基於離散倒頻譜之頻譜包絡估計架構及其於語音轉換之應用
A Discrete-cepstrum Based Spectrum-envelope Estimation Scheme and Its Application to Voice Transformation

古鴻炎　　蔡松峰
Hung-Yan Gu　and　Song-Fong Tsai

國立台灣科技大學 資訊工程系
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
e-mail: guhy@mail.ntust.edu.tw

摘要

基於前人以離散倒頻譜來逼近頻譜包絡之觀念、及其穩定化係數值之求解方法，本論文進一步研究實際實施時所面臨的兩個問題，其一是"頻譜峰點挑選"的問題，其二是"頻率軸尺度轉換"的問題，對這兩個問題我們提出了不錯的解決方法，然後用以建構一個頻譜包絡的估計架構(scheme)，測試實驗顯示該架構所估計出的頻譜包絡，確實比原始方法所估計出的準確不少。接著，我們應用所提出的頻譜包絡估計之架構，去製做出一個語音轉換系統，經由頻譜包絡估計、頻譜包絡伸縮、基頻移動、和信號重新合成等處理步驟，可把輸入語音信號的音色轉換成不同性別和年齡的其它音色。由聽測實驗的結果顯示，我們的語音轉換系統，的確可有效地達成音色轉換的功能。
關鍵詞: 頻譜包絡, 離散倒頻譜, 語音轉換, 音色轉換

一、前言

這裡"頻譜包絡"指的是頻譜振幅包絡(magnitude-spectrum envelope)，關於一個語音音框的頻譜包絡的估計，先前研究者已提出了一些方法，例如基於線性預測編碼(linear prediction coding, LPC)之方法[1, 2]，以全極(all pole)模型之頻率響應曲線來逼近語音的頻譜包絡，不過 LPC 頻響(頻率響應)曲線，在一個共振峰頻率的附近會比理想的頻譜包絡曲線低，而在頻譜變化較快速的頻率區段，則會比理想的頻譜包絡曲線高很多，如圖 1 裡一個/i/音音框的 LPC 頻響曲線所示，所以 LPC 頻響曲線和理想的頻譜包絡曲線之間會存在著不能忽略的誤差，這樣的誤差在一些應用裡(如語音轉換)，將會造成語音品質的衰退。

除了 LPC 之外，過去也有幾個以倒頻譜(cepstrum)爲基礎的頻譜包絡估計方法被提出，最簡單的一個是倒頻譜平滑法[1]，此法只保留倒頻譜係數的前幾個，而把後面的係數全部砍除(即令爲 0 值)，再作離散傅利葉轉換(discrete Fourier transform , DFT)，就可得到平滑的頻譜曲線，如圖 1 裡下方的那一條平滑曲線，很明顯地這樣的一條頻譜曲線並不是頻譜包絡，因爲它走在原始 DFT 頻譜的波峰與波谷之間，而不是沿著波峰行走。因此，Imai 和 Abe 兩人提出一個以倒頻譜爲基礎再作改進的方法[3, 4] ，稱爲 true envelope 估計法，然而此法的計算量很大而缺乏效率。另外，Galas 和 Rodet 兩人提出以離散倒頻譜(discrete cepstrum)來估計頻譜包絡的觀念[5]，後來 Cappé 和 Moulines 兩人則提出穩定化(regularization)的技術[6]，以解決使用離散倒頻譜來逼近頻譜包絡時所遇到的困難。我們覺得基於離散倒頻譜之估計法是一個不錯的方法，因此就著手研究此

方法，並且設法解決實際使用時所遇到的問題。

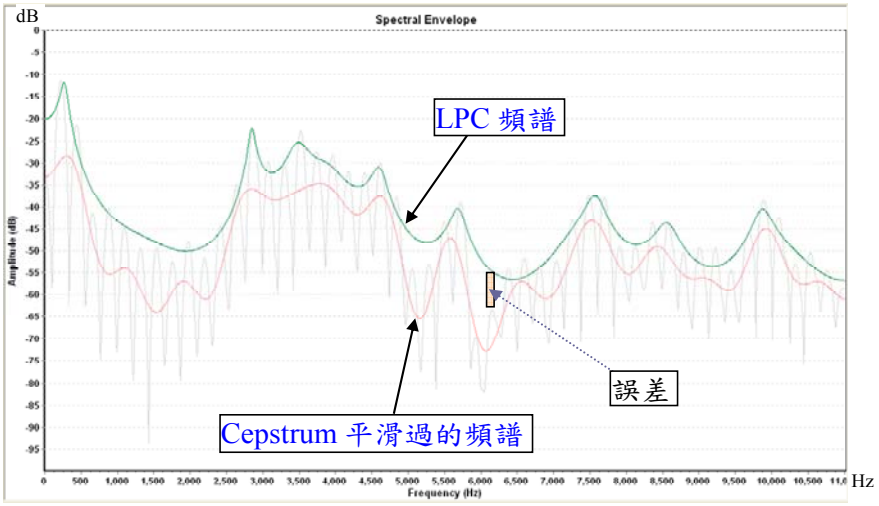


圖 1　/i/音音框之 LPC 頻譜包絡 和 倒頻譜平滑後之頻譜

本論文以離散倒頻譜之估計法爲基礎，研究、提出一個頻譜包絡估計的架構(scheme)，架構如圖 2 所示之處理流程，一個輸入的 20ms 之語音音框，首先進行基頻

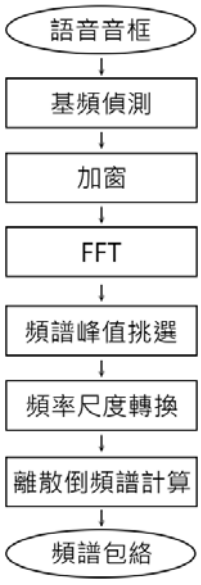


圖 2　頻譜包絡估計之架構

的偵測，以求出該音框的基頻值及判斷是否爲有聲(voiced)，求出的基頻值將會在"頻譜峰值挑選"之方塊內使用，在此我們使用了一種自相關(autocorrelation)函數搭配 AMDF (absolute magnitude difference function)的基頻偵測方法[1, 7, 8]。接著，將該音框信號乘上漢寧(Hanning)窗[1]，並於信號序列後面補上零，使序列長度成爲 1024 點，然後對該序列作 FFT (fast Fourier transform)計算而得到頻域上的頻譜振幅。之後，對頻譜振幅曲線作峰點(spectral peak)的挑選，並將選中的振幅峰點的頻率值作頻率尺度之轉換。使用挑選出的振幅峰點之振幅值和相對應的頻率值，就可帶入頻譜包絡的逼近準則(critera)[6]，去解出離散倒頻譜係數的數值，而離散倒頻譜係數則可用以算出所逼近的頻譜包絡。

在圖 2 的流程裡，我們主要作探討且提出新作法的是，"**頻譜峰點挑選**"和"**頻率尺度轉換**"兩方塊。雖然，核心的"**離散倒頻譜計算**"方塊，我們只是直接參照前人的成果

