分頻式調變頻譜分解於強健性語音辨識

Sub-band modulation spectrum factorization in robust speech recognition

范顯騰 Hao-teng Fan
國立暨南國際大學電機工程學系
Department of Electrical Engineering
National Chi Nan University
s99323904@mail1.ncnu.edu.tw

蔡益彰 Yi-zhang Cai
國立暨南國際大學電機工程學系
Department of Electrical Engineering
National Chi Nan University
s99323523@mail1.ncnu.edu.tw

洪志偉 Jeih-weih Hung
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

在本篇論文中，我們使用了非負矩陣分解(nonnegative matrix factorization, NMF)技術來強化語音特徵調變頻譜、藉此提升自動語音辨識系統之雜訊強健性，其中，NMF 法為語音之調變頻譜的強度求取一組基底向量，而我們藉由此組基底向量來擷取語音中重要的辨識成分，跟以往基於 NMF 之強健技術不同之處在於兩點：其一，我們利用了正交投影(orthogonal projection)的方式取代原先的迭代方式，使運算速度大幅增加。其二，我們採取分頻帶分解的方式取代原先全頻帶分解，藉此減少計算量。在 Aurora-2 之連續數字資料庫之辨識實驗顯示，上述的新方法相對於基礎實驗而言，能有效提升雜訊環境下語音辨識的精確度，可提供高達 58%的相對錯誤改善率，而跟原 NMF 法相較，新方法運算複雜度明顯降低，而能維持原辨識精確度、部分甚至有提升的效果。

Abstract

This paper proposes a novel scheme that enhance the modulation spectrum of speech features in noise speech recognition via non-negative matrix factorization (NMF). In the presented approach, we apply NMF to obtain a set of non-negative basis spectra vectors which derived from the clean speech to represent the important components for speech recognition. The difference compared to the conventional NMF-based scheme that leverages iterative search to update the full-band modulation spectra is two: first, we apply the orthogonal projection to update the low sub-band modulation spectra. Second, we process the low half-band of the

modulation spectrum rather than the full-band. The presented new process improves the computation efficiency without the cost of degarded recognition performance. In the Aurora-2 database and task, the presented new NMF-based approach can achieve the average error reduction rate of over 58% relative to the baseline MFCC.

關鍵詞：非負矩陣分解法、強健性、調變頻譜、語音辨識。

Keywords: nonnegative matrix factorization, modulation spectrum, speech recognition, noise robustness.

一、緒論

當一套語音辨識系統[1]應用在實際環境下時，環境的不匹配、語者變異性及發音的變異性通常會造成辨識效能的低落，為了降低這些變異性所造成的影響而發展的技術，一般而言統稱為強健性技術(robustness techniques)。

常見的語音特徵(speech features)強健性技術中，有一大類別是對於語音特徵的統計值做正規化(statistics normalization)，如倒頻譜平均值正規化法(cepstral mean normalization, CMN)[2]、倒頻譜平均值與變異數正規化法(cepstral mean and variance normalization, CMVN)[3]、倒頻譜增益正規化法(cepstral gain normalization, CGN)[4]、相對頻譜法(RelAtive SpecTra, RASTA)[5]、倒頻譜平均值與變異數正規化結合自回歸動態平均濾波器法(cepstral mean and variance normalization plus auto-regressive-moving average filtering, MVA)[6]、統計圖等化法(histogram equalization, HEQ)[7][8]、時間序列結構正規化法(temporal structure normalization, TSN)[9]等，這些正規化通常是作用於語音特徵的時間序列(temporal sequence)上。

本篇論文中，與上述正規化法主要不同的是，我們對於特徵時間序列的傅立葉轉換、即其調變頻譜(modulation spectrum)作強健性更新，其它常見的調變頻譜處理技術有頻譜統計圖等化法(spectral histogram equalization, SHE)[10]、強度比率等化法(magnitude ratio equalization, MRE)[10]、調變頻譜替代法(modulation spectrum replacement, MSR)與調變頻譜濾波法(modulation spectrum filtering, MSF)[11]等。作用於調變頻譜上其可能的好處是，我們可以直接針對不同的頻率成分加以處理。由N.Kanedera 的研究[12]發現，大部份的語音辨識資訊分布在 1 Hz 及 16 Hz 的中低調變頻率之間，因此如果若著重於處理此段調變頻率成分，而非整體頻帶，預期將可在不影響原始全頻帶方法之辨識精確度的前提下，有效減低計算上的複雜度、提升強健性技術的即時(real-time)性。

在分析多維資料的特性時，非負矩陣分解法(non-negative matrix factorization, NMF)[13-16]是近十幾年來相當新穎且有用的技術，起初，NMF 常運用在影像處理上，近年來，已有許多的相關 NMF 的應用及研究於語音辨識上。例如在文獻[17]中，利用了 NMF 對於語音特徵之全頻帶的調變頻譜作分解與更新，而達到了提升語音特徵強健性的效果。在本篇論文中，我們延伸了文獻[17][18]的觀念與方法，提出了兩種提升運算效能的新步驟：

1. 藉由正交投影(orthogonal projection)的方式取代原先 NMF 中求取新調變頻譜強度之迭代法(iteration approach)，如此可避免迭代法中費時的迭代運算及不確定的迭代數目，進而改進整體的運算速度。
2. 我們採取分頻帶分解的方式取代原先全頻帶分解，可只對於重要的中低頻帶作 NMF 的分解與更新，或分別對中低頻帶與高頻帶作 NMF 的分解與更新，藉此強調中低頻帶

的重要性、並可減少計算量。

除了上述採用 NMF 更新及分解調變頻譜之外，我們另行使用文獻[18]中的方法來進行討論，以主軸成分分析(principal component analysis, PCA)求取調變頻譜基底矩陣，由於主軸成分分析其主要目的在於針對一群資料找出其最佳投影方向，使得其投影後的資料點能夠獲得對大之變異量，因此，同樣以投影方法更新調變頻譜，並利用上述分頻帶分解進行更新的方法降低其運算複雜度。

在 Aurora-2 的數字資料庫的辨識實驗上，我們驗證了上述新方法的良好效果、可有效降低原始 NMF 處理法的複雜度，有時可附加提升辨識精確度的功能。

二、非負矩陣分解調變頻譜投影法

在本章節中我們將分成四個小節介紹基本的非負矩陣概念、及本篇論文所提出的方法之原理與步驟。首先第一小節介紹非負矩陣分解法，接著第二小節說明之前學者所提出的迭代式頻譜更新法，第三小節則說明本論文所提出的調變頻譜投影法，最後一小節則是藉由功率頻譜密度圖來討論其效能。

