The 2016 Conference on Computational Linguistics and Speech Processing
ROCLING 2016, pp. 181-193

非負矩陣分解法於語音調變頻譜強化之研究

A study of enhancing the modulation spectrum of speech signals via nonnegative matrix factorization

王緒翔 Xu-Xiang Wang[2,3]、鄭至皓 Zhi-Hao Zheng[1]、
曹昱 Yu Tsao[3]、洪志偉Jhih-Wei Hong [1]
[1] 國立暨南國際大學電機系
[2] 國立台灣大學通訊工程研究所
[3] 中研院資訊科技創新研究中心

摘要

在本論文中，我們使用了非負矩陣分解 (nonnegative matrix decomposition, NMF) 技術來強化語音特徵調變頻譜，並且分別訓練資料中乾淨語音及雜訊的調變頻譜強度基底(basis)，利用所得到的基底來分解測試語音之調變頻譜強度，最後搭配原始相角透過反傅立葉轉換(inverse Fourier transform)得到新的聲學頻譜並進而得到強化語音訊號。另外我們提出兩種變形以利降低演算複雜度:一種是將相鄰的聲學頻率點視為一體處理、另一種則是只處理低頻率區域的調變頻譜。最後，我們還與傳統的語音強化法做比較，如頻譜消去法和韋納濾波器法及最小期望平方誤差之短時頻譜強度估測法，驗證提出之方法的可行性。

在實驗資料庫的選擇上，我們引用 AURORA-2 連續數字語料庫之部分語句，其中的語音訊號受到加成性雜訊影響，實驗結果顯示上述之新方法對於基礎實驗而言，能有效提升雜訊環境下語音訊號的品質(PESQ)。

Abstract

In this paper, we propose to enhance the modulation spectrum of the spectrograms for speech signals via the technique of non-negative matrix factorization (NMF). In the training phase, the clean speech and noise in the training set are separately transformed to spectrograms and modulation spectra in turn, and then the magnitude modulation spectra are used to train the NMF-based basis matrices for clean speech and noise, respectively. In the test phase, the test signal is converted to its modulation spectrum, which is then enhanced via NMF with the basis matrices obtained in the training phase. The updated modulation spectrum is finally transformed back to the time domain as the enhanced signal. In addition, we propose two variants for the newly method in order to possess relatively high computation complexity One is to consider the several adjacent acoustic frequencies as a whole for the subsequent processing, and the other is to process the low modulation frequency components. These new methods are validated via a subset of the Aurora-2 noisy connected-digit database. Preliminary experiments have indicated that these methods can achieve better signal quality relative to the baseline results in terms of the Perceptual Evaluation of Speech Quality (PESQ) index, and they outperform some well-known speech enhancement methods including spectral subtraction (SS), Wiener filtering (WF) and minimum mean squared error short-time spectral amplitude estimation (MMSE-STSA).

關鍵詞：非負矩陣分解法、語音強化、時頻圖。

Keywords: non-negative matrix factorization, speech enhancement, modulation spectrum, spectrogram.

一、緒論

近數十年來科技一再的突破，電子產品的發展日新月異，大幅提升了生活的便利性也縮短了人與人之間的距離。言語，此一人類溝通最重要的媒介，重點在於使用雙方皆明白的語言，且儘量清晰、簡潔、明了地表達自己的觀點，交流彼此意見與想法。而在無法面對面以語音溝通的限制下，以往透過書信的方式，遠方傳來的文字訊息動輒需一週甚至一個月，或是需要高額的長途或國際電話費用，而拜當今網路科技所賜，現在幾乎可說是隨傳隨到，除了時間的縮減，傳送模式也有重大的改變，由傳統的文字傳送提升為語音傳送，讓使用者可以更快速簡捷且低成本地傳送自己的訊息，如微信、LINE 等著名的通訊應用程式 (APP)，而這些通訊 APP 的蓬勃發展與進步，更是讓消費者多了許多對話的選擇，由原先的文字訊息提升至語音通話、甚至進階到視訊通話，不管距離多遠，打開手機，對方彷彿就能出現在你面前與你交談，實可謂無遠弗屆。

然而在實際遠距離通訊中，語音傳遞的過程必然會受到通訊通道與收發音訊之周遭環境的干擾，對後者而言，意即當說話者在一個吵雜的環境中傳話，此語音會夾帶著許多雜訊導致接收者無法明確接收資訊。因此在本論文中，我們針對降低雜訊干擾而討論研發語音強化的技術，目的就是要降低雜訊對語音的干擾，只得人耳在聽覺方面能更加清晰接收語音。

特別一提的是，語音辨識可融入到汽車功能上，使汽車不再是一個單純交通運輸的機械產品。微處理器、電腦輔助等控制了汽車系統，包含引擎、傳動、ABS（Anti-lock Braking System） 剎車等原本基本的機械系統，而近年來更是增加了通訊娛樂導航系統及各式各樣的通訊設備，使汽車不再是以往單純的機械系統，但隨著汽車功能越來越多，相對的按鍵控制鈕也越來越複雜，駕駛人面對像是飛機駕駛艙的儀表按鈕，通常只會亂了手腳、不知如何輕鬆駕馭，使用這些五花八門的功能。為了簡化這些按鈕，許多車廠朝向使用語音的方式，目的就是要讓駕駛「眼睛不離開路面，手不離開方向盤」並且可和車子做直接的溝通，像是福特和微軟合作開發的 SYNC[1]，可以藉由語音的命令達到控制效果，如:FM 頻率選擇、調整冷氣溫度及開關、聲控撥號等，但是汽車語音辨識最大的難處之一，在於行駛中之車外雜訊干擾，如風切聲、引擎聲、輪胎的滾動雜訊等，這些聲音會破壞原始語音，進而降低車內語音辨識系統的辨識度，使整體智慧型系統做出錯誤的回應，反而違背當初設計的理念，由此可知語音強化在此的重要性。但在降低雜訊的過程中，過度的降噪又會造成原始語音失真，所以在降低雜訊的同時必須減少對原始語音的損壞，是語音強化技術需兼顧的條件。