[6]，但是，如果沒有挑選到正確的頻譜峰點，或者沒有使用正確的頻率尺度，則"**離散倒頻譜計算**"所逼近出的頻譜包絡曲線，仍然會出現不能忽略的誤差，而不能算是理想的頻譜包絡曲線。由於過去的文獻，並未記載"**頻譜峰點挑選"**和"**頻率尺度轉換"**的確實作法，因此我們便著手研究這兩個問題，詳細情形在第三節和第四節分別說明。此外，我們也把圖 2 的頻譜包絡估計架構，應用於作語音轉換(voice transformation)，例如把成年女性的原始發音，經由轉換處理而得到女小孩的聲音、或成年男生的聲音，詳細情形在第五節裡說明。

二、 基於離散倒頻譜之頻譜包絡估計

2.1 離散倒頻譜

離散倒頻譜之觀念是由 Galas 和 Rodet 所提出[5]，他們採取以頻域上的最小平方準則(least-squares criterion)來求取倒頻譜係數，這和原本的實數倒頻譜(real cepstrum)係數的求取方式不同。原本的求取方式是把對數頻譜振幅, $\log|X(k)|$, k=0,1, …, N-1, 去作反離散傅利葉轉換(IDFT)而得到，令所得的倒頻譜係數爲 $c_0, c_1, \ldots, c_{N-1}$，之後，再將這些倒頻譜係數作 DFT，就可還原求得對數振幅頻譜，其公式[8]爲

$$\log\left|X(k)\right| = \sum_{n=0}^{N-1} c_n e^{-j\frac{2\pi}{N}kn}, \quad 0 \le k \le N-1 \quad (1)$$

由於對數頻譜振幅 $\log|X(k)|$是偶對稱的，即 $\log|X(k)| = \log|X(N-k)|$，所以由對數頻譜振幅求出的倒頻譜係數也是偶對稱的，即 $c_k = c_{N-k}$，依據這個偶對稱的特性，公式(1)可被推導成爲

$$\log\left|X(k)\right| = c_0 + 2\sum_{n=1}^{N/2-1} c_n \cos(\frac{2\pi}{N}kn) + c_{n/2}\cos(\pi k), \quad 0 \le k \le N-1 \quad (2)$$

這是因爲在轉換核心的虛部(imaginary part) 奇函數 sin(·)和偶對稱的倒頻譜係數序列會加總成 0 值。

若公式(2)裡只保留前面少數幾個(例如 p+1 個)倒頻譜係數，則它可用以計算出一個平滑過的振幅頻譜，計算公式爲

$$\log S(f) = c_0 + 2\sum_{n=1}^{p} c_n \cos(2\pi fn) \quad (3)$$

但是，如果想要以公式(3)來逼近頻譜包絡，則公式(3)裡的倒頻譜係數 c_k，就不是使用 IDFT 來求取了，而是要先定義一些欲被滿足的頻譜包絡限制，然後在儘量滿足這些頻譜包絡限制的條件下，去求解出最佳的倒頻譜係數 c_k 的數值，如此求出的倒頻譜係數就稱爲離散倒頻譜係數。前述所謂的頻譜包絡限制，其實是從原始的 DFT 頻譜$|X(k)|$上，找出 L 組具有代表性的頻譜振幅峰值及其頻率(a_k, f_k), k=1, 2 …, L。由於 L 通常比倒頻譜階數 p 大許多，所以需要使用一個加權式最小平方準則(weigthed least-squares criterion)，來最小化這 L 個頻率點上(f_k, k=1, 2 …, L)，$S(f_k)$和 a_k 之間的誤差，也就是最小化

$$\varepsilon = \sum_{k=1}^{L} w_k \cdot \left|\log a_k - \log S(f_k)\right|^2 \quad (4)$$

其中 w_k 是一個權重值與頻率 f_k 有關，對於不同的頻率值給予不同的加權，可藉以求得較好的頻譜包絡。若把公式(4)以矩陣形式來表示，則可改寫成下式[6]

$$\varepsilon = \left|a - Mc\right|^2 \cdot W = (a - Mc)^T W (a - Mc) \quad (5)$$

其中 $a = \left[\log(a_1), ,\cdots \ \log(a_L)\right]^T$ ； $c = [c_0, \cdots; \ c_p]^T$ 是一個維度 p+1 的向量，代表未知的倒頻譜係數；

$W = \begin{bmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_L \end{bmatrix}$ 爲對角矩陣，對角原素爲加權值；

$$M = \begin{bmatrix} 1 & 2\cos(2\pi f_1) & 2\cos(2\pi f_1 2) & \cdots & 2\cos(2\pi f_1 p) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2\cos(2\pi f_L) & 2\cos(2\pi f_L 2) & \cdots & 2\cos(2\pi f_L p) \end{bmatrix}$$

若要得到一組最佳的倒頻譜係數 c，也就是要最小化公式(4)的誤差，這可由公式(5)推導出如下之線性解

$$c = (M^T W M)^{-1} M^T W a \quad (6)$$

如此，透過矩陣逆轉換與矩陣相乘，即可得到一組最佳的離散倒頻譜係數。

2.2 離散倒頻譜之穩定化(regularization)

由前一小節的說明可知，離散倒頻譜的原理是，在頻域上以最小平方準則去求解倒頻譜係數，這樣的求解方法在實務上會遭遇到一個問題，而使它變成不實用，那就是 ill-conditioning 問題。由於矩陣 $M^T W M$ 通常是條件很差的矩陣，這將會導致包絡曲線有非常大的誤差，這意味著只要數據(a_k, f_k) 有些微的改變(例如四捨五入)，則計算出的倒頻譜係數就會有劇烈的變化，而包絡曲線也會因爲 f_k 的改變而產生過大的起伏。

以圖 3 爲例，虛線曲線所表示的是，使用 40 階離散倒頻譜係數、和非線性頻率尺度所求得的頻譜包絡，雖然包絡曲線有正確的經過前 7 個振幅峰點，不過峰點與峰點之間的曲線變化得非常劇烈。在許多頻譜包絡的應用裡，相鄰峰點之間的振幅值必須是可以計算的，如果有可能出現如圖 3 所示的情況，那麼此曲線在實務上將無法作爲頻譜包絡曲線。

先前研究者也發現，在以下三種情況下，上述問題發生的機會將會提高，分別是：(a)當有很寬的頻率軸區間沒有可逼近的振幅點時；(b)當兩個相鄰的頻率點 f_k, f_{k+1} 很靠近，並且此兩點的振幅差異很大；(c)當離散倒頻譜係數的個數 p 很接近欲逼近的頻率點數 L 時，通常在基頻較大時發生。因此，Cappé 和 Moulines 提出一個穩定化技術[6]，用以排除包絡曲線的不合理起伏，他們在求解一組離散倒頻譜係數時，除了依據最小化平方誤差準則，還將包絡曲線的平滑程度納入考慮，如此修改過的準則就變成如下公式：

$$\varepsilon = \sum_{k=1}^{L} w_k \cdot \left|\log a_k - \log S(f_k)\right|^2 \ + \ \lambda \cdot R\big(S(f)\big) \quad (7)$$