（一）非負矩陣分解法之介紹

非負矩陣分解法(nonnegative matrix factorization, NMF)主要目的是將一個內容皆為非負實數的矩陣，分解為元素皆為非負實數的兩個基底矩陣之乘積。假設欲分解的非負矩陣表示為 $\mathbf{V}=[\mathbf{v}_1,...\mathbf{v}_M]$，其中 $\mathbf{v}_j$ 為矩陣 $\mathbf{V}$ 之第 j 行，而矩陣 $\mathbf{V}$ 的尺寸為 N×M。藉由 NMF 分解 $\mathbf{V}$，得到 $\mathbf{W}$ 及 $\mathbf{H}$ 兩個非負矩陣，如下式表示：

$$\mathbf{V} \approx \mathbf{W}_{N\times r}\mathbf{H}_{r\times M} \tag{1}$$

$\mathbf{W}$ 及 $\mathbf{H}$ 尺寸分別為 N×r 與 r×M，r 可決定 $\mathbf{W}$ 及 $\mathbf{H}$ 兩矩陣尺寸(一般而言 r 遠小於 N 與 M)，其中 $\mathbf{W}$ 帶有 $\mathbf{v}$ 的行向量的綜合資訊，當我們改寫成 $\mathbf{v}\approx\mathbf{W}\mathbf{h}$ 時，則 $\mathbf{v}$、$\mathbf{h}$ 和 $\mathbf{V}$、$\mathbf{H}$ 將呈現行相關的關係，換句話說，$\mathbf{v}$ 可視為 $\mathbf{W}$ 之行向量的線性組合而 $\mathbf{h}$ 為其權重比，然而最主要的目的在於使 $\mathbf{W}$ 與 $\mathbf{H}$ 兩矩陣的乘積逼近於 $\mathbf{V}$，如此一來才能得到一組可靠的基底，這個過程我們以下式表示：

$$(\mathbf{W},\mathbf{H})=\min_{\tilde{\mathbf{W}},\tilde{\mathbf{H}}}\sum_{i,\mu}\left(\mathbf{V}_{i,\mu}-\left(\tilde{\mathbf{W}}\tilde{\mathbf{H}}\right)_{i,\mu}\right)^2 \tag{2}$$

（二）NMF 使用於全頻帶調變頻譜之迭代式更新

無論是此小節將要介紹的迭代式更新法或是下一小節的投影法，所採用的皆是與壓縮感知(compressed sensing)[16]相同的概念，簡單來說，所謂的壓縮感知並不直接對訊號直接做採集，而是經由將信號投影至一組波形上，得到一組壓縮數據後，再藉由最佳化的方式進行解碼，進而估計出原始訊號的重要訊息。

在文獻[18]中，首先提及利用NMF於語音倒頻譜特徵調變頻譜的更新上。在此，我們簡要地介紹其更新步驟：

步驟 I. 對於特定項之語音特徵(如第一維倒頻譜特徵)而言，將用以訓練聲學模型的每一句乾淨語音特徵時間序列作作離散傅立葉轉換(discrete Fourier transform, DFT)，得到其調變頻譜序列，將這些不同語句所對應之調變頻譜序列的強度(magnitude)排成一個矩陣 $\mathbf{V}$ 的每一行(column)，因此若 $\mathbf{v}$ 的尺寸為 N×M，代表了我們共有 M 句語音，而其頻率

點數為 N。

步驟 II. 利用前述之 NMF 法分解矩陣 $\mathbf{V}$，即求取等式(1)中的兩個矩陣 $\mathbf{W}$ 與 $\mathbf{H}$。其中矩陣 $\mathbf{W}$ 包含了 r 個尺寸為 N×1 的行向量(column vector)，這 r 個行向量包含了每一句乾淨語音調變頻譜強度之基底向量，而 $\mathbf{H}$ 則是代表了每一句乾淨語音的權重。

步驟 III. 對於訓練與測試的語句其特徵時間序列的調變頻譜強度（以向量 $\tilde{\mathbf{v}}$ 表示），我們利用前步驟所得的矩陣 $\mathbf{W}$，對此向量 $\tilde{\mathbf{v}}$ 以 NMF 的迭代法作逼近進而求得最後的 $\mathbf{h}^{(L)}$，而 $\mathbf{h}^{(L)}$ 即為 $\tilde{\mathbf{v}}$ 與 $\mathbf{W}$ 之間的權重比關係，整個過程如下：

初始： 任意指定一非負向量 $\mathbf{h}^{(0)}$。

迭代：前後兩向量 $\mathbf{h}^{(j)}$ 與 $\mathbf{h}^{(j+1)}$ 的關係為：

$$\mathbf{h}_k^{(j+1)} = \mathbf{h}_k^{(j)} \frac{(\mathbf{W}^T\tilde{\mathbf{v}})_k}{(\mathbf{W}^T\mathbf{W}\mathbf{h}^{(j)})_k} \tag{3}$$

其中對向量使用下標 k 代表此向量中的第 k 項。

終止：在多次迭代之後，藉由最後一次迭代所得的向量 $\mathbf{h}^{(L)}$（假設迭代總次數為 L），新調變頻譜強度為：

$$\tilde{\mathbf{v}}' = \mathbf{W}\mathbf{h}^{(\mathbf{L})}. \tag{4}$$

此新的調變頻譜強度配合原始的相位成分(phase part)，經過反傅立葉轉換就可得到新的特徵時間序列。

（三）NMF 使用於調變頻譜之投影式更新

在這裡，我們為前述的 NMF 調變頻譜更新技術，提出了兩個降低計算複雜度的修正步驟，分述如下：

I. 使用正交投影(orthogonal projection)替代原始的迭代法：

根據之前的描述，新調變頻譜強度由於 $\tilde{\mathbf{v}}'$ 必須以迭代方式輾轉求得，因此極可能影響演算法的運算複雜度。因此，我們希望能找出"單次"的運算法來求取新調變頻譜強度。如式(4)所示，新調變頻譜強度 $\tilde{\mathbf{v}}'$ 是由乾淨語音所得的基底矩陣 $\mathbf{W}$ 中每一個行向量作線性加成(linear combination)而得，換言之，$\tilde{\mathbf{v}}'$ 必落在基底矩陣 $\mathbf{W}$ 之行空間(column space)中。根據線性代數的知識，我們直接採用原始調變頻譜強度 $\tilde{\mathbf{v}}$ 投影於基底矩陣 $\mathbf{W}$ 之行空間之分量，作為新的調變頻譜強度，可表示為：