在現實的環境中語音訊號容易受到外在環境的影響，降低了語音的可讀性 (intelligibility) 與品質 (quality)，依據雜訊特性是否隨著時間明顯變化可分為兩大類，若變化慢或沒有顯著的變化，稱作穩態雜訊 (stationary noise)，反之則稱作非穩態雜訊 (non-stationary noise)。而根據與語音和雜訊來源的組合關係，雜訊來源分為兩類：

（1）加成性雜訊 (additive noise)：亦稱作背景雜訊，雜訊在語音和時間上成線性相加 (linear addition) 的關係，此類雜訊可進一步分為非時變加成性雜訊與時變加成性雜訊，前者像是車聲、機器發出的聲音等，後者則包含市集雜訊 (babble) 與警報雜訊 (siren noise) 等。

（2）摺積性雜訊 (convolutional noise)：亦稱作通道雜訊，雜訊在語音和時間上可簡化成摺積 (convolution) 的關係，像是電話通道效應和麥克風通道效應，亦被稱之通道失真 (channel distortion)，假設麥克風是固定位置，則造成的通道效應近似為非時變，但若移動快速 （如在高速行駛的交通工具內） 的行動電話通訊，則通道效應明顯為時變。

語音強化目的是要降低雜訊對語音的干擾，把雜訊語音中的語音強調或還原，長期以來有許多學者提出此類的演算法，一般來說這些語音強化法可分為兩種: 監督式 (supervised) 和非監督式 (unsupervised)，簡單來說，兩者差異在於訓練資料本身之類別是否已知。

監督式的語音強化法通常同時使用語音和雜訊兩方的模型，且每個模型的參數由各自的訓練樣本所估測獲得，例如音素相關非負矩陣分解法 (phoneme-dependent NMF) [2]、基於碼簿(codebook)的強化法[3] 、基於貝氏非負矩陣分解結合隱藏馬可夫模型法 (BNMF-HMM)[4]、加入共稀疏性之摺積非負矩陣分解法 (convolutive nonnegative matrix factorization with cosparsity)[5]。非監督式的語音強化法並不要求事先求得語音特性及雜訊種類，目標是直接從雜訊語音中估測乾淨語音，此類著名的方法包括頻譜消去法 (spectral subtraction , SS)[6]、韋納濾波器法 (Wiener filter)[7]、卡爾曼濾波器 (Kalman filter)[8]、基於拉普拉斯最小均方誤差法語音強化法 (Laplacian-based MMSE estimator for speech enhancement)[9]、單通道週期訊號之強化法 (Enhancement of single channel periodic signals in the time-domain)[10]、壓縮語音強化 (compressive speech enhancement) [11]、基於貝氏非負矩陣分解之線上更新法 (Online BNMF)[4]、局部詞典混合 (mixtures of local dictionaries)[12]。監督式的方法因為事前資訊較豐富，若運用得當通常效果優於非監督式的方法，但仍需視方法本身的複雜度與假設是否吻合事實而定。

本論文主要是使用非負矩陣分解法 (nonnegative matrix factorization, NMF)[13] 來發展強化語音技術，NMF 法是由貝爾實驗室 (Bell Laboratory) 的 D.D. Lee 及麻省理工學院 (Massachusetts Institute of Technology, MIT) 的 H.S. Seung 所發展出來的演算法，起初用在影像處理，後續才漸漸被使用在語音強化上。基於 NMF 分析法，我們提出了更新語音調變頻譜 (modulation spectrum) 的強化技術。

對一段聲音而言，我們沿著時間軸把它切成音框 (frame) 的序列，當每個音框透過傅立葉轉換後，得到短時間聲學頻譜 (short-time acoustic spectrum)，接著再對每個聲學頻率點的時序列再取一次傅立葉轉換，即可得到各聲學頻率點之調變頻譜。而本論文所以出的方法是藉由 NMF 法對於上述之調變頻譜的強度做更新，藉由訓練資料中的乾淨語音及雜訊，分別訓練兩方之調變頻譜強度的 NMF 基底 (basis)，再利用兩方的基底來分解測試語音之調變頻譜強度，得到近似乾淨語音的成分後，配合原始相角、經由反傅立葉轉換 (inverse Fourier transform) 得到新的聲學頻譜時序列，從而可得強化語音訊號。另外，為了降低運算複雜度，我們另外提出了兩種變型：一種是將相鄰的聲學頻率點一併處理、另一種則是只處理低頻率區域的調變頻譜，根據實驗結果，這些新方法都可以有效提升雜訊語音的品質、亦即降低雜訊干擾的效應。

最後值得一提的是，我們將所提的新方法與三種著名語音強化法做比較，分別為頻譜消去法 (spectral subtraction, SS)[6]、韋納濾波器法 (Wiener filter)[7] 以及最小均方誤差短時間頻譜振幅估測法 (minimum mean-squared error short-time spectral amplitude estimation, MMSE-STSA) [14]，初步實驗結果可看出，我們所提的新方法在某些雜訊環境下能比這三種方法得到更佳的強化效果。

