acoustic, Speech, and Signal Processing*, *5*, 2443-2446.

Gales, M. J. F. & Young, S. J. (1996). Robust continuous speech recognition using parallel model combination. *IEEE Transactions on Speech and Audio Processing*, *4*(5), 352-359.

Gauvain, J.-L. & Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, *2*(2), 291-298.

Gemmeke, J. F., Viratnen, T. & Hurmalainen, A. (2011). Exemplar-based sparse representations for noise robust automatic speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(7), 2067-2080.

Gersho, A. & Gray, R. M. (1991). *Vector qusntization and signal compression*. Norwell, MA: Kluwer Academic.

He, Y., Sun, G. & Han, J. (2015). Spectrum enhancement with sparse coding for robust speech recognition. *Journal of Digital Signal Processing*, *43*, 59-70.

Hirsch, H. G. & Pearce, D. (2000). The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In *Proc. of ISCA ITRW ASR 2000*, 181-188.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, *5*, 1457-1469.

Huang, S. Y., Tu, W. H. & Hung, J. W. (2009). A study of sub-band modulation spectrum compensation for robust speech recognition. In *Proceeding of ROCLING XXI:Conference on Computational Linguistics and Speech Processing*, 39-52.

Kim, D. Y., Un, C. K. & Kim, N. S. (1998). Speech recognition in noisy environments using first-order vector Taylor series. *Speech Communication*, *24*(1), 39-49.

Leggetter , C.J. & Woodland, P.C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Computer Speech Language*, *9*(2), 171-185.

Li, J., Deng, L., Gong, Y. & Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *22*(4), 745-777.

Lu, C., Shi, J. & Jia, J. (2013). Online robust dictionary learning. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR2013)*, 415-422.

Mairal, J., Bach, F., Ponce, J. & Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, *11*, 19-60.

Mallat, S. & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, *41*(12), 3397-3415.

Mukherjee, S., Basu, R. & Seelamantula, C. S. (2016). ℓ 1-K-SVD: A robust dictionary learning algorithm with simultaneous update. *Signal Processing*, *123*, 42-52.

Pati, Y. C., Rezaiifar, R. & Krishnaprasad, P. S. (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*.

Tabrikian, J., Fostck, G. S. &Messer, H. (1999). Detection of environmental mismatch in a shallow water waveguide. *IEEE Transactions on Signal Processing*, *47*(8), 2181-2190.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267-288.

Tosic, I. & Frossard, P. (2011). Dictionary learning. *IEEE Signal Processing Magazine*, *28*(2), 27-38.

Van Segbroeck, M. & Van Hamme, H. (2011). Advances in missing feature techniques for robust large-vocabulary continuous speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(1), 123-137.

Viikki, O., Bye, D. & Laurila, K. (1998). A recursive feature vector normalization approach for robust speech recognition in noise. In *Proc. of ICASSP*, 733-736.

Viikki, O. & Laurila, K. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, *25*(1-3), 133-147

Wipf, D. P. & Rao, B. D. (2004). Sparse Bayesian learning for basis selection. *IEEE Transactions on Signal Processing*, *52*(8), 2153-2164.

Yaghoobi, M., Daudet, L. & Davies, M. E. (2009). Parametric dictionary design for sparse coding. *IEEE Transactions on Signal Processing*, *57*(12), 4800-4810.

張庭豪(2015)。**調變頻譜分解之改良於強健性語音辨識**(碩士論文)。取自 http://etds.lib.ntnu.edu.tw/cgi-bin/gs32/gsweb.cgi/ccd=rs0dbQ/record?r1=1&h1=0
[Chang, T.-H. (2015). *Several Refinements of Modulation Spectrum Factorization for Robust Speech Recognition* (Master's thesis). Retrieved from http://etds.lib.ntnu.edu.tw/cgi-bin/gs32/gsweb.cgi/ccd=rs0dbQ/record?r1=1&h1=0]