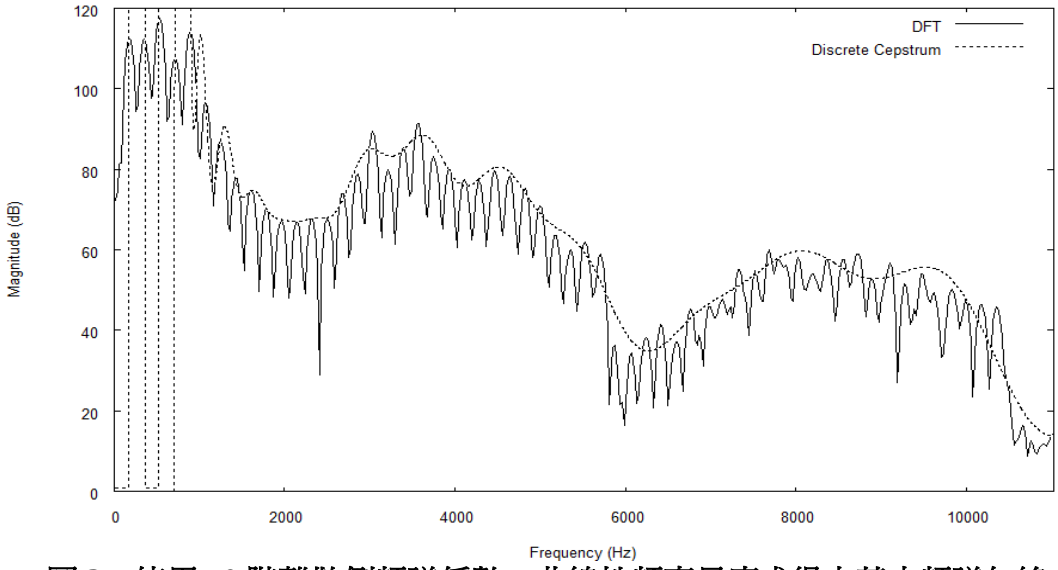


圖 3　使用 40 階離散倒頻譜係數、非線性頻率尺度求得之基本頻譜包絡

其中 $R(S(f))$是一個補償函數，用以量測包絡曲線的平滑程度，若包絡曲線越平滑則其值越小，反之越大；λ 是正規化參數，用以控制平滑程度和最小平方誤差準則之間的相對權重，λ 值越大則逼近出的頻譜包絡越平滑。一個典型的補償函數如下式[6]：

$$R\left(S(f)\right)=\int_0^{\pi}\left[\frac{d}{df}S(f)\right]^2 df \tag{8}$$

當把公式(3)式代入公式(8)中，可推導得

$$R\left(S(f)\right)=c^T U c\ ,\quad U=8\pi^2\begin{bmatrix}0 & & 0\\ & 1^2 & \\ & & \ddots\\ 0 & & p^2\end{bmatrix} \tag{9}$$

如此從公式(7)推導出的最佳解如下式[6]

$$c=(M^T W M+\lambda U)^{-1}M^T W a \tag{10}$$

其中正規化參數λ的較佳值在 0.0001 附近，如此矩陣 ill-conditioning 的問題就可獲得解決，並且依然能逼近出良好的頻譜包絡曲線，圖 4 裡的虛線曲線就是使用穩定化離散倒頻譜估計方法，所逼近出的頻譜包絡，很明顯地在低頻部分過度起伏的情況已經獲得改善，至於圖 4 裡的實線 DFT 曲線則與圖 3 裡的相同。

三、 頻譜峰值挑選

一般來說，頻譜包絡可看成是 DFT 振幅頻譜上連結各峰點(spectral peak)的曲線，因此離散倒頻譜的估計，採取以最小化所選出的峰點振幅值 a_k 與包絡曲線 $S(f)$之間的平方誤差作爲準則。由此可推知，峰點的挑選是一個相當重要的前置步驟，如果採取的是簡單的峰點挑選方法，例如選出全部的頻譜峰點，這將會導致不良的頻譜包絡曲線被逼近出來，而使用此種頻譜包絡進行語音編碼或音高調整，也將會降低語音品質和造成音色的不一致。

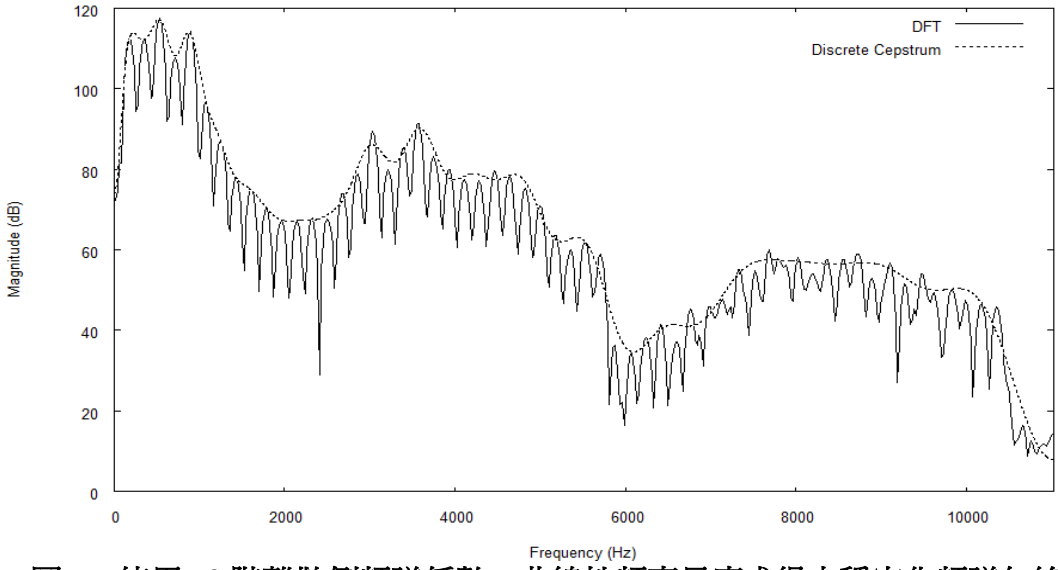


圖 4　使用 40 階離散倒頻譜係數、非線性頻率尺度求得之穩定化頻譜包絡

因此，我們採用 HNM (harm onic-plus-noise model) 之觀念[9, 10] ，當輸入的音框作基週偵測後判斷爲有聲時，就依據 HNM 偵測出的最大有聲頻率(maximum voiced frequency, MVF)，把該音框的 DFT 頻譜切割成低頻的諧波(harmonic)部分、和高頻的雜音(noise)部分。然後，對於諧波部分，依據偵測出的基頻值 F_0，在頻率範圍 [0 .5×F_0, 1.5×F_0] 內尋找出峰點的振幅值 a_1、及其對應的頻率值 f_1；接著在頻率範圍 [f_1+0.5×F_0, f_1+1.5×F_0] 尋找出另一峰點的振幅值 a_2、及其對應的頻率值 f_2；如此繼續找出其它峰點。如果在尋找的頻率範圍內沒有峰點，則會把尋找的範圍往後移動 0.5×F_0，再嘗試尋找峰點。此外，我們也設定了峰點振幅的門檻，以排除振福較小的峰點。