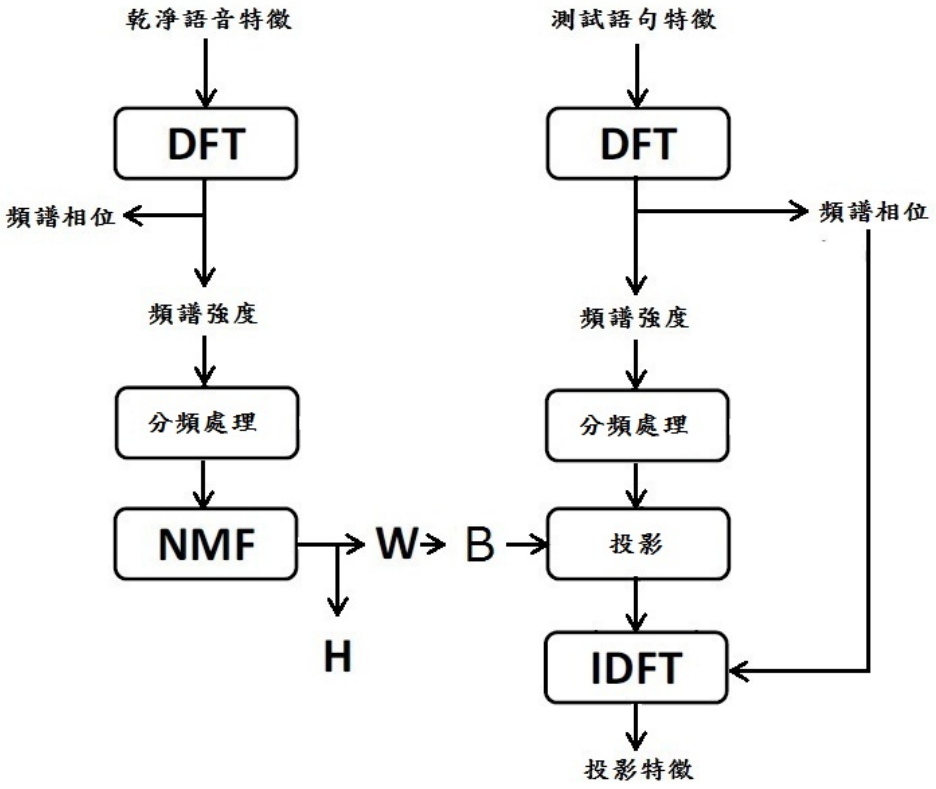
$$\tilde{\mathbf{v}}' = proj_{\mathbf{W}}(\mathbf{v}) = \mathbf{W}(\mathbf{W}^{\mathbf{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathbf{T}}\mathbf{v} \tag{5}$$

或

$$\tilde{\mathbf{v}}' = proj_{\mathbf{W}}(\mathbf{v}) = \mathbf{B}\mathbf{B}^{\mathbf{T}}\mathbf{v} \tag{6}$$

其中，矩陣 $\mathbf{B}$ 包含了基底矩陣 $\mathbf{W}$ 之行空間的正交基底(orthogonal basis)，接下來我們將採用(6)式來求得新的調變頻譜強度而不是(5)式的主要原因，是因為 $\mathbf{W}$ 有可能為稀疏矩陣(sparse matrix)且秩數小於 r，若此情況一旦發生，則 $(\mathbf{W}^{\mathbf{T}}\mathbf{W})^{-1}$ 項將呈現無解的情形，因此採用(6)式將可避免此問題，並且帶有額外的好處為矩陣 $\mathbf{B}$ 可在更新每一句語音調變頻譜前就事先由矩陣 $\mathbf{W}$ 求得，相較之下複雜度相對較低。採用投影法的最大優點在於無須隨著不同的語句而重新計算，所以並不會影響更新每一句特徵之運算複雜度。同時，在等式(6)的運算中，我們可以先計算向量 $\mathbf{B}^{\mathbf{T}}\mathbf{v}$，再將其左乘上矩陣，這樣的作法會遠比直接將事先算好的投影矩陣 $\mathbf{B}\mathbf{B}^{\mathbf{T}}$ 乘上向量 $\mathbf{v}$ 來的少。例如，矩陣 $\mathbf{B}$ 的尺寸(至多)

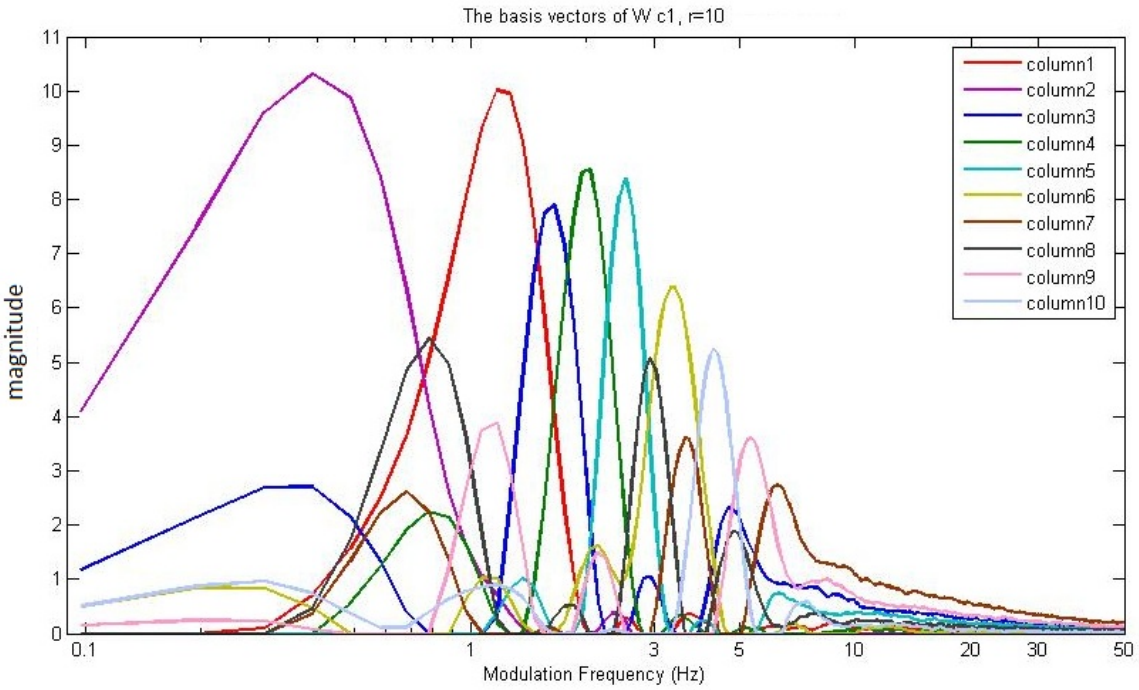
為 N×r，矩陣 $\mathbf{BB}^{\mathbf{T}}$ 的尺寸為 N×N，故若直接把 $\mathbf{BB}^{\mathbf{T}}$ 右乘尺寸為 N×1 的向量 $\mathbf{v}$，所需的乘法數目為 N^2；然而，先計算 $\mathbf{B}^{\mathbf{T}}\mathbf{v}$，再計算 $\mathbf{B}(\mathbf{B}^{\mathbf{T}}\mathbf{v})$ 所需的乘法數目為 Nr＋Nr=2Nr，而實際運用上，2Nr 通常低於 N^2，這是因為由於在 NMF 運算中，求取的基底矩陣 **W** 其行向量個數 r 通常遠低於其尺寸 N。（在我們的實驗設定中，N=513, r=10，則 2Nr=5120 < 65536= N^2），整體過程如下圖一所示。