二、基於非負矩陣分解之調變頻譜強化法

（一）非負矩陣分解法之介紹

NMF 基本方法即是將一個非負矩陣分解成另外兩個非負矩陣，前矩陣近似為後兩個矩陣的乘積，所謂非負矩陣，是指此矩陣內的元素 (element) 都是大於或等於零的實數。若以數學描述，即為任一尺寸為$N \times M$的非負矩陣，透過 NMF 分析，可求得兩個非負矩陣$\mathbf{W}$和$\mathbf{H}$，如下式表示:

$$\mathrm{V} \cong \mathbf{WH} \tag{1}$$

其中：

- 矩陣$\mathbf{V}$通常稱為資料矩陣 (data matrix) 或樣本矩陣 (sample matrix)，每行資料樣本的維度為$N \times 1$，相當於$\mathbf{V}$包含了筆$N \times 1$的向量樣本
- $\mathbf{W}$通常稱為基底矩陣 (Basis matrix)，尺寸為$N \times r$ (一般來說$\mathbf{r}$遠小於N與M)，其中 r 為基底個數。由式 (1) 可看出，資料矩陣$\mathbf{V}$的每一個行向量可近似為基底矩陣$\mathbf{W}$的所有行向量之線性組合。
- $\mathbf{H}$通常稱為編碼矩陣 (Encoding matrix)，尺寸為$r \times M$，其每一行代表了上述之線性組合的係數，詳細來說，當選擇基底 （座標軸） 為矩陣$\mathbf{W}$的行向量時，樣本矩陣$\mathbf{V}$的任何第j個行向量 (即第j筆資料) 所對應的編碼（座標系數） 為矩陣$\mathbf{H}$的第j行向量。

根據前述，NMF 主要目的之一是要由資料矩陣$\mathbf{V}$由分析出資料的主要分布方向（座標軸，即所得之基底矩陣$\mathbf{W}$），進而得到座標值 （置於編碼矩陣$\mathbf{H}$中）。在這過程中，求得兩矩陣$\mathbf{W}$和$\mathbf{H}$的方式有很多種，一般而言，我們需定義一個成本函數 (cost function)，代表$\mathbf{WH}$與$\mathbf{V}$的差距或逼近程度，藉由最小化此成本函數的方式使$\mathbf{WH}$更有效的近似$\mathbf{V}$，下述所提到的是兩種比較常用到的成本函數:

- 歐幾里德距離平方 (Squared Euclidean distance)：

$$d_1 = \sum_{i,j}\left(\mathbf{V}_{ij} - (\mathbf{WH})_{ij}\right)^2 \tag{2}$$

- KL 散度 (Kullback-Leibler divergence)：

$$d_2 = \sum_{i,j}\left(\mathbf{V}_{ij}\log\frac{\mathbf{V}_{ij}}{(\mathbf{WH})_{ij}} - \mathbf{V}_{ij} + (\mathbf{WH})_{ij}\right) \tag{3}$$

接著使用乘法法則 (multiplication rule) 更新$\mathbf{W}$與$\mathbf{H}$，將 (2) 式或 (3) 式所表示的成本函數逐步縮小，若使用 (2) 式，則$\mathbf{W}$與$\mathbf{H}$的迭代更新式如下:

$$\mathbf{W}_{ik} \leftarrow \mathbf{W}_{ik}\frac{(\mathbf{VH}^T)_{ik}}{(\mathbf{WHH}^T)_{ik}} \tag{4}$$

$$\mathbf{H}_{kj} \leftarrow \mathbf{H}_{kj}\frac{(\mathbf{W}^T\mathbf{V})_{kj}}{(\mathbf{W}^T\mathbf{WH})_{kj}} \tag{5}$$

若使用 (3) 式則為:

$$\mathbf{W}_{ik} \leftarrow \mathbf{W}_{ik} \frac{\sum_j \mathbf{H}_{kj}\mathbf{V}_{ij}/(\mathbf{WH})_{ij}}{\sum_j \mathbf{H}_{kj}} \quad (6)$$

$$\mathbf{H}_{kj} \leftarrow \mathbf{H}_{kj} \frac{\sum_j \mathbf{W}_{ik}\mathbf{V}_{ij}/(\mathbf{WH})_{ij}}{\sum_i \mathbf{W}_{ik}} \quad (7)$$

特別一提的是，本文後續所使用的 NMF 法，其成本函數固定為 (2) 式之歐幾里德距離平方，因此**W**與**H**的迭代更新公式為 (4) 式和 (5) 式。

（二）基於非負矩陣分解之調變頻譜強化法 (NMF-MSE) 之介紹

在此我們將介紹本論文所提的新方法—基於非負矩陣分解之調變頻譜強化法 (NMF-based modulation spectrum enhancement, NMF-MSE)。一般而言，NMF 常用以強化語音之時頻圖 (spectrogram) [15]，亦即其聲學頻譜(acoustic spectrum) 的時序列 (time series)，而本章的新方法裡，簡單來說，即我們將各聲學頻率的頻譜強度時序列取其傅立葉轉換 (Fourier transform)、得到其調變頻譜 (modulation spectrum) [19,20,21]後，在對其強度成分作 NMF 的強化。

由於在此新方法中，所要強化更新的是語音聲學頻譜強度時序列之的調變頻譜其強度成分 (magnitude part)，在此我們先簡介如何求得調變頻譜。對一段語音而言，我們沿著時間軸把它切成一連串音框 (frame)，形成音框的時序列，再對每個音框分別取短時間傅立葉轉換 (short-time Fourier transform, STFT)，即可以得到每個音框的短時間聲學頻譜 (short-time acoustic spectrum)，式子如下:

$$X[n,k] = \sum_{u=0}^{L-1} x_n(\ell) e^{-j\frac{2\pi k\ell}{L}}, \quad (8)$$

$$0 \le n \le N-1, 0 \le k \le K-1,$$

其中，$\{x_n(\ell), 0 \le \ell \le L-1\}$ 代表了第n個音框的時間訊號，N與K分別為音框總數與聲學頻率點數。式(8)中的$X[n,k]$通常稱為語音的時頻圖 (spectrogram)。接著，我們對任一聲學頻率點k的聲學頻譜強度$|X[n,k]|$、沿著音框時間序列軸 （即n軸）再做一次傅立葉轉換 (Fourier transform)，即可得到各聲學頻率點之調變頻譜，如下式(9)。

$$\mathcal{X}[k,m] = \sum_{n=0}^{N-1} |X[n,k]| e^{-j\frac{2\pi nm}{M}} \quad (9)$$

$$0 \le m \le M-1, 0 \le k \le K-1,$$

其中M為調變頻率點數。在我們所新提出的 NMF-MSE 法中，即是針對式(9)之調變頻譜其強度（即$|\mathcal{X}[k,m]|$）加以強化，藉此降低雜訊在其中的效應。與一般基於 NMF 的強化法相似，NMF-MSE 法包含了訓練階段與測試階段，訓練階段用以得到訓練資料矩陣的基底 （具有代表性的方向），測試階段則利用上述的基底來凸顯測試資料中屬於特定方向上的分量，降低非相關的成分。詳述如下：

