應用特徵偏移向量的線性調整作環境調適之語音辨認

Linear adaptation of feature deviation vector for environment adaptive speech recognition[1]

李立民　林耕弘

Lee-Min Lee　　Geng-hong Lin

摘要

在安靜環境中所發展出來的語音辨識系統移植到雜訊環境時，其辨認率將會急遽下降，故雜訊效應的補償方法將為語音辨認之重要課題。在本文中，我們對語音之逆頻譜特徵係數在外加雜訊下的行為進行探討。我們發現逆頻譜向量的值會隨著外加雜訊的增加而漸縮小，當環境為白色雜訊時，其變動的方向會與乾淨語音的逆頻譜向量的方向相反，且會縮小到原點，但是在非白色雜訊下，變動的方向則往純雜訊與乾淨語音間的特徵向量差值方向行進，基於這個行為，我們提出特徵偏移向量，將其引入參考模型中，我們利用少量的混雜語音與參考模型中每個狀態之逆頻譜均值向量間的差值，求得特徵偏移向量，但考慮到求出特徵偏移向量所使用的混雜語音，可能與實際使用時的環境並不匹配，所以我們再利用一最佳純量係數，乘上特徵偏移向量作為最佳線性調整量，加至原始的語音參考模型，使模型能夠更適應的在雜訊環境下使用，實驗結果顯示這個方法能夠有效的使用在白色雜訊與其它非白色雜訊環境下。

關鍵詞：語音辨識、環境適應、外加雜訊、特徵偏移向量

ABSTRACT

When a speech recognition system in quiet environment is moved to a noisy environment, the recognition rate drops drastically. The compensation of noise effect becomes an important task for noisy speech recognition. In this study, we investigate the behavior of speech cepstral vector due to additive noise. We find that the cepstral vector deviates as the level of additive noise increases. In the case of white noise, the direction of cepstral vector deviation is approximately opposite to the direction of the cepstral vector of the clean speech. As power level of the white noise increases, the cepstral vector of the noisy speech will converge to the zero vector. However, for other types of noise, the change of cepstral vector is approximately at the direction of the difference vector of the noise cepstral vector and clean speech cepstral vector. Base on this behavior, we include a feature deviation vector into the reference model to compensate for the noise effect. The deviation vector is calculated according to the difference value of the cepstral vector of a few noisy speech and the corresponding model state cepstral mean vector. During the pattern matching phase, an optimally scaled deviation vector is added to the state mean vector of the clean speech model so that the clean speech model is adapted to the noisy environment. Experimental results show that the proposed method is effective for white noise and color noises.

Keywords：speech recognition, environment adaptation, additive noise, feature deviation vector

[1] 本研究係國科會編號 NSC90-2213-E-212-022 支持之研究計畫，本文作者感謝中華電信研究所提供語料庫。

李立民　林耕弘　大葉大學電機工程研究所

1. 前言

語音對人類來說是最自然且方便的溝通方式，而語音辨識系統可將此方式應用到人機介面，語音辨識技術在安靜的環境中已可達到不錯的辨識效果，不過，仍需解決的問題還是很多，比如說，使用在有雜訊的實際環境中，其辨認率即會迅速下降，由於不可能要求使用者在絕對安靜的環境下使用系統，因此，雜訊效應的補償問題就顯得格外重要。在過去的研究中，在逆頻譜特徵的補償方面，Mansour 和 Juang 提出一系列強健式的距離量測方法[8]。在模型的補償方法方面，Guan 等人提出的適應性的線性預估係數[2]，Moreno 等人提出利用近似的環境函數使模型更強健[9]， Gales 和 Young 提出平行模型補償法[3]。