對於有聲音框的雜音部分，由於頻譜已無明顯的諧波結構，如圖 4 裡 5800Hz 之後的 DFT 頻譜曲線，相鄰峰點之間的頻率間隔變得隨機而非固定值，且峰點的振幅高度也變得隨機地起伏，因此我們認爲高頻雜音的頻譜，不能夠再使用與低頻部分相同的峰點挑選方法。在此，我們先使用 30 階的實數倒頻譜係數去計算出一個平滑過的頻譜曲線，然後依據此平滑的頻譜曲線，把 MVF 之後所有 DFT 頻譜振幅大於平滑頻譜振幅的峰點，都算是要選出的峰點。至於無聲的音框，我們就直接把 MVF 設爲 0，然後採取上述雜音頻譜峰點的相同挑選方法。圖 5 裡顯示了一個以我們的頻譜峰點挑選方法，所挑選出的頻譜峰點結果，選出的頻譜峰點以符號"+" 表示。

四、 離散倒頻譜階數與頻率軸尺度

4.1 離散倒頻譜之階數

由第二節的說明可知，離散倒頻譜係數的個數 p 必須先固定，然後才能去解 p 個聯立方程式來求得離散倒頻譜係數。至於 p 值要設爲多少? 若使用太小的 p 值(如 $p < 10$)，則包絡曲線的起伏次數會較少，而無法準確地逼近大多數的頻譜包絡形狀。然而隨著 p 值的增加，解聯立方程式的計算量也會增加，不過爲了準確地逼近大多數的頻譜包絡形狀，以避免音質下降及保持音色的一致，我們認爲加大 p 值是必要的。那麼 p 值應設爲多少? Shiga 和 King 曾提到[11]，若要得到精確的頻譜包絡，則較高階數(如 48~64)的倒頻譜係數是需要的。

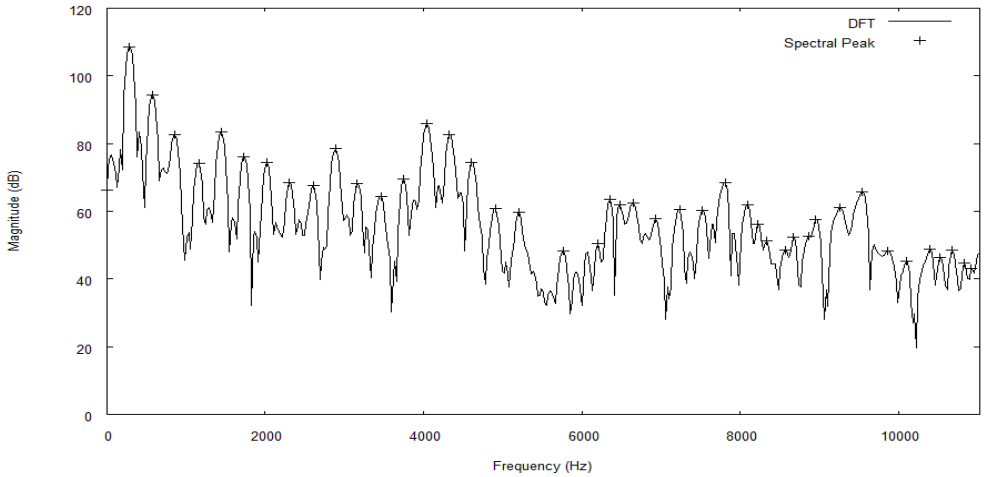


圖 5　一個頻譜峰點挑選的例子

在此，我們以實驗的方式來探討離散倒頻譜階數和頻譜包絡之逼近誤差的關係，實驗裡我們使用的誤差量測公式如下:

$$Es = \frac{1}{Nr}\sum_{t=0}^{Nr-1}\left[\frac{1}{L}\sum_{k=1}^{L}\left|20\log_{10} a_k^t - 20\log_{10} S(t, f_k)\right|\right] \qquad (11)$$

其中 Nr 表示語音音框的總數，a_k^t 表示第 t 個音框裡的第 k 個頻譜峰點的振幅，$S(t, f_k)$ 表示第 t 個音框以離散倒頻譜所逼近出的頻譜包絡。實驗後計算出的的逼近誤差如圖 6 所示，橫軸表示離散倒頻譜的階數，縱軸則是量得的逼近誤差 Es 的數值，觀察圖 6 可發現， 隨著階數的增加 Es 值會明顯地下降，直到階數高於 30 時，Es 值下降的幅度才趨於緩和，因此我們決定把 p 值設定爲 40，以確保逼近出的頻譜包絡的準確性。

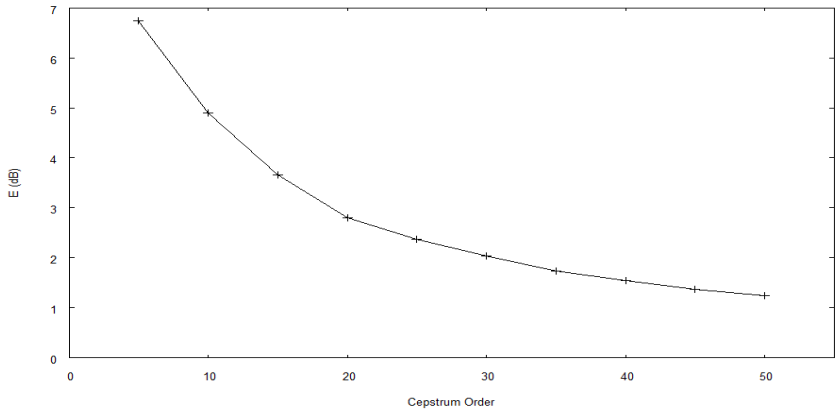
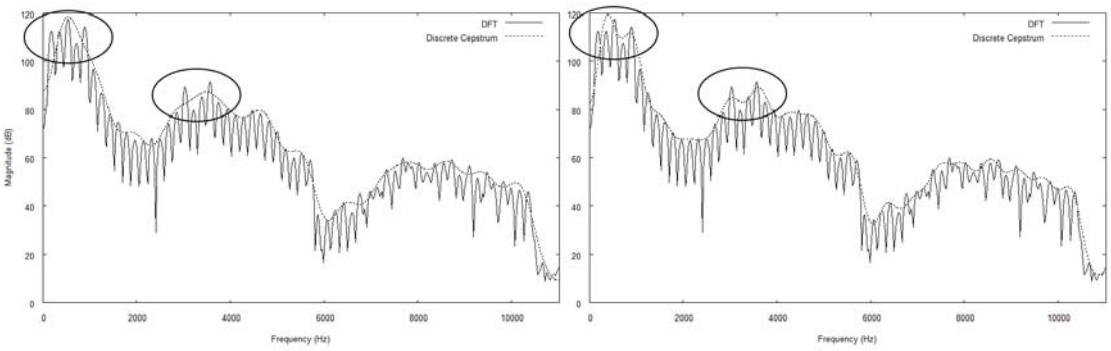