Step 1:訓練階段 (training phase)

將用以訓練的乾淨語音 (clean speech) 與雜訊 (noise)，通過短時間傅立葉轉換 (STFT) 後轉為頻域，得到了乾淨語音時頻圖與雜訊時頻圖，將不同語句之乾淨語音時

頻圖對應相同聲學頻率的調變頻譜強度（對應相同聲學頻率索引k之$|\mathcal{X}[k,m]|$，$0 \le m \le M/2$）排成一個矩陣，即這些頻譜強度藉由行向量 (column vector) 的形式排成乾淨語音資料矩陣 $\mathbf{V}_S$，同理，雜訊在切割成數段後，每段雜訊的時頻圖對應相同聲學頻率的調變頻譜強度也排成另一個矩陣$\mathbf{V}_N$，接著將$\mathbf{V}_S$與$\mathbf{V}_N$透過 NMF 分解得到基底矩陣 ($\mathbf{W}_S$與$\mathbf{W}_N$) 以及編碼矩陣 ($\mathbf{H}_S$與$\mathbf{H}_N$)，訓練程序步驟如下：

1. 決定基底矩陣 ($\mathbf{W}_S$與$\mathbf{W}_N$) 其行向量個數 r （行空間維度） 。
2. 將基底矩陣 ($\mathbf{W}_S$與$\mathbf{W}_N$) 與編碼矩陣 ($\mathbf{H}_S$與$\mathbf{H}_N$) 初始化，矩陣內所有數值皆不為負。
3. 對$\mathbf{W}_S$與$\mathbf{W}_N$的行向量做正規化，使每個行向量的元素總和為 1。
4. 對基底矩陣與編碼矩陣做迭代更新：使用先前所提到的 NMF 法，對語音矩陣$\mathbf{V}_S$與雜訊矩陣$\mathbf{V}_N$分別加以分解，並配合前兩步驟得到的初始值，透過 (4) 式與 (5) 式迭代得到收斂後的基底矩陣$\mathbf{W}_S$、$\mathbf{W}_N$及編碼矩陣$\mathbf{H}_S$、$\mathbf{H}_N$。

值得注意的是，上述步驟是針對每一個聲學頻率單獨求取，因此當時頻圖有K個聲學頻率點時，上述步驟需做K次，得到K組不同的基底矩陣 ($\mathbf{W}_S, \mathbf{W}_N$)。

STEP 2：測試階段 (testing phase)：

將測試雜訊語音 (noisy speech)，透過短時間傅立葉轉換 (STFT) 轉換成時頻圖，再求取單一聲學頻率對應的調變頻譜強度成分，以向量$\mathbf{v}$表示，接著將 STEP 1 得到兩基底矩陣$\mathbf{W}_S$與$\mathbf{W}_N$左右串聯而成並得到一個複合基底矩陣，即得到$\mathbf{W}_C^{Tn} = [\mathbf{W}_S \quad \mathbf{W}_N]$。然後，藉由 NMF 法分解向量$\mathbf{v}$，即$\mathbf{v} \approx \mathbf{W}\mathbf{h}$，但此時只迭代更新編碼向量$\mathbf{h}$，基底矩陣$\mathbf{W}$維持不變、固定為上述之複合矩陣$\mathbf{W}_C^{Tn}$，其用意在於能利用 STEP 1 之$\mathbf{W}_S$與$\mathbf{W}_N$所分別包含的乾淨語音和雜訊資訊。當迭代完成後我們可得：

$$\mathbf{v} \approx \mathbf{W}\mathbf{h} = [\mathbf{W}_S \quad \mathbf{W}_N]\begin{bmatrix}\mathbf{h}_S \\ \mathbf{h}_N\end{bmatrix} = \mathbf{W}_S\mathbf{h}_S + \mathbf{W}_N\mathbf{h}_N, \tag{10}$$

最後，取得近似乾淨語音頻譜之強度，如$\mathbf{W}_S\mathbf{h}_S$或

$$(\mathbf{W}_S\mathbf{h}_S / (\mathbf{W}_S\mathbf{h}_S + \mathbf{W}_N\mathbf{h}_S)) \times \mathbf{v}, \tag{11}$$

（其中，"/"與"×" 分別為矩陣單一元素點的除法與乘法運算），此強度配合$\mathbf{v}$的原始相位成分，可得到更新過後的調變頻譜，再經由反傅立葉轉換 (Inverse Fourier transform) 即可獲得強化的 （單聲學頻率） 聲學頻譜強度時序列，最後，強化後之所有聲學頻率之聲學頻譜強度配合原始相位，得到新的時頻圖後，各音框經由反短時間傅立葉轉換 (Inverse STFT) 就可得到強化後的語音訊號。