而在我們的研究與文獻資料中，發現逆頻譜的值會隨外加白色雜訊之訊雜比的增加而減小，亦是因為這個原因，使乾淨語料所訓練出來的語音參考模型，不適合在雜訊環境下，故讓辨認率遽然下降，為了克服辨識系統效能降低的問題，Lee 提出對逆頻譜向量之線性調整的方法[1]，其根據逆頻譜在雜訊環境下會縮小的行為，使用一個最佳縮小因子，將參考模型的逆頻譜均值向量隨測試的混雜語料，進行動態的線性調整，使其辨識系統能更強健的適用於雜訊環境。我們發現逆頻譜線性調整法，的確可以有效的改善系統在外加白色雜訊下辨認率不佳的結果，但在非白色雜訊下的改善效果，就沒有較佳的辨認率，尤其在訊雜比相對較低的雜訊環境下，逆頻譜線性調整法的改善效果就更明顯的不足，探討其原因我們可以發現，因為在非白色雜訊下的語音逆頻譜係數，所表現出來的行為，明顯的與白色雜訊下有所不同，當外在環境為非白色雜訊時，其逆頻譜係數不但會有縮小的行為，變動的方向亦有明顯的改變。故線性調整法對參考模型的補償就會略顯不足，所以我們對於逆頻譜的這些特性，提出了特徵偏移向量補償的方法，我們利用少量的混雜語料與參考模型中每個狀態之逆頻譜均值向量間的差值，得到每個狀態的特徵偏移向量，但必須考慮到實際使用的環境，可能與訓練特徵偏移向量時所使用的混雜語料有不匹配的情形，讓補償的效果大打折扣，所以我們在利用一個最佳純量係數，以作偏移向量的線性調整，使我們的參考模型能夠更逼近於實際系統操作的環境。實驗的結果亦顯示我們所提出來的方法比線性調整法有更好的辨認率，尤其是在非白色雜訊的環境下。

2. 語音逆頻譜向量在雜訊環境下之行為

在時域上，我們可以將混雜語音表示為：

$$y[n]=x[n]+w[n] \qquad (1)$$

其中 $x[n]$ 代表乾淨語音信號，$w[n]$ 為外加雜訊，$y[n]$ 代表含雜訊語音，n 為時間上的指數。首先，我們假設 $x[n]$ 與 $w[n]$ 為互不相關的信號，且外加雜訊的平均值為零，則 $y[n]$ 的自相關係數可以寫成：

$$r_{yy,m}=r_{xx,m}+r_{ww,m}\ ,\ \mathbf{0}\le m\le P \qquad (2)$$

其中，P 代表線性預估分析的階數，$r_{xx,m}$ 與 $r_{ww,m}$ 分別表示乾淨語音信號與外加雜訊的第 m 階自相關係數。我們將混雜語音利用自迴歸處理，並求出其線性預估係數（Linear predictive coefficients：LPC）為

$$\mathbf{a_y}=(\mathbf{R_x}+\mathbf{R_w})^{-1}(\mathbf{r_x}+\mathbf{r_w}) \qquad (3)$$

其中，$\mathbf{a_y}=\left[a_{y,1},a_{y,2},\cdots,a_{y,P}\right]^T$ 代表混雜語音的線性預估向量，$\mathbf{R_x}=\begin{bmatrix} r_{xx,0} & r_{xx,1} & \cdots & r_{xx,P-1} \\ r_{xx,1} & r_{xx,0} & \cdots & r_{xx,P-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx,P-1} & r_{xx,P-2} & \cdots & r_{xx,0} \end{bmatrix}$ 代表乾淨語音信號的自相關係數所形成之矩陣，相同的，$\mathbf{R_w}$ 則是代表雜訊信號的自相關係數所形成之矩陣

，$\mathbf{r_x}=\left[r_{xx,1},r_{xx,2},\cdots,r_{xx,P}\right]^T$ 則為乾淨語音信號的自

相關係數所形成之向量，相同的，$\mathbf{r}_w$ 則為雜訊信號的自相關係數所形成之自相關向量，我們將式（3）利用正規化自相關係數表示為

$$\mathbf{a}_y = (\bar{\mathbf{R}}_x + \rho \cdot \bar{\mathbf{R}}_w)^{-1}(\bar{\mathbf{r}}_x + \rho \cdot \bar{\mathbf{r}}_w) \quad （4）$$

上式（4）中 $\bar{r}_{ww,m} = r_{ww,m} / r_{ww,0}$

$\bar{r}_{xx,m} = r_{xx,m} / r_{xx,0}$ ，$1 \le m \le P$，分別代表雜訊及語音之正規化自相關係數，$\rho = r_{ww,0} / r_{xx,0}$ 為雜訊比。