圖 6　不同階數之離散倒頻譜的頻譜包絡逼近的誤差

4.2 頻率軸尺度

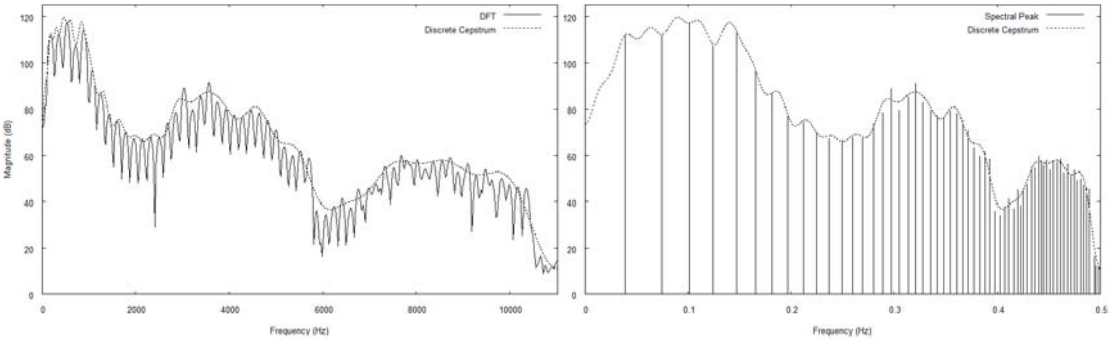
雖然高階數的離散倒頻譜已經可以逼近出不錯的頻譜包絡，不過當我們觀察某些語音信號的頻譜時，發現高階數的離散倒頻譜所逼近出的頻譜包絡上，仍然會出現不小的、不能忽略的逼近誤差。例如圖 7(a)裡的頻譜包絡曲線，它是使用 30 階的離散倒頻譜所逼近出的，圖中圈起來的部分顯示發生較大誤差的地方，如果我們將階數提高到 40，則會得到如圖 7(b)裡的頻譜包絡曲線，雖然 1,000Hz 與 3,000Hz 附近的誤差，已經獲得了一些改善，不過還是不夠理想。



(a)使用 30 階之離散倒頻譜　　(b)使用 40 階之離散倒頻譜

圖 7 以線性頻率尺度之離散倒頻譜所逼進之頻譜包絡

圖 7 裡的情況通常是發生在基頻較低的語音音框裡，由於相鄰峰點的頻率值非常接近，而發生頻譜包絡快速變化的情況，這種快速變化的頻譜包絡無法由低階數的離散倒頻譜去作準確的逼近，尤其是在低頻的區段。要解決這種問題，一個普遍被採取的觀念是，使用非線性的頻率軸來擴大低頻區段在整個頻率軸所佔的比率，而常見的非線性頻率尺度如梅爾尺度(Mel Scale)或巴克尺度(Back Scale)。實際上的作法是，在挑選出頻譜峰點之後，先將各峰點對應的頻率 f_k 作頻率尺度的轉換 $\hat{f}_k = warp(f_k)$，才去求解離散倒頻譜係數，之後當要計算所逼近的頻譜包絡時，也需作頻率軸尺度的轉換。圖 8(a)裡的頻譜包絡曲線就是使用梅爾尺度所逼近出的，相較於圖 7(b)的線性尺度，低頻的峰點都有確實地被包絡曲線通過，但是 3,000Hz 附近的峰點仍然存在誤差。

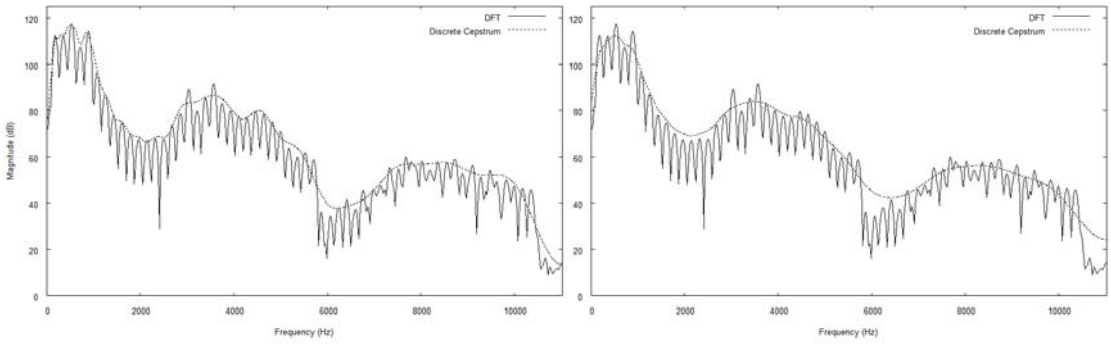


(a)橫軸爲線性頻率之包絡曲線　　(b)橫軸爲梅爾頻率之包絡曲線

圖 8　以梅爾頻率尺度和 40 階之離散倒頻譜所逼進的頻譜包絡

此外，在圖 8(a)的低頻部分，也可看到頻譜包絡曲線過度起伏的情況，原因是梅爾頻率尺度會把低頻部分的相鄰峰點的間隔擴大而引起過度起伏，這種過度起伏可從圖 8(b)裡觀察到。如果改成使用巴克頻率尺度，則將使包絡曲線過度起伏的情況更加嚴重，因爲巴克尺度會讓低頻部分所佔的比率更爲變大，並且相鄰的高頻峰點間的間隔也會變得更窄，而會使高頻部分的包絡曲線變得更爲平滑。

第二節裡曾提到包絡曲線過度起伏的解決方法，亦即穩定化之技術。到目前爲止，我們都只把正規化參數 λ 設爲 0.0002，爲了改善上述的過度起伏情形，我們嘗試加大正規化參數 λ，結果得到如圖 9 所示的頻譜包絡曲線，圖 9(a)裡設定 $\lambda = 0.001$，而在圖 9(b)裡設定 $\lambda = 0.01$。雖然隨著正規化參數 λ 的增加，低頻部分的頻譜包絡曲線會漸趨平滑，不過頻譜包絡的起伏程度也受到了限制，而導致整體 L 個峰點的逼近誤差也隨著增加。



(a) $\lambda = 0.001$ (b) $\lambda = 0.01$。

圖 9　在梅爾頻率尺度和 40 階倒頻譜之條件下加大穩定化參數 λ

因此，我們嘗試調整頻率軸尺度，也就是調整低頻部分在整個頻率軸所佔的比率，目的是讓離散倒頻譜階數大於 30 時，頻譜包絡曲線不會在低頻部分出現過度起伏的情況，並且在 2,000Hz ~ 6,000Hz 的頻率軸區段可以減小峰點振幅的逼近誤差。經過多次的實驗觀察，我們歸納出一種如下列公式所示之頻率尺度，

$$warp(f) = \log(1 + \frac{f}{1{,}750}) \tag{12}$$

圖 4 裡的頻譜包絡曲線就是使用此種頻率尺度來求解離散倒頻譜而逼近出的，可以發現小於 1,000Hz 之頻率部分，頻譜包絡曲線沒有過度起伏之情形，並且可準確地通過各個峰點，此外在 2,000Hz ~ 6,000Hz 的頻率部分，頻譜包絡曲線比起梅爾尺度所求出者，有著更佳的起伏能力。

爲了比較兩種頻率尺度，即我們提出的公式(12)之頻率尺度和梅爾頻率尺度，對於以離散倒頻譜逼近頻譜包絡所產生的誤差，在此我們就分別在不同的頻率範圍做逼近誤差的量測實驗，四種頻率範圍分別是 (a) 0 ~ 2,000Hz，(b) 0 ~ 4,000Hz，(c) 0 ~ 6,000Hz，(d) 0 ~ 11,025Hz，而誤差量測的方式仍然如公式(11)。實驗後我們得到如圖 10 所顯示的結果，也就是說在 0 ~ 2,000Hz 之頻率範圍，兩種頻率尺度有著類似的逼近誤差，但是在其它三種頻率範圍作量測時，梅爾尺度的頻譜包絡逼近誤差，明顯地會高於我們提出的頻率尺度，並且頻率範圍愈大時，我們所提的頻率尺度的逼近誤差會和梅爾尺度的逼近誤差愈來愈拉開。