（三）NMF-MSE法的兩種變形

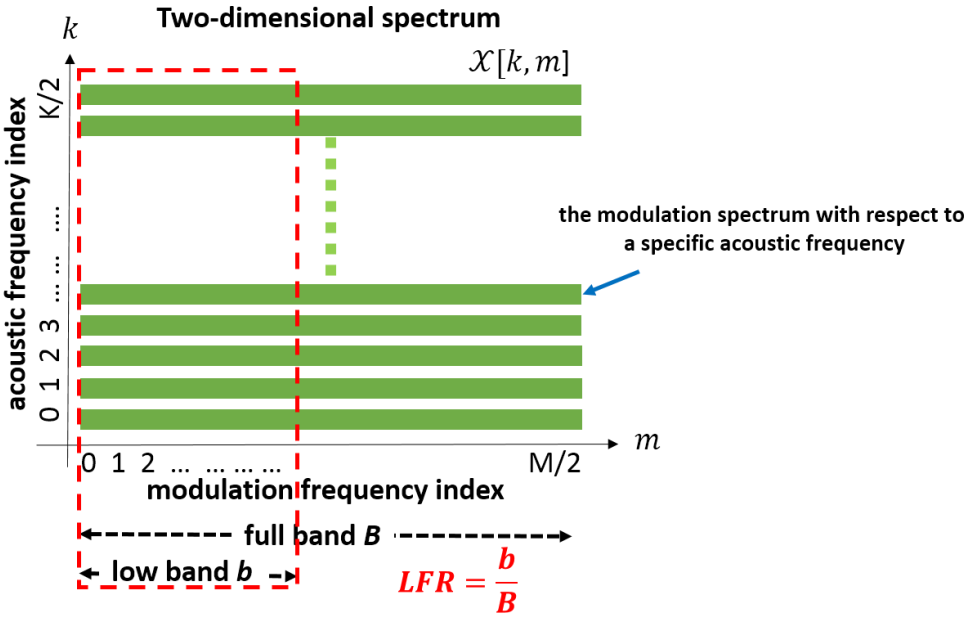
- 低調變頻帶之NMF-MSE

在諸多文獻中，皆提到語音主要的資訊是集中在低頻率的調變頻譜中，例如當音框取樣率為 100 Hz 時，雖然整體調變頻譜的頻率範圍為 0 到 50 Hz，對語音辨識之關鍵頻

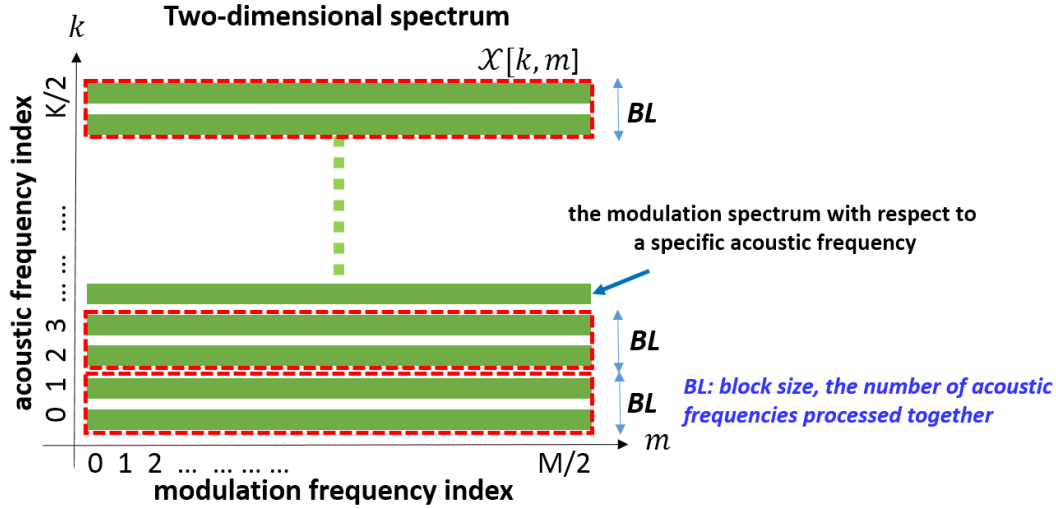
率範圍為 0 至 16 Hz，且以 4 Hz 的成分為最重要。我們使用這樣的觀念，來簡化 4.1 節中所提的 NMF-MSE 法，亦即我們只對於調變頻譜的低頻帶成分加以強化，而將剩餘的高頻帶成分保留不動。圖二為示意圖，我們在此定義了一個參數，稱做低頻帶相對全頻帶的頻寬比例 (Low-to-full ratio)，簡寫為LFR，例如當$LFR = 0.25$，則代表只更新全調變頻帶前 25%的低頻帶成分。使用低於 1 的LFR帶來的好處當然是可以降低 NMF-MSE 法的演算複雜度，因為需強化的調變頻譜強度向量的尺寸變小了。但LFR當然不能無限制地變小，否則處理過窄的低調變頻帶將無助於降低雜訊的成分。

- 降低頻率解析度之NMF-MSE

在前一節所述的 NMF-MSE 法中，我們是針對個別聲學頻率點的調變頻譜加以強化，但由於相鄰聲學頻率的特性通常相似，如果我們將相鄰聲學頻率視為一體、其調變頻譜強度共同用以 NMF 的基底求取的訓練上，此時因為訓練資料量的增加，基底矩陣本身應該更具代表性、有助於測試階段的調變頻譜強化上，圖三為示意圖，我們在此定義一個參數BL (block)，代表了一併處理之相鄰聲學頻率的點數。在原始的 MSE-NMF 模式中，$BL = 1$，我們將在之後的實驗裡，將此值變化為 2, 3 與 4。很明顯的，當BL越大，代表越多相鄰聲學頻率被視為一體來加以處理，意即頻率解析度變低，因此，雖然增加BL可使 NMF 訓練資料變多，但可能因為犧牲頻率解析度而降低整體強化的效果。



圖一：參數LFR之示意圖



圖二：參數BL之示意圖

三、實驗環境設定

（一）所使用的語音資料：

本論文用以評估強化方法之實驗所使用的語音資料，是取自歐洲電信標準協會(European-Telecommunications Standards Institute, ETSI) 所發行的 AURORA 2.0 語音資料庫[16]，發音語者為美國成年男女，內容為一系列連續的英文數字字句，此語料庫起初是用在雜訊環境之語音辨識評估上，訓練聲學模型的環境有兩種:乾淨訓練 (clean-condition training) 環境：使用的訓練語料是不包含雜訊的乾淨語音 (clean speech);另一種則是複合訓練 (multi-condition training) 環境：訓練語料是不同程度之訊雜比 (Signal-to-noise ratio, SNR) 的雜訊語音。由於我們在本論文中，重視的是語音強化技術的發展與評估，並未訓練聲學模型用以語音辨識，因此與上述兩種訓練環境無關。