由方程式(4)可以發現混雜語音的線性預估係數為 ρ 的函數，故我們將 $\mathbf{a}_y$ 表示為 $\mathbf{a}(\rho)$，接著利用線性預估係數，可以求出逆頻譜係數為：

$$c_m(\rho) = a_m(\rho) + \sum_{k=1}^{m-1} \frac{k}{m} c_k(\rho) \cdot a_{m-k}(\rho)$$

，其中 $1 \le m \le P$ （5）

由上式可以發現，逆頻譜係數與雜訊成一非線性關係，且語音受雜訊干擾後，其逆頻譜的大小值與方向均會有所改變，如圖 1 所示。

觀察語音受到雜訊干擾後，其逆頻譜係數所表現出來的行為有兩點特質：（1）逆頻譜向量的大小

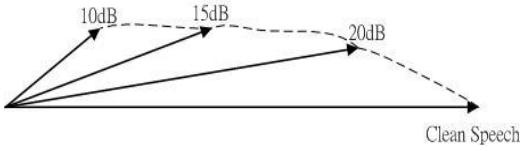


圖 1 乾淨語音的逆頻譜向量與其在不同雜訊強度下的版本

會隨雜訊增強而縮小（2）逆頻譜向量方向的改變不明顯[1,5]。

由圖 2 觀察逆頻譜係數在外加白色雜訊不同訊雜比下的改變，我們可從圖 2（a）乾淨語音至圖 2（d）訊雜比 10dB 之間的變化發現，逆頻譜係數的大小，隨著外加雜訊變大，慢慢的縮小，但方向的改變並不明顯

圖 3 顯示不同雜訊環境下逆頻譜係數的變動軌跡，從圖 3（a）中可以發現在外加白色雜訊時，逆頻譜係數會隨雜訊增加變動到原點，故使用前人所提到的線性調整法，即可對參考模型有所補償，但我們亦發現在外加雜訊為非白色雜訊時，逆頻譜係數並不會隨雜訊增加而變動到原點，如圖 3（b）、圖 3（c）所示，故線性調整法可能就無法有效的對參考模型有所補償，在這裡我們提出的偏移向量法就可有效的修正這個差值。

線性調整法與偏移向量調整法之比較如圖 4 所示，其中△表示經線性調整法調整後的位置，○表示經特徵偏移向量法調整後的位置，$\Delta\mathbf{c}$ 表示偏移向量，α 表示純量係數。

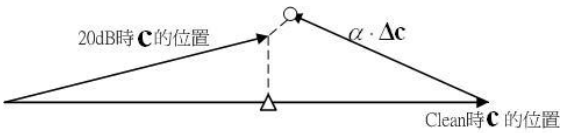
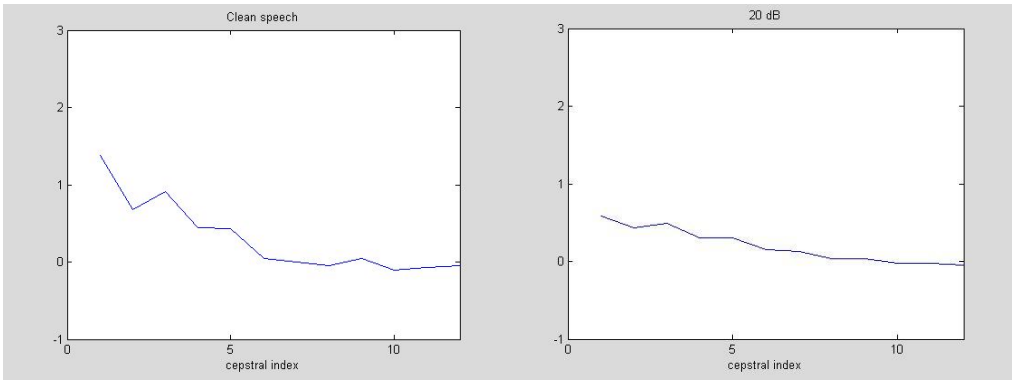
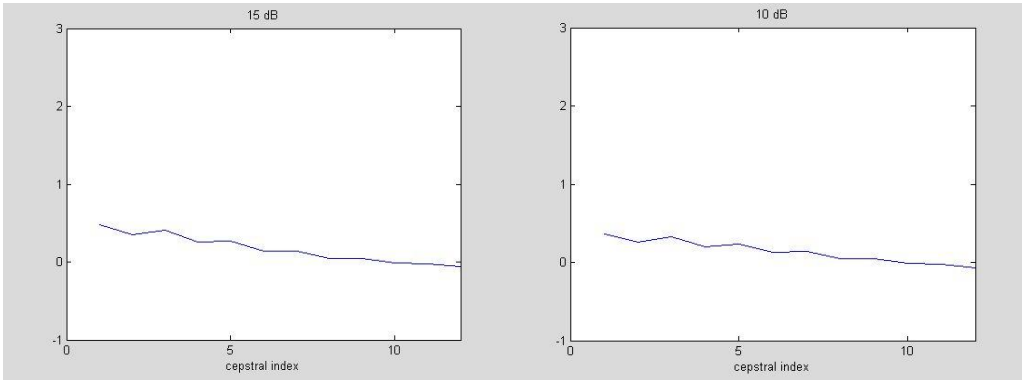