圖一 投影式調變頻譜更新法流程圖

II. 使用子頻帶更新替代原始的全頻帶更新：

如同先前所提到的，不同頻帶的語音特徵時間序列調變頻譜對語音辨識的重要性並不一致，圖二為利用 NMF 所求得的基底矩陣 **W** 之調變頻譜強度分布圖，我們可觀察到其分佈區域大多集中於 10 Hz 之前，由此可知低頻帶的成分尤其重要，等量的雜訊干擾若存在於低頻帶，相較於存在於高頻帶，對於語音辨識的精確度影響更大，原始的 NMF 更新法則是對於全頻帶一併更新，而在這裡的新步驟裡，我們只針對前半段子頻帶的調變頻譜作更新，相較於原方法，不僅可以降低一半的複雜度，我們也預期這樣的處理並不會影響其後的辨識精確度，此點可在之後的實驗數據中證實。接下來我們將原始的 NMF 更新法以 $\text{NMF}^{(i,f)}$表示，其中的代號"i"與"f"分別代表了迭代(iteration)與全頻帶(full-band)兩個詞），及本論文提出的兩種改良式 NMF 法的排列組合，分別以 $\text{NMF}^{(p,f)}$ 、$\text{NMF}^{(i,low)}$ 與 $\text{NMF}^{(p,low)}$表示，其中的代號"p"與"low"分別代表了投影 projection 與低頻帶 low-band 兩個詞）。