在評估我們所提出的基於 NMF 之調變聲學頻譜強化法中，我們使用了來自 Aurora-2 資料庫中標示 FAK 之女性語者的乾淨語音字句，其中 39 句用以訓練乾淨語音 NMF 基底矩陣$\mathbf{W}_S$，而另外的 10 句添加上警報器雜訊 (siren noise) 形成雜訊語音，作為待強化的測試語句，且其訊雜比 (signal-to-noise ratio, SNR) 共有 20 dB、15 dB、10 dB、5 dB、0 dB 五種。純警報器雜訊則用以訓練乾淨語音 NMF 基底矩陣$\mathbf{W}_N$，我們將兩基底矩陣 （$\mathbf{W}_S$與$\mathbf{W}_N$） 的行數固定為 20。

（二）所使用的語音品質估測方法

目前已有的語音品質估測方式分為兩類，主觀 (subjective) 類別及客觀(objective) 類別。主觀類別的語音品質評估其中著名的方法為平均意見得分 (mean opinion score, MOS)法[17]。客觀類別的語音品質評估法，是透過電腦演算法分析語音訊號或頻譜、用量化數據的方式來表示語音訊號受到雜訊損壞的程度高低，著名的方法如：Perceptual Evaluation of Speech Quality (PESQ)法[18]。在本論文中，將採用客觀類別的 PESQ 法來評估語音品質，PESQ 所得的分數範圍為 1.0~4.5 之間，主要對應到主觀評分的 MOS 法，越高分代表語音品質越佳。一般來說分數超過 4 分代表著聽者覺得滿意或非常滿意。

四、實驗數據與討論

本節將由四部分所組成。

（一） 原始 NMF-MSE 之實驗結果

在這裡，我們驗證所提出之原始 NMF-MSE 法對於受到雜訊干擾之語音的強化效果。特別說明的是，此時原始 NMF-MSE 法採用了單一聲學頻率（即BL參數設為 1）並執行於全調變頻帶（即LFR參數設為 1）。表一列出了受雜訊干擾之語音其強化後的 PESQ 值。

從表一明顯看出，雜訊對於語音品質皆有明顯的影響，原始 NMF-MSE 法在雜訊干擾的五種訊雜比 (SNR) 程度之環境下，都能明顯提升 PESQ 值，初步驗證了此新方法確實在抑制雜訊效應上有所幫助。例如與基礎實驗相比，NMF-MSE 法使 PESQ 值平均提升了 10%，其中又以 SNR 為 0dB 時最為明顯，提升了 21%。

表一：基礎實驗及原始 NMF-MSE 法所得的 PESQ 值

雜訊為 siren，固定$BL = 1$與$LFR = 1$之 NMF-MSE 法之 PESQ 結果					
SNR	**0 dB**	**5 dB**	**10 dB**	**15 dB**	**20 dB**
Baseline (未處理)	1.8266	2.1700	2.5367	2.8015	3.1490
NMF-MSE	2.2215	2.3892	2.7011	2.9321	3.2008

（二）不同比例之低調變頻帶之 NMF-MSE 之實驗結果

在本節，我們將呈現不同比例之低調變頻帶之 NMF-MSE 更新後之 PESQ 值，如前所述，參數 LFR 代表了被 NMF-MSE 更新之低頻帶頻寬相對於全頻帶頻寬之比例 (low-to-full ratio)。在本節實驗中此參數LFR分別被設定為 1、0.75、0.50 與 0.25，隨著LFR值的遞減，所更新處理的頻寬也跟著變小。同時，NMF-MSE 法中的BL參數值固定為 1，代表每個單一聲學頻率之調變頻譜被獨立強化。

表二列出了受雜訊干擾之語音經由 NMF-MSE 強化後的 PESQ 值。從這表，我們有以下的觀察：

1. 當處理的調變頻帶寬越小（即LFR越小）時，NMF-MSE 所對應的 PSEQ 提升程度一般會越低，唯一例外是 15 dB 的訊雜比時，$LFR = 0.75$的設定比$LFR = 1$的設定得到更高的 PESQ 值。
2. 雖然降低LFR相對帶來較低的語音強化效果，但其實 PESQ 變差的效應並不顯著，特別是當訊雜比較高的情形下，例如當把LFR從 1 降為 0.5 時，SNR 為 10 dB 的狀態下，PSEQ 大約只下降了 0.01 至 0.02，但運算複雜度卻可以因此下降了一半（因為只處理 50%的調變頻譜）。
3. 在高 SNR 時，處理低至$1/4$的調變頻譜寬度也能達到近似處理全頻帶的效果，但當 SNR 很低（如 0 dB 或 5 dB 時），處理全頻帶的 NMF-MSE 仍是較適當的選擇。

表二：警報雜訊環境下，基礎實驗及不同LFR值之 NMF-MSE 法所得的 PESQ 值

雜訊為 siren，固定$BL = 1$、變化LFR值之 NMF-MSE 法之 PESQ 結果					
SNR	**0 dB**	**5 dB**	**10 dB**	**15 dB**	**20 dB**

Baseline		1.8266	2.1700	2.5367	2.8015	3.1490
NMF-MSE	$LFR = 1$	**2.2215**	**2.3892**	**2.7011**	2.9321	**3.2008**
	$LFR = 0.75$	2.2176	2.3815	2.6931	**2.9387**	3.1991
	$LFR = 0.5$	2.1632	2.3476	2.6819	2.9267	3.1967
	$LFR = 0.25$	2.0703	2.2486	2.6280	2.8880	3.1885