圖 4 線性調整法與偏移向量法調整示意圖



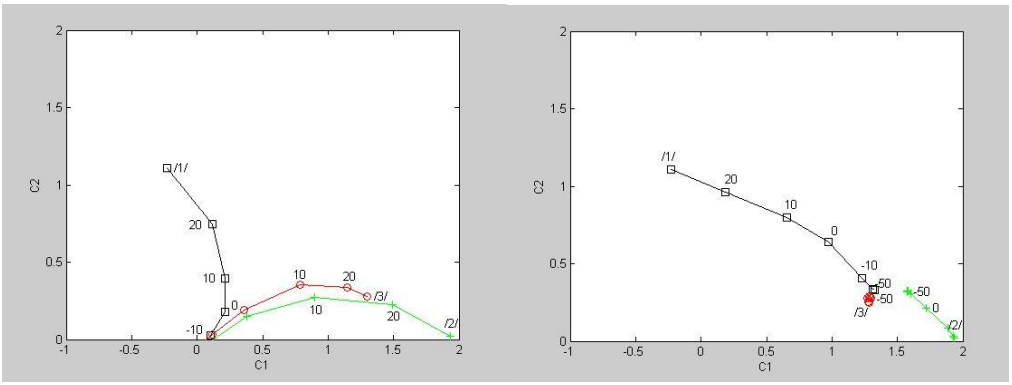
（a）　（b）



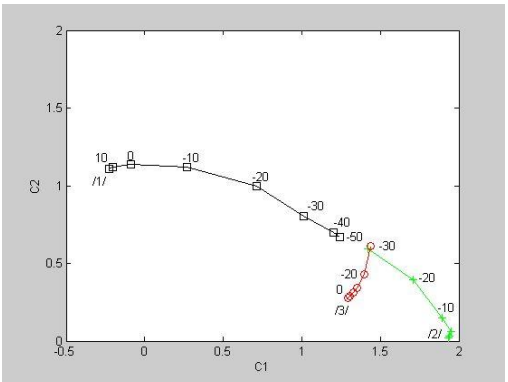
（c）　（d）

（a）乾淨語音　（b）SNR = 20dB（c）SNR = 15dB　（d）SNR = 10dB

圖 2　逆頻譜係數在不同訊雜比下的變化



（a）　（b）



（c）

圖 3　不同雜訊環境下逆頻譜係數的變動軌跡

3. 逆頻譜向量在雜訊下之表示

將混雜語音的逆頻譜向量利用一階泰勒級數表示為：

$$\mathbf{c}(\rho) \cong \tilde{\mathbf{c}}(\rho) = \mathbf{c}(0) + \left.\frac{d\mathbf{c}(\rho)}{d\rho}\right|_{\rho=0} (\rho - 0) \tag{6}$$