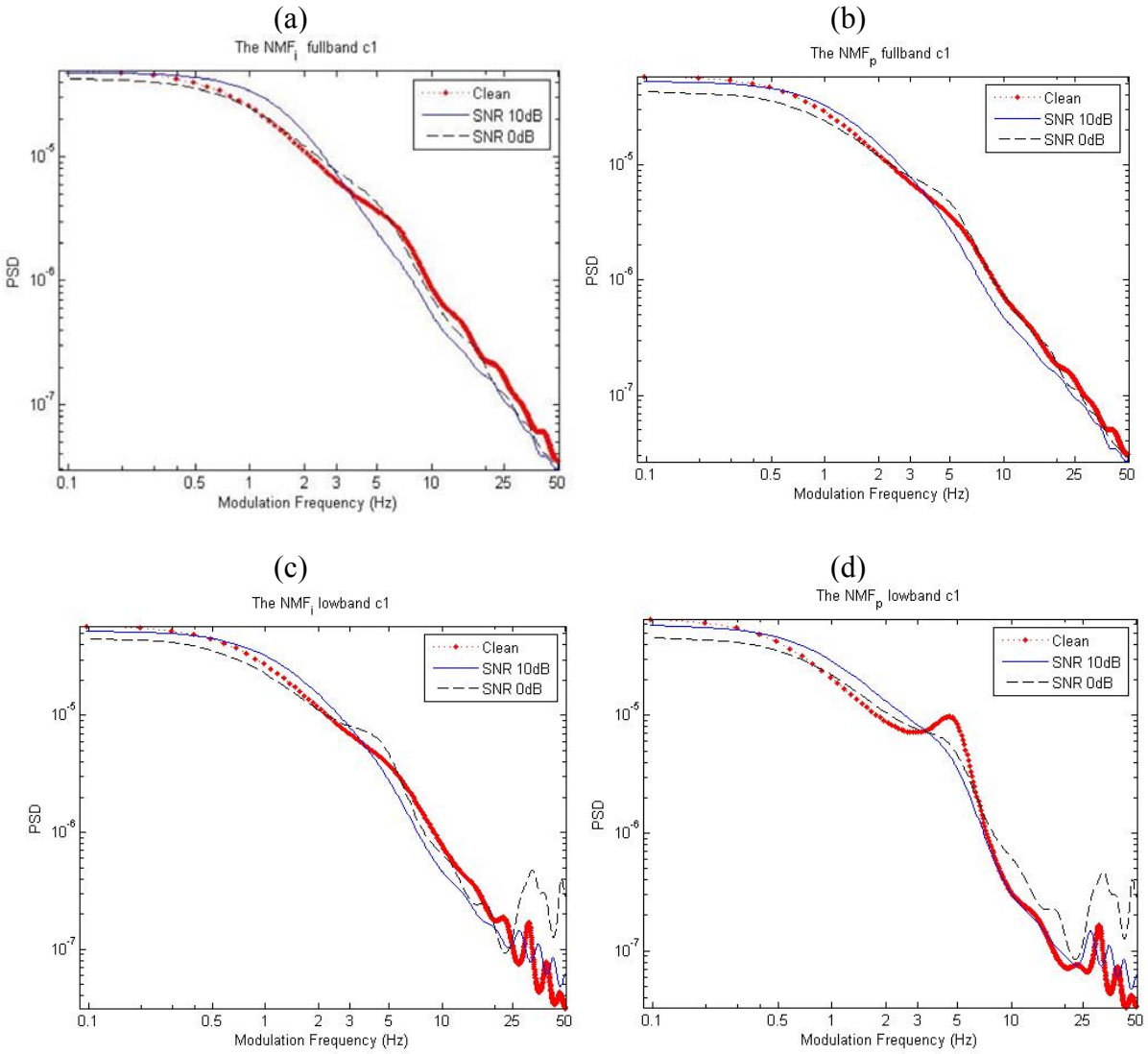


圖二 基底矩陣 **W** 之調變頻譜強度分布圖

（四）迭代法及投影法之初步效能討論

此小節主要在於比較迭代法與投影法的調變頻譜失真改善程度，藉由功率頻譜密度(power spectral density, PSD)圖來評估這兩個方法的效能。在此我們採用 AURORA 2.0[19]資料庫中的 MAH_27O6571A 語音檔，加上不同訊雜比(SNR)的地下鐵 (subway) 雜訊，使用的 NMF 法，其參數 r 設為 10。圖三的(a)(b)(c)(d)分別代表經過全頻帶迭代法、全頻帶投影法、低頻迭代法及低頻投影法處理後的第 1 維特徵序列的功率頻譜密度圖。

根據四個功率調變頻譜密度圖的結果，我們可發現四種結果都表現出相當好的失真改善性能，無論是經過迭代法或投影法處理過後的功率頻譜密度圖都明顯集中於調變頻率範圍[0, 25Hz]之間，因此，針對此段調變頻率範圍進行處理的效果可相當接近於全頻帶處理，在之後的辨識率實驗中也將證明此點。此外，在圖 3.1(c)(d)中可觀察到經過半頻處理後的兩種方法在 25Hz 前的失真改善程度皆相當良好。



圖三 各種方法作用於不同訊雜比下的 c1 特徵序列之功率頻譜密度圖：(a)全頻帶迭代法 (b)全頻帶投影法 (c)低頻帶迭代法 (d)低頻帶投影法

三、實驗環境設定

本論文之實驗中所採用的語音資料庫為歐洲電信標準協會(European Telecommunication Standard Institute, ETSI)所發行的 AURORA 2.0[19]語音資料庫，內容包含美國成年男女以人工方式錄製的一系列連續英文數字字串，在我們所採取的乾淨模式訓練、多元雜訊模式測試(clean-condition training, multi-condition testing)之實驗架構中，用以訓練聲學模型之語句為 8440 句乾淨語句，唯其包含了 G.712 通道之通道效應。測試語料則包含了三個子集合：Sets A 與 B 的語句摻雜了加成性雜訊，Set C 則同時包含加成性雜訊與摺積性雜訊，Sets A 與 B 各包含 28800 句語音，Sets C 包含 14400 句語音。加成性的雜訊種類分別為：地下鐵(subway)、人類嘈雜聲(babble)、汽車(car)、展覽館(exhibition)、餐廳(restaurant)、街道 (street)、機場 (airport)、火車站(train station)等雜訊，並以不同程度的訊雜比(signal-to-noise ratio, SNR)摻雜，分別為：clean、20 dB、15 dB、10 dB、5 dB、0 dB 與-5 dB；而通道效應分為 G.712 與 MIRS 兩種通道標準，由國家電信聯盟

(international telecommunication Union, ITU)[20]所訂定而成。

上述之用以訓練與測試的語句，我們先將其轉換成梅爾倒頻譜特徵參數(mel-frequency cepstral coefficients, MFCC)，作為之後各種強健性方法的基礎特徵(baseline feature)，建構 MFCC 特徵的過程主要是根據 AURORA 2.0[19]資料庫中的設定，最終 MFCC 特徵包含了 13 維的靜態特徵(static features)附加上其一階差分與二階差分的動態特徵(dynamic features)，共 39 維特徵。值得一提的是，本論文之後所提的強健性技術，皆是作用於 13 維的靜態特徵上，再由更新後的靜態特徵求取 26 維的動態特徵。同時，為了初步降低雜訊對於語音特徵的干擾，我們先以簡易但效果卓著的 MVN 法處理 MFCC 特徵，之後再配搭所新提出的各種基於 NMF 的演算法，藉此得到更佳的辨識精確度。

在聲學模型上，我們採取連續語音辨識中常見的隱藏式馬可夫模型(hidden Markov model, HMM)，並採用由左到右(left-to-right)形式的 HMM，意即下一個時間點所在的狀態只能停留在當下的狀態或下一個鄰近的狀態，狀態的變遷隨著時間由左至右依序前進。此外，模型中的狀態觀測機率函數為連續式高斯混合機率函數(Gaussian mixtures)，所以此模型又稱為連續密度隱藏式馬可夫模型(continuous-density hidden Markov model, CDHMM)。我們採用了 HTK[21]軟體來訓練上述的 HMM，在模型單位的選取上，採用前後文獨立(context independent)的模型樣式，所得之聲學模型包含了 11 個數字(oh, zero, one, …, nine)與靜音的隱藏式馬可夫模型，每個數字的 HMM 皆包含了 16 個狀態，而每個狀態由 3 個高斯混合函數組成。

四、實驗數據與討論

本節將由四部分所組成。

（一） 基於迭代方式之 NMF 頻譜強度更新法其辨識率實驗結果與討論

在本節中，我們列出兩種基於迭代方式之 NMF 頻譜強度更新法：$NMF^{(i,f)}$ (全頻帶更新)與 $NMF^{(i,low)}$ (半頻帶更新)所得的實驗結果並加以討論，我們變化式(1)所設定的矩陣行向量數 r，來觀察其帶來的影響。為了比較方便起見，我們先於表一與表二分別列出各種方法於「同一組別、不同訊雜比」與「同一訊雜比、不同組別」之平均辨識率，且先固定參數 r=10，接著變化參數 r=5, 10, 15，觀察 $NMF^{(i,f)}$ (全頻帶更新) 與 $NMF^{(i,low)}$ (半頻帶更新)兩法的總平均辨識率，呈現於表三中。

從表一、表二與表三之數據，我們可發現：

1. 只處理低頻帶之 $NMF^{(i,low)}$法在三組不同的雜訊環境下，對於 MVN 欲處理的語音特徵在辨識率上的提升，幾乎等同於原始處理全頻帶之 $NMF^{(i,f)}$法，在 Set A 與 Set B 中的辨識表現上，$NMF^{(i,low)}$法略低於 $NMF^{(i,f)}$法，但在 Set C 上，$NMF^{(i,low)}$法則略優於 $NMF^{(i,f)}$法，此結果呼應之前我們的推測，即由於語音鑑別資訊大部分集中於低頻帶區域，只對低頻帶處理就可獲得很好的效能，幾乎等同於全頻帶的處理。