五、語音轉換之應用

5.1 語音轉換系統

這裡所說的語音轉換(voice transform ation)，指的是作頻譜包絡曲線的伸展或收縮(相當於作頻率軸的 scaling)，及基本頻率的移動(frequency shifting)，以把輸入語音信號的音色改變成另一個人的音色，例如把成年女生的語音轉變成成年男生的語音、或小孩的語音。過去，作語音轉換常被使用的是 phase vocoder (PV) 之技術[12, 13]，但是基本的 PV 技術，並不能讓**頻譜包絡伸縮**和**基頻移動**兩者作獨立的控制。在本論文裡，我們決定採取電腦音樂之加法式合成法(additive synthesis) [12]及 HNM 的觀念[9]，來讓頻譜包絡伸縮和基頻移動兩者可以被分別地控制，而實作上則需要應用前面說明的頻譜包絡之估計方法，來求得輸入語音各音框的頻譜包絡曲線，然後才據以作頻譜包絡的伸縮。我們所製做的語音轉換系統，其程式介面如圖 11 所示，pitch contour 區塊顯示從原始語音所分析出的基週軌跡，waveform 區塊上下分別顯示原始語音及轉換過語音的波形。

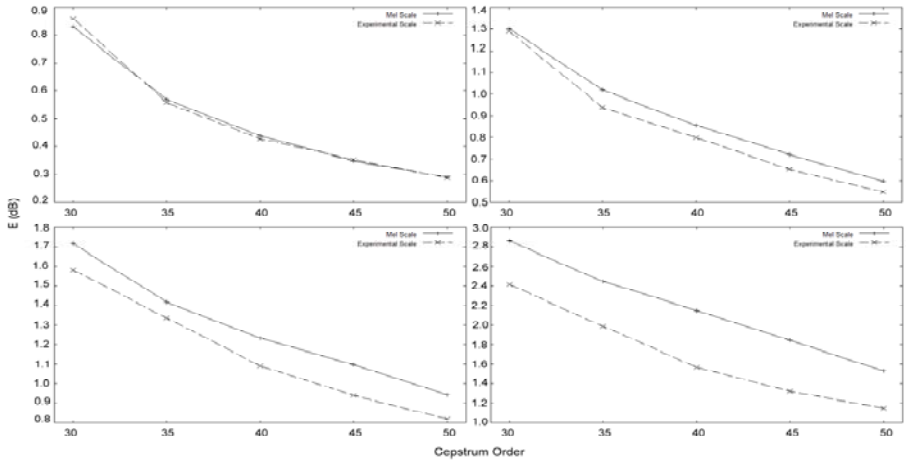


圖 10　在四種頻率範圍比較本論文尺度、梅爾尺度的頻譜包絡逼近之誤差: (左上) 0~2,000Hz，(右上) 0~4,000Hz，(左下) 0~6,000Hz，(右下) 0~11,025Hz

至於此系統的處理流程則如圖 12 所示，輸入的語音信號，先切割成長 20ms 重疊 10ms 的音框序列，然後對各個音框作如圖 2 所示的處理步驟，接著依據公式(10)去求取 40 個離散倒頻譜係數，再依據公式(3)就可計算出頻譜包絡曲線；至於"頻譜包絡伸縮"、"基頻移動"、"信號重新合成"等三個方塊的細節，將於下面各子節分別作說明。關於此系統的處理速度是，在 Intel T5600 1.83GHz CPU 的筆記本電腦上，處理 1 秒鐘的語音，平均需花 0.75 秒的時間。

5.2 頻譜包絡伸縮

不同年齡與性別的語者所發出的語音信號，在聲學上的主要差異是，語音頻譜上之共振峰頻率(formant frequency)值的高低會有明顯的差別。一般來說男生由於聲道(vocal track)較女生的長，所以男生語音的共振峰頻率值會比女生的低。因此，若要把輸入的語音信號轉換成不同性別與年齡的語音信號，則共振峰峰頻率的調整是必需的，不過，要求取出正確的共振峰頻率值、及直接修改它，並不是容易的事，因此我們採取對頻譜包絡作伸展或收縮的處理，以達到共振峰頻率的移動。

對頻譜包絡作伸、縮處理一個例子如圖 13 所示，圖 13(a)畫的是原始的頻譜包絡，令表示此包絡的函數是 $vs(f)$，如果我們對頻譜包絡作收縮，例如令收縮過的包絡的表示函數爲 $vc(f)$，且令 $vc(f) = vs(\frac{10}{7}f)$，則收縮的包絡將如圖 13(b)所示，如此就可把全部的共振峰頻率都調低；相反地如果我們對頻譜包絡作伸展，例如令伸展過的包絡的表示函數爲 $ve(f)$，且令 $ve(f) = vs(\frac{7}{10}f)$，則伸展的包絡將如圖 13(c)所示，如此就可把全部的共振峰頻率都調高。

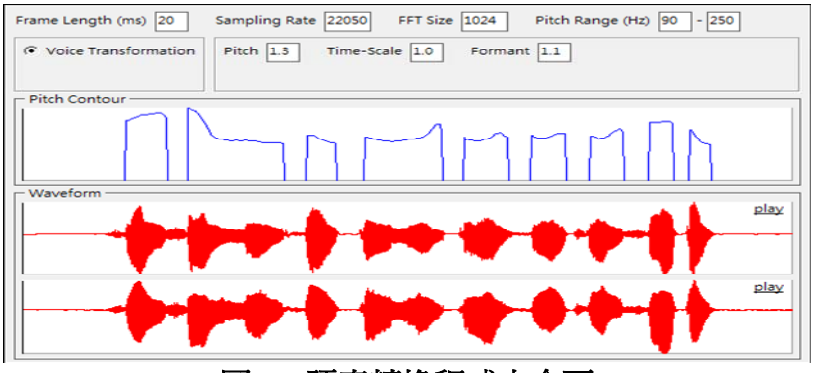


圖 11　語音轉換程式之介面

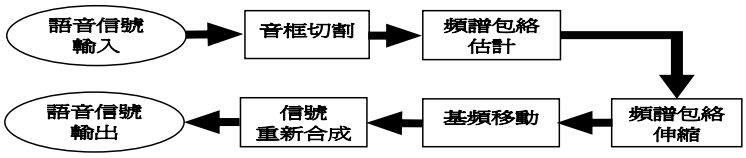
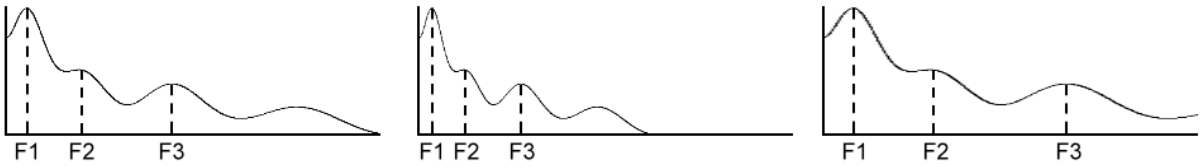