其中 $\tilde{\mathbf{c}}(\rho)$ 代表混雜語音逆頻譜向量 $\mathbf{c}(\rho)$ 的一階線性近似值。

將逆頻譜向量的變化，簡化為線性的關係[6]，給定一雜訊比 δ ，則 $\left.\frac{d\mathbf{c}(\rho)}{d\rho}\right|_{\rho=0}$ 可近似為

$$\left.\frac{d\mathbf{c}(\rho)}{d\rho}\right|_{\rho=\mathbf{0}} \cong \frac{\mathbf{c}(\delta) - \mathbf{c}(\mathbf{0})}{\delta} \tag{7}$$

因此，（6）式可化為

$$\tilde{\mathbf{c}} = \mathbf{c}(0) + \frac{\rho}{\delta}\{\mathbf{c}(\delta) - \mathbf{c}(0)\} \tag{8}$$

令 $\alpha = \rho/\delta$ ，則式（8）可為

$$\begin{aligned}\tilde{\mathbf{c}} &= \mathbf{c}(0) + \alpha \cdot \mathbf{c}(\delta) - \alpha \cdot \mathbf{c}(0) \\ &= (1-\alpha)\mathbf{c}(0) + \alpha \cdot \mathbf{c}(\delta)\end{aligned} \tag{9}$$

我們可以定義偏移向量 $\Delta\mathbf{c}(\delta)$ 為

$$\Delta\mathbf{c}(\delta) = \mathbf{c}(\delta) - \mathbf{c}(0) \tag{10}$$

則雜訊環境下的逆頻譜向量亦可以表示為

$$\tilde{\mathbf{c}} = \mathbf{c}(0) + \alpha \cdot \Delta\mathbf{c}(\delta) \tag{11}$$

4. 雜訊效應下參考模型的補償方法

4.1 逆頻譜線性調整法

我們可以利用逆頻譜係數在雜訊環境下的特性，將均值向量的逆頻譜部份乘上一個純量係數以補償雜訊對語音的影響，使其能夠適應當時的雜訊環境[1]。假定某音框落在狀態 i ，則對觀測向量 $\mathbf{c}_t$ 所得的相似分數為

$$\begin{aligned}L(\mathbf{c}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = -\tfrac{1}{2}\{&D\ln(2\pi) + \ln(|\boldsymbol{\Sigma}_i|) \\ &+ (\mathbf{c}_t - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{c}_t - \boldsymbol{\mu}_i)\}\end{aligned} \tag{12}$$

其中 D 表示維度， $\boldsymbol{\mu}_i$ 為均值向量， $\boldsymbol{\Sigma}_i$ 為變異數矩陣，我們對隱藏式馬可夫模型（Hidden Markov Models：HMM）中[7]的均值向量的逆頻譜部份乘上一純量係數，以補償逆頻譜縮小的特性，如下式：

$$\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\} \rightarrow \{\lambda\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\} \tag{13}$$

可產生最大相似分數的純量係數，定義為最佳縮小係數 λ 。將 $L(\mathbf{c}_t; \lambda\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ 對 λ 取微分並令微分式為零，得最佳縮小因子之計算式如下：

$$\lambda = \frac{\langle \mathbf{c}_t, \boldsymbol{\mu}_i \rangle}{\langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_i \rangle} = \frac{\sum_{k=1}^{P} c_{t,k} . u_{i,k} / \sigma_{i,k}^2}{\sum_{k=1}^{P} u_{i,k} . u_{i,k} / \sigma_{i,k}^2} \tag{14}$$

4.2 特徵偏移向量之線性調整法

由式（11）觀察可發現混雜語音逆頻譜向量的變化，我們利用少量的混雜語料計算出偏移向量，並依照實際使用時的環境對 $\Delta\mathbf{c}(\delta)$ 作線性調整，使參考值能更接近於測試值[4]。

$$\alpha_{opt} = \arg\min_{\alpha} \{Dist(\mathbf{c}_T, \tilde{\mathbf{c}}_R)\} \tag{15}$$

$\mathbf{c}_T$ 為測試語音的逆頻譜向量，$\tilde{\mathbf{c}}_R$ 為在當時雜訊環境下，最適宜的逆頻譜參考值。由測試與參考的逆頻譜，可得距離如下式：

$$Dist(\mathbf{c}_T, \tilde{\mathbf{c}}_R) = \left|\mathbf{c}_T - (\mathbf{c}_R(0) + \alpha\Delta\mathbf{c}_R(\delta))\right|^2 \tag{16}$$