2. 上述之 $NMF^{(i,low)}$法與 $NMF^{(i,f)}$法不僅在三組不同的雜訊環境下之平均效能相似，在不同的訊雜比(SNR)環境下，得到的辨識率也十分相近，這顯示辨識環境中無論雜訊程度的多寡，低頻帶處理的 $NMF^{(i,low)}$法都不會顯著低於原始全頻帶處理的 $NMF^{(i,f)}$法的強健性效能。

3. 當變化基底頻譜向量數 r 時，可發現就總平均辨識率而言，r 值的三種設定(5, 10, 15)所對應的兩種 NMF 法效果也是十分接近，較小的 r 值對應之平均辨識率甚至是較好的，這顯示了我們可以使用少量的基底頻譜，就可使這兩種 NMF 法發揮其近乎最佳的效果。

表一、使用 10 個基底頻譜向量(r=10)時， 原始 NMF(i,f)法與新提出之 NMF(i,low)法在不同組別之下、取 5 種訊雜比(20 dB, 15 dB, 10 dB, 5 dB 與 0 dB)之辨識率(%)平均比較

	Set A	Set B	Set C	Avg	RR
MVN	73.81	75.02	75.08	74.55	36.76
$NMF^{(i,f)}$	82.65	83.94	81.75	82.99	57.73
$NMF^{(i,low)}$	82.80	84.08	81.89	83.13	58.08

表二、使用 10 個基底頻譜向量(r=10)時， 原始 $NMF^{(i,f)}$法與新提出之 $NMF^{(i,low)}$法在不同組別之下、取 5 種訊雜比(20 dB, 15 dB, 10 dB, 5 dB 與 0 dB)之辨識率(%)平均比較

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
MVN	99.08	96.58	93.07	84.56	65.24	33.73	13.39
$NMF^{(i,f)}$	98.71	96.57	94.48	89.52	78.23	55.11	26.64
$NMF^{(i,low)}$	98.76	96.74	94.65	89.74	78.61	54.87	25.60

表三、使用三種基底頻譜向量數的設定(r=5, 10, 15)時， 總平均辨識率(%)之比較

	r = 5	r = 10	r = 15
$NMF^{(i,f)}$	83.36	82.99	82.87
$NMF^{(i,low)}$	83.19	83.13	83.09

（二）基於正交投影方式之 NMF 頻譜強度更新法其辨識率實驗結果與討論

在本節中，我們列出了所新提出之兩種基於**正交投影方式**之 NMF 頻譜強度更新法：$NMF^{(p,f)}$ (全頻帶更新) 與 $NMF^{(p,low)}$ (半頻帶更新)所得的實驗結果並加以討論，類似前一節，我們變化式(1)所設定的矩陣行向量數 r，來觀察其帶來的影響。且為了方便起見，我們先於表四與表五中分別列出各種方法於「同一組別、不同訊雜比」與「同一訊雜比、不同組別」之平均辨識率，且先固定參數 r=10，接著變化參數 r=5. 10, 15，觀察 $NMF^{(p,f)}$ (全頻帶更新) 與 $NMF^{(p,low)}$ (半頻帶更新)兩法的總平均辨識率，呈現於表六之中。從這幾個表中，我們得到了與上一節十分類似的結論，亦即：

1. 利用正交投影方式之全頻處理法 $NMF^{(p,f)}$與半頻處理法 $NMF^{(p,low)}$，無論於不同類別或是不同程度的雜訊環境，對於 MVN 預處理之 MFCC 語音特徵都有十分接近的辨識率改善效能。

2. 少量的基底頻譜向量(r=5)即可使全頻處理法 $NMF^{(p,f)}$與半頻處理法 $NMF^{(p,low)}$都達到很好的辨識精確度。

表四、使用 10 個基底頻譜向量(r=10)時， 原始 $NMF^{(p,f)}$法與新提出之 $NMF^{(p,low)}$法在不同組別之下、取 5 種訊雜比(20 dB, 15 dB, 10 dB, 5 dB 與 0 dB)之辨識率平均比較

	Set A	Set B	Set C	Avg	RR
MVN	73.81	75.02	75.08	74.55	36.76
$NMF^{(p,f)}$	82.66	83.82	81.63	82.92	57.56
$NMF^{(p,low)}$	82.65	83.97	81.90	83.03	57.84

表五、使用 10 個基底頻譜向量(r=10)時， 原始 $NMF^{(p,f)}$法與新提出之 $NMF^{(p,low)}$法在不同組別之下、取 5 種訊雜比(20 dB, 15 dB, 10 dB, 5 dB 與 0 dB)之辨識率平均比較

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
MVN	99.08	96.58	93.07	84.56	65.24	33.73	13.39
$NMF^{(p,f)}$	98.73	96.61	94.48	89.53	78.20	54.70	25.92
$NMF^{(p,low)}$	98.74	96.65	94.57	89.72	78.55	54.72	25.07

表六、使用三種基底頻譜向量數的設定(r=5, 10, 15)時， 總平均辨識率之比較

	r = 5	r = 10	r = 15
$NMF^{(p,f)}$	83.33	82.92	83.20
$NMF^{(p,low)}$	83.34	83.03	82.97

（三）各種強健性方法之效能比較

接下來我們將彙整前述不同的調變頻譜更新方法之實驗結果，除了我們先前提及的四種基於 NMF 的調變頻譜更新法：$NMF^{(i,f)}$、$NMF^{(i,low)}$、$NMF^{(p,f)}$與 $NMF^{(i,low)}$之外，在這裡我們也同時作了兩個特徵時間序列更新法：HEQ 與 MVA，及基於 PCA 的調變頻譜更新法，為了精簡比較起見，在各種 NMF 法與 PCA 法中，基底向量數目 r 固定為 10。

此外，我們也進一步去分析，在 NMF 法與 PCA 法中，僅更新整段頻帶之前 1/3（約前 170 個頻率點）與前 1/4（前 128 個頻率點）之頻譜強度對於辨識率之影響為何。這些實驗數據皆彙整於表七。藉由表七可以觀察到下列幾點：