（三）不同聲學頻率點數之全調變頻帶之 NMF-MSE 之實驗結果

在本節，我們將呈現不同聲學頻率點數之全調變頻帶之 NMF-MSE 更新後之 PESQ 值，如前章所述，參數 BL 代表了同時被 NMF-MSE 更新之相鄰聲學頻率的點數。在本節實驗中此參數BL分別被設定為 1、2、3 與 4，隨著BL值的增加，代表了 NMF-MSE 所處理之聲學頻率解析度相對變小。此時，NMF-MSE 法中的LFR參數值固定為 1，代表處理的對象是全頻帶之調變頻譜。

表三列出了受雜訊干擾之語音經由 NMF-MSE 強化後的 PESQ 值。從這表，我們看到在較低聲學頻率解析度的設定（即BL大於 1）時，NMF-MSE 強化的效果通常比原設定（原頻率解析度，即BL等於 1）來的好，此大致呼應了我們前面所提，相鄰聲學頻率之調變頻譜的特性相近，因此一併處理是可行的，且因為增加樣本數，使 NMF 法能得到較精確的基底矩陣，進而使 NMF-MSE 的效果更明顯。此外，採用大於 1 的BL值（即較低的頻率解析度）可使 NMF-MSE 運算複雜度降低、卻仍可以帶來相似甚至更好的強化效果，此為一顯著的優點。

表三：警報雜訊環境下，基礎實驗及不同BL值之 NMF-MSE 法所得的 PESQ 值

雜訊為 siren，固定$LFR = 1$、變化BL值之 NMF-MSE 法之 PESQ 結果						
SNR		**0 dB**	**5 dB**	**10 dB**	**15 dB**	**20 dB**
Baseline		1.8266	2.1700	2.5367	2.8015	3.1490
NMF-MSE	$BL = 1$	2.2215	2.3892	2.7011	2.9321	3.2008
	$BL = 2$	2.2445	**2.3981**	2.7020	**2.9372**	3.1991
	$BL = 3$	**2.2453**	2.3969	**2.7062**	2.9356	**3.2015**
	$BL = 4$	2.2322	2.3882	2.6960	2.9319	3.1870

（四）NMF-MSE 法與三種語音強化法之效能比較

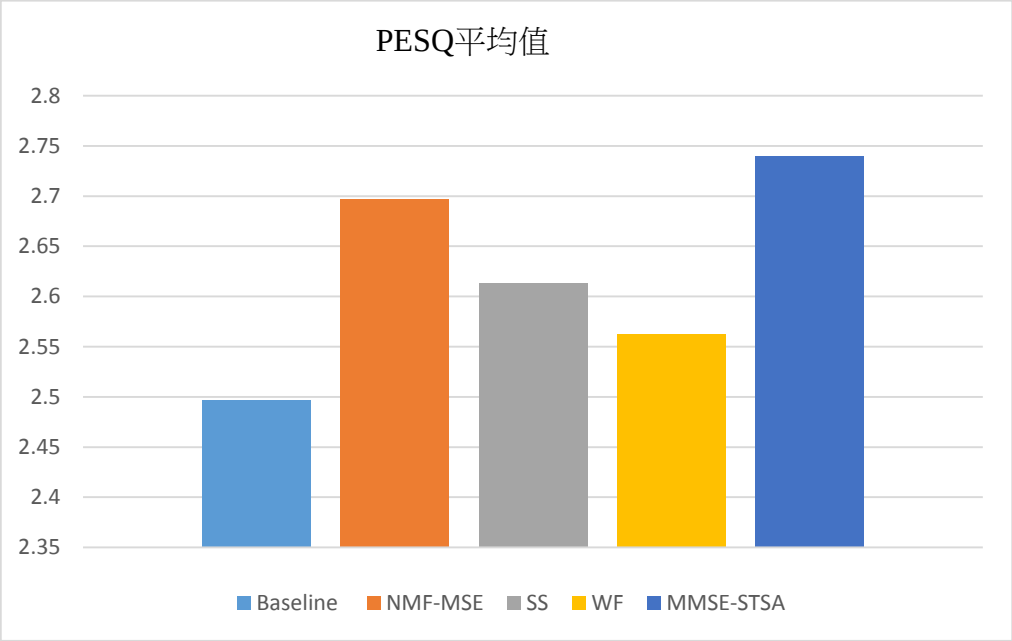
在本節，我們將呈現前一節最佳設定的 NMF-MSE 法與三種著名語音強化法做比較，分別為頻譜消去法(SS)、韋納濾波器法(WF)以及最小均方誤差之短時頻譜振幅估測法(MMSE-STSA)，實驗結果（PESQ 值）如表四。同時，我們也將表四各方法跨不同 SNR 之 PESQ 平均值轉成圖一之長條圖以方便比較。

根據表四及圖一，我們可以觀察到三個著名的語音強化法及本論文提出的 NMF-MSE 法在五種 SNR 狀態下均可明顯提升語音品質，平均而言，MMSE-STSA 的效能表現最佳，其次就是本論文所提出的 NMF-MSE，WF 次之，SS 居末。特別的是，在 SNR 為0 dB 時，NMF-MSE 法得到的 PESQ 值是所有方法裡面最高的，相對於基礎實驗 (baseline)

相比，提升了 0.4187。而當雜訊程度很低（SNR 為 20dB 時）， WF 法所對應的 PESQ 為最佳，而 NMF-MSE 和基礎實驗相比仍高了 0.0525。值得注意的是此時 NMF-MSE 的 BL參數設定為 3，相當於聲學頻率解析度降為原始的$1/3$，在此較低的運算複雜度條件下，仍可得到顯著的語音強化效果。