令上式（16）為零，可得最佳縮小係數 α 值為

表 1 本文所使用之特徵參數擷取規格

取樣頻率	8k Hz
音框長度	256 × 8k Hz ＝ 32ms
音框平移	128 × 8k Hz ＝ 16ms
漢明視窗	$0.54-0.46\cos(2.n\pi/N-1)$，N 為視窗長度
濾波器階數	12 階（P=12）
特徵向量 $\mathbf{v}^T=[\mathbf{c}^T \quad \mathbf{d}^T \quad e]$	12 維 LPC 逆頻譜係數 ***c***（Cepstral coefficients）＋ 12 維 LPC 逆頻譜係數差分 ***d***（Delta Cepstral coefficients）＋ 1 維對數能量差分 *e*（Delta log-energy）

$$\alpha_{opt}=\frac{\langle \Delta\mathbf{c}_R(\delta),(\mathbf{c}_T-\mathbf{c}_R(0))\rangle}{\langle \Delta\mathbf{c}_R(\delta),\Delta\mathbf{c}_R(\delta)\rangle} \qquad (17)$$

我們將觀察的結果應用到 HMM 參考模型中的均值向量之逆頻譜部份，可調整如下所示：

$$\{\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\}\rightarrow\{\boldsymbol{\mu}_i+\alpha\Delta\boldsymbol{\mu}_i(\delta),\boldsymbol{\Sigma}_i\} \qquad (18)$$

可得到最大相似分數 $L(\mathbf{c}_t;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$，其最佳純量係數為：

$$\alpha_{opt}=\frac{(\mathbf{c}_t-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}\Delta\boldsymbol{\mu}_i(\delta)}{\Delta\boldsymbol{\mu}_i{}^T(\delta)\boldsymbol{\Sigma}_i^{-1}\Delta\boldsymbol{\mu}_i(\delta)} \qquad (19)$$

5. 實驗環境與結果分析

我們將電信研究所提供之國語獨立數字分為四個群組，每個群組內有 25 個人，每人錄有 0-9 獨立數字語音，每個數字各 3 次，總共 3000 個聲音檔。混雜語音利用程式合成為不同訊雜比的語音檔。實驗中所使用的隱藏式馬可夫模型，其狀態的轉換是採由左至右（Left – to - Right）的形式，也就是在狀態的轉移上，只允許從某一狀態跳到下一狀態或是停留在原狀態上。在每個獨立數字語音的參考模型上包含前後 2 個靜音與 6 個短停留共 8 個狀態，而每個狀態分佈的觀測是由單一或兩個混合（Mixture）的高斯機率分佈表示[7]。表 1 為本文所使用之特徵參數擷取規格，表 2 為各個實驗時，使用語料的統計表，其中每個獨立數字語音的參考模型之特徵偏移向量 $\Delta\boldsymbol{\mu}_i(\delta)$，利用 30 個雜訊環境下的語音檔與原始的參考模型訓練所得。圖 5 到圖 10 為不同系統在不同雜訊環境下的辨識結果比較圖。

表 2 各系統使用語料統計表

	混合數 = 1；混合數 = 2		
	訓練語料	測試語料	調整語料
基本系統	1500	1500	0
線性調整法	1500	1500	0
偏移向量法	1500	1500	300