1. MVN 法在處理 MFCC 特徵上，可使平均辨識率從 59.75%大幅提昇至 74.55%，而 HEQ 及 MVA 更可分別提供 22.46%及 19.00%的絕對錯誤率改善。
2. PCA 法無論處理全頻帶或低頻帶，都能有十分顯著的辨識率改進，且跟 NMF 的效能幾乎不相上下，甚至於更高，此原因極可能是雖然 PCA 並未加入「所求得之調變頻譜強度必不為負」的限制，但其實得到的頻譜強度仍然絕大部分都是大於或等於零，相位失真的可能性極小，所以效能優越。這將在之後的章節作說明。
3. HEQ 與 MVA 相對於 MVN 法能夠帶來更佳的辨識率，但本論文所討論之四種 NMF 頻譜更新法則明顯都優於 HEQ 與 MVA。

4. 之前所討論的兩種低頻帶更新法$NMF^{(i,low)}$與$NMF^{(f,low)}$皆是更新半頻帶的頻譜強度(相當於更新 256 個低頻率點)，如果我們將更新範圍縮減至全頻帶的 1/3 與 1/4（分別為三個表中所列之 170 點與 128 點)，其辨識率相對於全頻帶與半頻帶的更新法而言，確實會逐漸變低，但是變低的幅度並未很顯著，PCA 法也是如此，這代表了對於 NMF 與 PCA 法而言，我們可以更新比一半更少的頻率點，就足以趨近更新全部頻率點之效能。
5. 我們所提的三種新方法：$NMF^{(i,low)}$、$NMF^{(p,f)}$與$NMF^{(p,low)}$若與原始的$NMF^{(i,f)}$法相較，全頻式的$NMF^{(p,f)}$、半頻式的$NMF^{(i,low)}$與$NMF^{(p,low)}$皆可得到十分接近的平均辨識率。
6. 整體而言，在本論文中所提出的兩種改善步驟（正交投影與低頻處理)，確實能降低原始 NMF 用於更新語音調變頻譜之方法的運算複雜度，同時，辨識率並不會因此而降低。在使用正交投影法的改善步驟時，所得到的辨識率甚至可超越原步驟之結果，意味了此兩種新步驟可同時使 NMF 法的效率與效能同時提升。

表七、不同調變頻譜正規化法與各種特徵時間序列強健性技術之辨識率(%)比較表

方法	更新之低頻率點個數	基底向量個數	Set A	Set B	Set C	平均	AR	RR
MFCC	—	—	59.24	56.37	67.53	59.75	-	-
MVN	—	—	73.81	75.02	75.08	74.55	14.80	36.76
HEQ	—	—	79.96	81.74	80.45	80.77	21.02	52.22
MVA	—	—	78.15	79.17	79.12	78.75	19.00	47.21
迭代式 NMF	Full	r=10	82.65	83.94	81.75	82.99	23.24	57.73
	256	r=10	82.80	84.08	81.89	83.13	23.38	58.08
	170	r=10	82.55	83.93	81.88	82.97	23.22	57.68
	128	r=10	82.24	83.62	81.40	82.62	22.87	56.83
投影式 NMF	Full	r=10	82.66	83.82	81.63	82.92	23.17	57.56
	256	r=10	82.65	83.97	81.90	83.03	23.28	57.84
	170	r=10	82.66	83.90	81.84	82.99	23.24	57.74
	128	r=10	82.37	83.86	81.64	82.82	23.07	57.32
PCA	Full	r=10	83.10	84.13	82.22	83.34	23.59	58.60
	256	r=10	82.23	83.57	81.63	82.65	22.90	56.89
	170	r=10	82.41	83.66	81.71	82.77	23.02	57.20
	128	r=10	82.49	83.71	81.73	82.82	23.07	57.32

（四）迭代法及投影法之運算複雜度及負值頻譜強度之討論

在之前提到，我們所提出的兩種改良式 NMF 調變頻譜更新法，可有效降低原始 NMF 法的運算複雜度，在這裡，我們藉由實驗中各種參數的設定，來具體觀察這些基於 NMF 之調變頻譜更新法(NMF$^{(i,f)}$、NMF$^{(i,low)}$、NMF$^{(p,f)}$與 NMF$^{(p,low)}$)的運算量與所需時間、藉此比較它們的複雜度。

在參數的設定上，求取調變頻譜的 FFT 點數為 1024 點，由於共軛對稱特性的關係，因此式(1)中的調變頻譜點數 N=513，基底矩陣 W 的行向量數 r=10，迭代式 NMF 法(NMF$^{(i,f)}$與 NMF$^{(i,low)}$)之迭代數 L 設為 100，則四種演算法 NMF$^{(i,f)}$、NMF$^{(i,low)}$、NMF$^{(p,f)}$與 NMF$^{(p,low)}$對於單一調變頻譜向量之更新所需的的乘法運算數目及運作於 MATLAB 程式所需的時間詳列於表八，在此，執行時間改善率的定義如下：

$$\text{執行時間改善率} = \left(1 - \frac{\text{其他方法執行時間}}{\text{NMF(i,f)執行時間}}\right) \times 100\% \qquad (7)$$

表八 四種基於 NMF 之調變頻譜更新法的複雜度比較

方法	乘法總數（代數表示與參數帶入後的實際值）		在 MATLAB 執行所需時間(msec)	執行時間改善率(%)
NMF$^{(i,f)}$	$L(r^2+2r)+2Nr$	22260	5.52	—
NMF$^{(p,f)}$	$2Nr$	10260	3.30	40.22
NMF$^{(i,low)}$	$L(r^2+2r)+Nr$	17130	4.45	19.38
NMF$^{(p,low)}$	Nr	5130	1.98	64.13

由表八我們可看出，我們提出的三種新方法：NMF$^{(p,f)}$ 、NMF$^{(i,low)}$ 與 NMF$^{(p,low)}$ ，相對於原始的迭代式全頻帶處理之 NMF$^{(i,f)}$ 法可得到較低的運算複雜度（較少的乘法數目）及執行時間，其中，投影方式可減少 $L(r^2+2r)$次的乘法運算，半頻處理後則可減少 Nr 次的運算量。此外，實際程式運作的時間也都有相當大的改善，其中 NMF$^{(p,low)}$可減少約 64%的執行時間。