圖 12 語音轉換處理之主流程



(a)原始之頻譜包絡　　(b)收縮之頻譜包絡　　(c)伸展之頻譜包絡

圖 13 頻譜包絡之收縮與伸展

5.3 基頻移動

調整音高(pitch)的高低可讓聲音轉爲尖銳或低沉，但只調整男性語音的音高，並不能轉換出具有女性或孩童音色的語音，因此，若要把語音轉換成不同性別與年齡的語音，則共振峰頻率的調整(也就是頻譜包絡伸縮的調整)是必需先作的，然後才來作音高的調整。

當作完 5.2 小節的頻譜包絡伸縮之後，就可使用該頻譜包絡 *vc*(*f*) (或者 *ve*(*f*))來設定新的諧波參數，假設目前音框的原始基頻是 180Hz，而現在要把基頻調高到 250Hz, 則在此基頻的各個倍頻上的諧波，它們的振幅高度可根據 *vc*(*f*)來求得，也就是 *vc*(250), *vc*(500), *vc*(750), …, 這相當於在新的基頻及其倍頻上對頻譜包絡取樣，用以取代原先的諧波頻率和振幅，結果會得到如圖 14 所示的新諧波結構。在此一個諧波的參數只有頻率和振幅，我們暫時不使用相位之資訊。

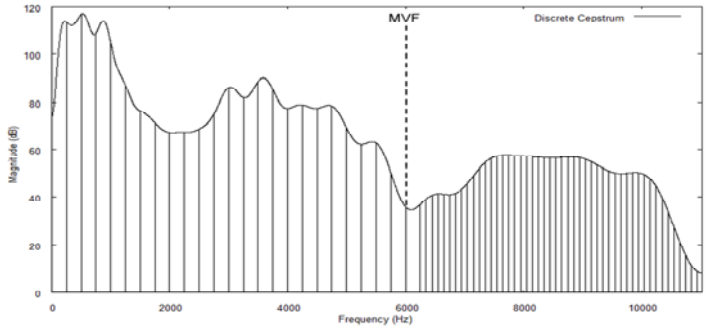


圖 14 基頻爲 250Hz 之諧波結構

5.4 信號重新合成

由於我們是根據語音音框在頻域上分析出的頻譜包絡，來製做語音轉換的功能，所以相對地也必須採取以頻域參數建構的信號模型，來作語音信號的重新合成。本論文採取的信號模型是諧波加雜音模型(harmonic-plus-noise m odel, HNM) [9, 10] ，它是由 Y. Stylianou 所提出，除了考慮語音信號裡低頻部分的諧波特性之外，還考慮了高頻部分的雜音特性，所以可以更確切地掌握語音信號的特性。HNM 模型提供了一個最大有聲頻率(maximum voiced frequency , M VF)的偵測方法，找出 MVF 之後，就可把一個語音音框的頻譜分割成低頻的諧波部分、和高頻的雜音部分，如圖 14 所示。

令第 *i* 個語音音框由 5.3 子節所求出的諧波參數是 f_k^i , a_k^i , k=1, 2, …, L^i, f_k^i 與 a_k^i 分

別表示第 k 個諧波的頻率與振幅；再令第 i+1 個語音音框由 5.3 子節所求出的諧波參數是 f_k^{i+1}, a_k^{i+1}, k=1, 2, …, L^{i+1}。如此，當要合成第 i 和第 i+1 音框之間時刻 t 的諧和(harmonic)信號之樣本 $h(t)$時，我們先以如下公式作線性內差，

$$f_k(t) = f_k^i + \frac{f_k^{i+1} - f_k^i}{N} t, \quad k = 1,2,...,L$$
$$a_k(t) = a_k^i + \frac{a_k^{i+1} - a_k^i}{N} t, \quad k = 1,2,...,L \tag{13}$$

以求取時刻 t 時各諧波的頻率與振幅，其中 N 表示相鄰音框之間的位移樣本數(frame shift)，L 是 L^i 和 L^{i+1} 的較大者，因此當 L^i 小於 L^{i+1} 時，就要把 a_k^i, k= L^i +1,…, L^{i+1} 設爲零值。然後，以如下公式計算 $h(t)$，

$$h(t) = \sum_{k=1}^{L} a_k(t) \cdot \cos(\phi_k(t)), \quad 0 \le t < N$$
$$\phi_k(t) = \phi_k(t-1) + 2\pi \cdot f_k(t) / 22,050 \tag{14}$$

其中 $\phi_k(t)$ 表示第 k 個諧波累積到時刻 t 時的相位量，關於初始值 ϕ_k(-1) ，我們令其等於前一音框裡的 $\phi_k(N\text{-}1)$ 以便維持相位的連續性，而當音框編號 i 爲 0 時就以亂數來設定，此外 22,050 是取樣率。

關於雜音(noise)信號的合成，我們採取 HNM 文獻上提到的一個作法，就是把雜音當作是 MVF 之後頻率間隔固定爲 100Hz、但振幅會隨時間改變之一些弦波的加總。令 mvf 爲第 i 和第 i+1 音框的 MVF 的較大者，則依 mvf 可決定頻率 index 之下限 KL=mvf / 100, 而其上限明顯地是 KU=22,050 / 100，如此，對於第 i 和第 i+1 音框之間時刻 t 的雜音信號樣本 $g(t)$，我們以如下公式來計算，

$$g(t) = \sum_{k=KL}^{KU} b_k(t) \cdot \cos(\psi_k(t)), \quad 0 \le t < N$$
$$\psi_k(t) = \psi_k(t-1) + 2\pi \cdot k \cdot 100 / 22,050 \tag{15}$$

其中 $b_k(t)$ 表示時刻 t 時第 k 個弦波的振幅，其值也是以類似公式(13)之線性內差來求得，$\psi_k(t)$ 表示第 k 個弦波累積到時刻 t 時的相位量，其初始值也是以亂數來設定。最後，將 $h(t)$與 $g(t)$相加，即可得到時刻 t 的合成信號樣本。

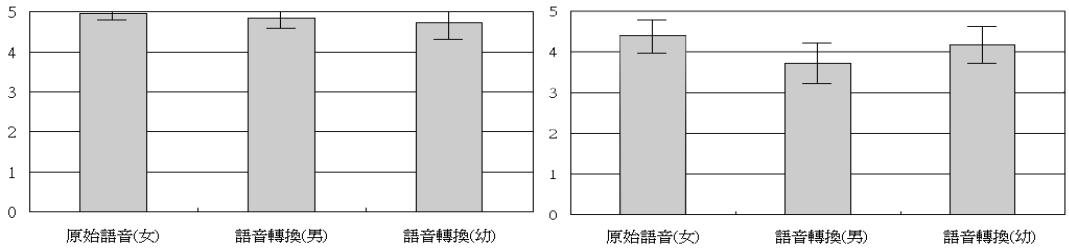
5.5 聽測實驗

爲了評估我們的語音轉換系統的效能，接著就進行主觀的聽測評估實驗。聽測的語料包含了成年女性發音的 3 句原始語句，和成年男性發音的 2 句原始語句。依據女性原始語句，我們系統進行了兩種轉換處理，第一種是設定頻譜包絡收縮成 80%、且基頻移動到原基頻的 60%，以轉換出男性的語音；第二種是設定頻譜包絡伸展成 130%、且基頻移動到原基頻的 140%，以轉換出孩童的語音。另外，依據男性原始語句，我們系統也進行了兩種轉換處理，第一種是設定頻譜包絡伸展成 120%、且基頻移動到原基頻的 210%，以轉換出女性的語音；第二種是設定頻譜包絡伸展成 130%、且基頻移動到原基頻的 150%，以轉換出孩童的語音。如此，對於女性和男性原始語句，各有 2 組轉換出的語句，可供作聽測評估，這些語句(原始的和轉換出的)，可從網頁 http://guhy.csie.ntust.edu.tw/dcc/vt.html 去下載和試聽。在此評估的項目有兩項，分別是音色辨識度、和語音品質，音色辨識度在於評估轉換出來的語音音色與目標音色的接近