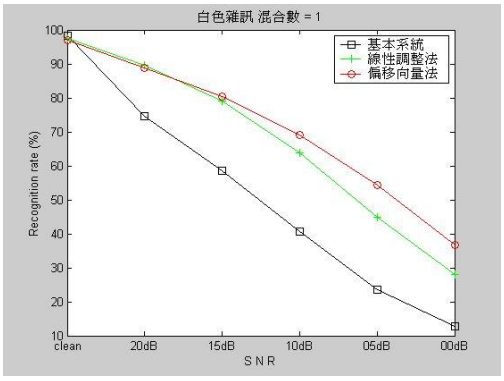


圖 5 白色雜訊 混合數 = 1 各系統辨識率

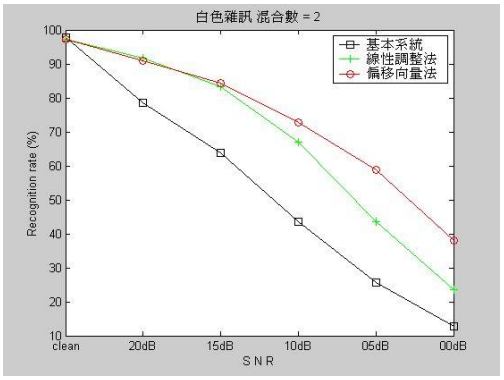


圖 6 白色雜訊 混合數 = 2 各系統辨識率

觀察圖 5 與圖 6，我們可以發現基本系統在雜訊環境下的辨認率相當的低，尤其在訊雜比 0dB 時，辨認率只有約 12%，使用線性調整法對參考模型補償後，0dB 時的辨認率提高到 23% ~ 27%，偏移向量法更提高到 37%，實驗結果表示這兩種補償的方法，在白色雜訊下都能夠使辨認率有所提昇，而偏移向量法的效果更高於線性調整法。接下來，我們觀察在非白色雜訊環境下的補償效果，圖 7 與圖 8 為人聲雜訊環境下的辨認率。

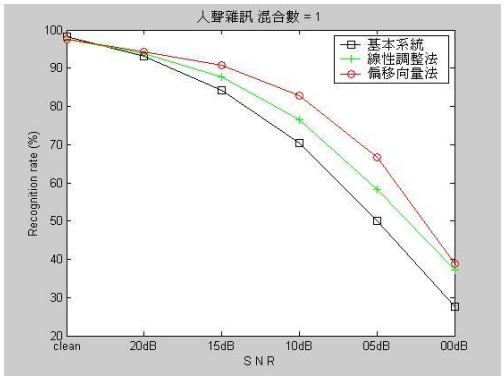


圖 7 人聲雜訊 混合數 = 1 各系統辨識率

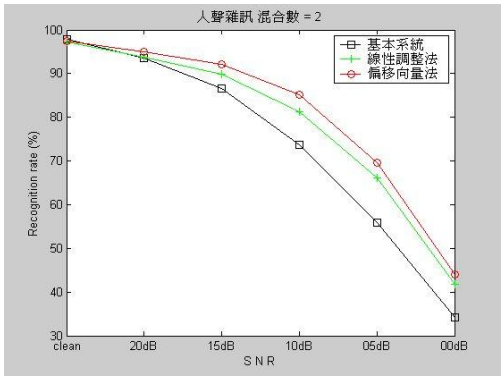


圖 8 人聲雜訊 混合數 = 2 各系統辨識率

由圖 7 與圖 8 可得知，線性調整法與偏移向量法在非白色雜訊之人聲雜訊環境下，亦能夠對辨認率的提昇有所幫助，而在另一種非白色雜訊之汽車雜訊下的辨認效果，我們可以從圖 9 與圖 10 觀察。