表四：警報雜訊環境下，基礎實驗、NMF-MSE （其BL設為 3、LFR設為 1）與三種語音強化法（SS, WF, MMSE-STSA）所得的 PESQ 值

SNR	**0 dB**	**5 dB**	**10 dB**	**15 dB**	**20 dB**
Baseline	1.8266	2.1700	2.5367	2.8015	3.1490
NMF-MSE	**2.2453**	2.3969	2.7062	2.9356	3.2015
SS	1.9320	2.2362	2.7176	2.9483	3.2350
WF	1.6569	2.2042	2.6856	2.9632	**3.3036**
MMSE-STSA	2.0098	**2.4355**	**2.7472**	**3.0262**	3.2873



圖一、 基礎實驗、NMF-MSE（其BL設為 3、LFR設為 1）與三種語音強化法（SS, WF, MMSE-STSA）各方法所得之跨不同 SNR 的 PESQ 平均值

五、結論

在本論文中，我們提出了基於非負矩陣分解法 (NMF) 在語音強化上的應用技術，使用 NMF 法對語音時頻圖之調變頻譜的強度做更新，簡稱為 NMF-MSE (NMF-based modulation spectrum enhancement)。在 NMF-MSE 法中，藉由分開訓練語句中乾淨語音及雜訊的調變頻譜強度的 NMF 基底 (basis)，將所得基底用以分解測試語音時頻圖的調變頻譜強度，再將強化後的調變頻譜透過反傅立葉轉換 (inverse Fourier transform) 得到

新的時頻圖，進一步得到強化後的語音訊號。同時，我們對上述新方法進一步提出了兩種方式來降低演算複雜度，分別為不同比例之低調變頻帶之 NMF-MSE 更新及不同聲學頻率點數之全調變頻帶之 NMF-MSE 更新，從 PESQ 為指標的語音品質評估上，上述方法皆能有效提升雜訊語音的清晰度。

特別一提的是，NMF-MSE 法是針對語音時頻圖的調變頻譜加以更新，就我們所知，目前基於 NMF 的語音強化法大多是針對時頻圖本身作強化，甚少有進一步處理調變頻譜，因此我們的新方法相當於拓展了 NMF 法在語音強化上的應用。在未來展望中，我們希望可以將 NMF-MSE 法結合其他種語音強化法達到抑制雜訊或提升語音品質，如本論文中用以比較的的頻譜消去法 (spectral subtraction, SS)、韋納濾波法 (Wiener filtering, WF) 與平均最小化誤差短時頻譜振幅估測法 (minimum mean-square error short-time spectral amplitude estimation, MMSE-STSA)。此外，我們將測試擴展 NMF-MSE 法的運用，包括用在多語者與多類型雜訊環境的語音強化上，另外也可以更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) ，以探討其實際層面應用的價值。

參考文獻

[1] Ford Sync - Official Site http://www.ford.com/technology/sync/

[2] B. Raj, R. Singh, T. Virtanen, “Phoneme-dependent NMF for speech enhancement in monaural mixtures” in Proceedings of Interspeech, 2011.

[3] S. Srinivasan, J. Samuelsson, and W. Kleijn, “Codebook driven short-term predictor parameter estimation for speech enhancement,” IEEE Transactions on Audio, Speech, and Language Processing, 14(1), pp. 163–176, 2006.

[4] N. Mohammadiha, "Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization," IEEE Trans on Audio, Speech, and Language Processing, vol 21, Oct. 2013

[5] M. Mirbagheri, Y. Xu, S. Akram and, S. Shamma, “Speech enhancement using convolutive nonnegative matrix factorization with cosparsity regularization” in Proceedings of Interspeech, 2013

[6] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Transactions on Acoustics, Speech, and Signal Processing, 27(2), pp. 113–120, 1979.

[7] N. Wiener, "The Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications," New York: Wiley, 1949.

[8] V. Grancharov and J. S. B. Kleijn, “On causal algorithms for speech enhancement,” IEEE Transactions on Audio, Speech, and Language Processing, 14(3), pp. 764-773, 2006.

[9] B. Chen and P. C. Loizou, “A Laplacian-based MMSE estimator for speech enhancement,” Speech Communication, vol. 49, no. 2, pp. 134–143,Feb. 2007.

[10] J. Jensen, J. Benesty, M. Christensen, and S. Jensen, "Enhancement of single-channel

periodic signals in the time-domain," IEEE Trans on Audio, Speech, and Language Processing, vol. 20, no. 7, pp. 1948–1963, 2012.

[11] S. Y. Low, D. S. Pham, and S. Venkatesh, “Compressive speech enhancement,” Speech Communication, vol. 55, no. 6, pp. 757–768, July 2013.

[12] M. Kim, P. Smaragdis, “Mixtures of local dictionaries for unsupervised speech enhancement, ”IEEE Signal processing letters, vol. 22, no. 3, pp. 293-297, 2015

[13] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in Proceedings of INIP, pp.556–562, 2000

[14] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” IEEE Transactions on Audio, Speech, and Language Processing, 32(6), pp. 1109–1121, 1984.

[15] H-T. Fang, J-H. Hung, X. Lu, S-S. Wang, Y. Tsao, “Speech enhancement using segmental nonnegative matrix factorization,” in Proceedings of the IEEE ICASSP, pp. 4483 – 4487, 2014.

[16] H. G. Hirsch and D. Pearce, “The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions,” in Proceedings of the ASR, 2000

[17] Mean Opinion Score (MOS) — A Measure of Voice Quality http://voip.about.com/od/voipbasics/a/MOS.htm

[18] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ) — a new method for speech quality assessment of telephone networks and codecs,” in Proceedings of ICASSP, pp. 749-752, 2001.

[19] Shihab A. Shamma and Les Atlas, "Joint acoustic and modulation frequency", EURASIP Journal on Applied Signal Processing, Volume 2003, Jan 2003

[20] L. Atlas, Q. Li, and J. Thompson, "Homomorphic modulation spectra", in Proceedings of the IEEE ICASSP, 2004.

[21] C.-C. Hsu, T.-H. Lin and T.-S. Chi, " FFT-based spectro-temporal analysis and synthesis of sounds", in Proceedings of the IEEE ICASSP, 2011.