在我們之前的辨識實驗裡，主要是比較四種基於 NMF 的頻譜強度更新技術，但因我們所提出的正交投影法與線性代數理論中的 PCA 法密切相關，因此實驗裡我們也同時檢視基於 PCA 的頻譜強度更新技術的效果，簡單來說，PCA 亦是如同 NMF 一般，對於式(1)的資料矩陣 V 求取一組基底矩陣 W，並符合式(2)之最佳化準則。但跟 NMF 不同的點在於，PCA 並無限制限制矩陣 W 之行向量其中元素必為非負實數，經由 PCA 所求出之基底矩陣，其所包含的行向量恰為資料矩陣 V 所對應之共變異矩陣(convariance matrix)其 r 個固有向量(eigenvector)，前這些固有向量分別對應至共變異矩陣從大至小排序之前 r 個固有值(eigenvalue)。在無非負的限制下，PCA 比 NMF 能夠達到更小的平方誤差和（如式(2)所示），但潛在缺點是，PCA 求出之基底向量與之後使用正交投影方式求取出的調變頻譜強度可能出現負值，此違背了頻譜強度必為非負值的前提。

以下，我們將討論各種方法（包含 PCA 法）在更新頻譜強度時，所可能產生負值

的情形，其中，原始利用迭代方式的 NMF$^{(i,f)}$法並不會造成負值的產生，而利用正交投影方式的 NMF$^{(p,f)}$法在求得正交化矩陣 **B** 時可能產生負值，此外，如前所述，PCA 法也無法保證所求取出的頻譜強度必不小於零。在此，我們統計在 Aurora-2 資料庫之三個測試集裡，藉由不同方法其更新後的頻譜強度中所有負數總數，並且平均後得到每一句中每一維特徵之負數平均數量，列於表八。由表八中可發現到，經由迭代公式所求得之頻譜強度由於 NMF 法分解非負矩陣的本質，確實不會產生負值，但 NMF 投影法與 PCA 皆有少許的機會得到負值的頻譜強度，其中 PCA 得到負頻譜強度的比率略高於 NMF 投影法約 0.0006%，雖然頻譜強度為負值並不合理、會引入相位的失真（增加相位 π），但我們發現，由於上述兩種方法得到負值頻譜強度的機率相當低，因此可能對辨識性能的影響也就不明顯。此結果已在前面的實驗章節中呈現。

表八 異常之負值頻譜強度數量比較表

方法	NMF$^{(i,f)}$	NMF$^{(p,f)}$	PCA $^{(p,f)}$
正值平均數量	513	512.53	512.24
負值平均數量	0	0.47	0.7596
負值率(負值平均數量/總數量)	0%	0.0009162%	0.0015%

五、結論

本論文的重點在於改善原始基於迭代方式、非負矩陣分解（NMF）之全頻帶調變頻譜更新法之複雜度，同時保有其對語音特徵的雜訊強健化效能，所提出之方法與原始法最大不同在於，我們採取一次性的正交投影方式來求取在基底頻譜矩陣之展開空間(the spanned subspace)的新調變頻譜強度，而原始的迭代方式則是利用逐次逼近的方式來求得基底頻譜之權重。同時，我們根據語音調變頻譜其主要辨識資訊都集中在低頻區域的瞭解，提出了單獨更新低頻調變頻譜的模式，同是提升調變頻譜更新法的執行效率。就實際運行於 MATLAB 程式中、執行時間的改善程度來看，投影法可降低 40.22%的執行時間，半頻處理則可降低 19.38%以上的時間；特別的是，我們發現在 Aurora-2 資料庫的辨識實驗中，上述兩種改善運算複雜度之方式幾乎不會影響 NMF 調變頻譜更新法對應之辨識精確度，仍能提供原始 MVN 預處理後的 MFCC 特徵顯著的辨識率提升。

在未來的展望中，我們可將投影法與其他特徵時間序列域強健技術做結合，或是藉由 NMF 找出各種雜訊的基底，再藉由頻譜消去法(spectral subtraction, SS)或其他消噪法達到抑制雜訊或提升辨識率的效果。此外，也可進一步在其他資料庫上處理(如中文數字語音或是更多字彙的資料庫)，使其在現實層面中能有更多實際的應用。

參考文獻

[1] 王小川, “語音訊號處理,” 全華科技圖書, 2004.

[2] S. Furui, “Cepstral analysis technique for automatic speaker verification,” IEEE Trans.

On Acoustics, Speech and Signal Processing, pp.254-272, 1981.

[3] S. Tiberewala and H. Hermansky, “Multiband and adaptation approaches to robust speech recognition,” 1997 European Conference on Speech Communication and Technology (Eurospeech 1997).

[4] S. Yoshizawa, N. Hayasaka, N. Wada and Y. Miyanaga, “Cepstral gain normalization for noise robust speech recognition,” 2004 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2004), pp.1021-1024, 2004.

[5] H. Hermansky and N. Morgan, “RASTA Processing of speech features,” IEEE Trans-actions on Industrial Electronics, IEEE Trans. On Speech and Audio Processing, 1994.

[6] C. Chen and Bilmes, “MVA Processing of speech features,” IEEE Trans. on Audio, Speech and Language Processing, pp.257-270, 2006.

[7] 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.

[8] A. 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. Speech Audio Processing, vol. 13, no. 3, pp. 355–366, May 2005.

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

[10] L. C. Sun, C. W. Hsu and L. S. Lee, “Modulation spectrum equalization for robust speech recognition,” 2007 Automatic Speech Recognition and Understanding (ASRU2007).

[11] C. C. Lai, “Study of modulation spectrum normalization for robust speech recognition,” Master’s thesis, National Chi Nan University, Taiwan, July. 2011.

[12] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, “On the importance of various modulation frequencies for speech recognition,” Eurospeech 1997.

[13] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, 401:788–791, 1999.

[14] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” Ad-vances in Neural Information Processing Systems 13, 2000.

[15] B. Schuller, F. Weninger, M. Wollmer, Y. Sun, and G. Rigoll, “Non-negative matrix factorization as noise-robust feature extractor for speech recognition,” ICASSP 2010.

[16] David L. Donoho, “Compressed sensing,” IEEE Transactions on Information Theory, 52(4): 1289-1306, 2006.

[17] Wen-Yi Chu, Jeih-weih Hung, and Berlin Chen, “Modulation spectrum factorization for robust speech recognition,” APSIPA 2011.

[18] Jan-Yee Lee, Jeih-weih Hung, “Exploiting principal component analysis in modulation spectrum enhancement for robust speech recognition,” Fuzzy Systems and Knowledge Discovery (FSKD), vol 3, pp 1947 – 1951, 2011.

[19] H. G. Hirsch and D. Pearce, “The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions,” Proceedings of ISCA IIWR ASR2000, Paris, France, 2000

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

[21] The hidden Markov model toolkit (HTK): http://htk.eng.cam.ac.uk