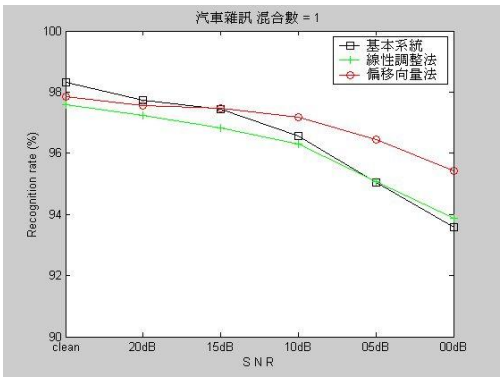


圖 9 汽車雜訊 混合數 = 1 各系統辨識率

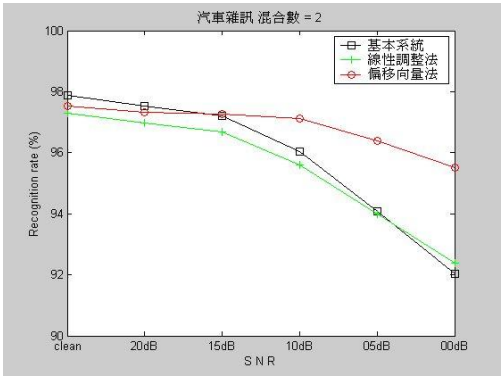


圖 10 汽車雜訊 混合數 = 2 各系統辨識率

在雜訊環境為汽車雜訊時的觀察，可發現線性調整法並不能夠對辨認率的提昇有所貢獻，反而使系統的辨認率下降，只有在 0dB 時，相較於基本系統多出了約 0.3%的辨認率，在這裡不同於線性調整法無法在汽車雜訊下對參考模型有所補償，我們提出的偏移向量法能夠較有效的補償參考模型，使辨認率有所提昇。

6. 結論

在本文中，我們將重點放在雜訊環境下語音參考模型的補償方法，探討的過程，我們觀察在不同的外加雜訊下對參考模型作補償調整。當外加雜訊為白色雜訊時，逆頻譜變動的方向會與乾淨語音的逆頻譜向量的方向相反，且會縮小到原點，所以線性調整法對參考模型乘上一最佳縮小因子來近似混雜語音就能夠對辨認率有所提昇。但是當外加雜訊為非白色雜訊時，補償的方法只

有最佳縮小因子是不夠的，例如在汽車雜訊的實驗裡，我們就可發現，因為逆頻譜變動的方向往純雜訊與乾淨語音間的特徵向量差值方向行進，線性調整法不但無法有效補償參考模型，反而使辨認效果變差，故我們引入特徵偏移向量，我們利用少量的混雜語音與參考模型中每個狀態之逆頻譜均值向量間的差值，求得特徵偏移向量，並考慮特徵偏移向量可能與實際使用時的環境並不匹配，所以我們再利用一最佳純量係數，乘上特徵偏移向量作為最佳線性調整量，加至原始的語音參考模型，使模型能夠更適應的在雜訊環境下使用。實驗的結果亦驗證了我們提出的特徵偏移向量之線性調整，較只有對逆頻譜作線性調整的方法，能夠有效的在各種雜訊環境下操作，且有更高的辨認率。

7. 參考文獻

[1] 李立民，(1995)，雜訊環境下語音辨認之研究，國立清華大學電機工程研究所博士論文。

[2] Guan, C., Chen, Y. and Wu, B., (1993), "Direct modification on LPC coefficients with application to speech enhancement and improving the performance of speech recognition in noise", Proceedings of IEEE international conference on Acoust., Speech Signal Processing, Vol. Ⅱ, pp. 107-110.

[3] Gales, M. J. F. and Young, S. J., (1993), "Cepstral parameter compensation for HMM recognition in noise", Speech Commun., No. 12, pp. 231-239.

[4] Hwang, T.-H., Lee, L.-M. and Wang, H.-C., (1997), "Feature adaptation using deviation vector for robust speech recognition in noisy environment", Proceedings of IEEE international conference on Acoust., Speech Signal Processing, pp. 1227-1230.

[5] Hwang, T.-H., Lee, L.-M. and Wang, H.-C., (1998), "Cepstral behaviors due to additive noise and a compensation scheme for noisy speech recognition", IEE Proc.-Vis. Image Sig. Process., Vol. 145, No. 5, pp. 316-321.

[6] Lee, L.-M., Chen, J.-K. and Wang, H.-C., (1994), "Nonlinear cepstral equalization method for noisy speech recognition", IEE Proc.-Vis. Image Signal Process., Vol. 141, No. 6, pp. 397-402.

[7] Rabiner, L. R., (1989), "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of IEEE, Vol.77, No.2, pp.257-286.

[8] Mansour, D. and Juang, B. H., (1989), "A family of distortion measures based upon projection operation for robust speech recognition", IEEE Trans. Acoust. Speech Sig. Process., No. 37, pp. 1659-1671.

[9] Moreno, P. J., Raj, B., and Stern, R. M., (1996), "A vector Taylor series approach for environment-independent speech recognition", Proceedings of IEEE international conference on Acoust, Speech Signal Processing, pp. 733-736.