程度，而語音品質在於評估轉換出來的語音信號聽起來是否清楚且無失真。

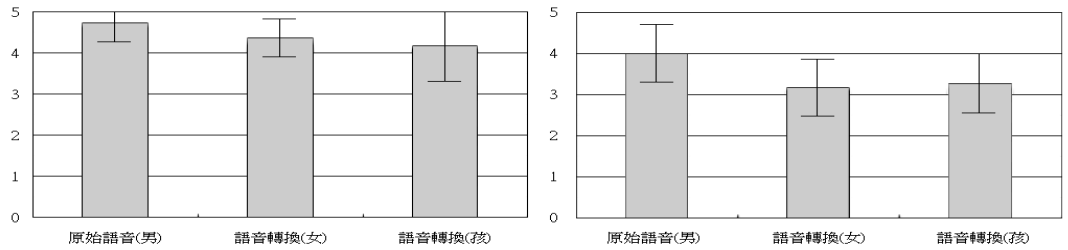
我們邀請了 13 位受測者來進行聽測評估，首先是對女性原始語句及轉換出的語句作聽測，實驗時一次讓一位受測者聆聽 3 組語句(女性原始、男性轉換、孩童轉換)，然後請他對這 3 組語句各給一個評分，評分的範圍由最高 5 分到最低 1 分，5 分代表非常相似(or 好) 而 1 分代表非常不相似(or 差)，可以打至小數點下一位。關於音色辨識度的評估，以女性語句爲例是詢問「聽測音檔的音色和女性語音音色的相似度爲何?」，至於男性語句、孩童語句可依此類推。關於語音品質的評估，詢問的方式則是「聽測音檔的語音品質是非常好、普通、或非常差?」。聽測實驗後，我們將 13 位受測者的評分作平均，所得到的平均值如圖 15 所示，在音色辨識度方面，轉換過的語句皆有相當高的辨識度，並且很接近原始語音的辨識度，如圖 15(a) 所示；在語音品質方面，轉換過的孩童語音跟原始音檔的語音品質比較接近，但下降一些，而轉換過的男性語音之語音品質則約比原始音檔的低 0.5 分。

另外，我們也請這 13 位受測者來對男性原始語句及轉換出的語句作聽測評估，實驗時一次讓一位受測者聆聽 3 組語句(男性原始、女性轉換、孩童轉換)，然後請他對這 3 組語句各給一個評分，評分方式如前段所述。實驗後將 13 位受測者的評分作平均，得到的平均分數如圖 16 所示，在音色辨識度方面，轉換出的孩童音色的分數比另二者的低一些，原因應是頻譜包絡伸展比例 130%不夠高，而使得音色感覺像國中生男孩而非孩童；在語音品質方面，轉換出的女性和孩童語句的分數都比原始男性語句的低，而且圖 16(b)三者的分數都分別比圖 15(b)三者的低，這似乎暗示，男性原始音作語音轉換所得的語音品質會比較差。



(a)音色辨識度　　(b)語音品質

圖 15　使用女性原始語句之聽測結果



(a)音色辨識度　　(b)語音品質

圖 16　使用男性原始語句之聽測結果

六、結論

雖然以離散倒頻譜來逼近頻譜包絡的觀念，多年以前就被提出了，但是實際實施時會面臨到三個問題，其一是求解穩定的頻譜包絡(無劇烈起伏)所對應的倒頻譜係數，這個問題已由前人解決，本論文則研究了另外二個問題，即"頻譜峰點挑選"和"頻率軸尺度轉換"的問題。關於頻譜峰點之挑選，我們應用 HNM 的觀念，先把頻譜分割成低頻

諧波和高頻雜音兩個部分，然後在低頻諧波部分依據所求出的基頻值去偵測諧波頂點，而在高頻雜音部分，則依據一般 cepstrum 平滑過的頻譜曲線去找出高過曲線的頻譜峰點。此外，關於頻率軸尺度的轉換，文獻上雖然提到 mel 尺度和 bark 尺度，但是我們從觀察逼近出的頻譜包絡曲線，發現 mel 尺度和 bark 尺度所得到的頻譜包絡仍然是不夠理想的，在中頻帶(3KHz~6KHz)常常會出現錯誤的頻譜包絡，因此我們在一番嘗試後，設計了一種頻率軸的尺度轉換公式，測試實驗顯示我們的轉換公式可明顯地降低頻譜包絡和頻譜峰點之間的逼近誤差。使用上述的解決方法，我們建構了一個頻譜包絡的估計架構。

此外，我們應用所提出的頻譜包絡估計之架構，去製做出一個語音轉換系統，該系統經由頻譜包絡估計、頻譜包絡伸縮、基頻移動、和信號重新合成等處理步驟，可把輸入語音信號的音色轉換成不同性別和年齡的其它音色。在信號重新合成步驟裡，我們採用了 HNM 信號模型，來分別合成諧波信號和雜訊信號，再作相加。爲了評估此系統的效能，我們進行了聽測實驗，由 13 位受測者的平均評分來看，我們系統的確可以有效地達成音色轉換之功能。根據這樣的音色轉換之表現，未來我們將會把本論文研究的頻譜包絡估計之架構，應用於特定語者之間的音色轉換的研究。

參考文獻

[1] D. O'Shaughnessy, *Speech Communications: Human and Machine*, IEEE Press, Piscataway, NJ, 2000.

[2] D. Schwarz and X. Rodet, "Spectral e nvelope estim ation and representation for sound analysis-synthesis", *Int. Computer Music Conference*, Beijing, China, pp. 351-354, Oct. 1999.

[3] S. Im ai and Y . Abe, "Spectral enve lope extraction by im proved cepstral m ethod", Electron. and Commun. in Japan, Vol. 62-A, No. 4, pp. 10–17, 1979. (in Japanese)

[4] A. Robel and X. Rodet, "Ef ficient spectra l en velope es timation and its app lication to pitch shifting and envelope preservation", *Int. Conference on Digital Audio Effects*, Madrid, Spain, pp. 1-6, September 2005.

[5] T. Galas and X. Rodet, "An im proved cepstral method for deconvolut ion of source filter systems with discre te s pectra: Applica tion to m usical sou nd signals ", *Int. Computer Music Conference* (ICMC), Glasgow, Scotland, pp. 82-44, 1990.

[6] O. Cappé and E. Moulines, "Regularization techniques for discrete cepstrum estimation", *IEEE Signal Processing Letters*, Vol. 3, No. 4, pp. 100–102, 1996.

[7] 古鴻炎、張小芬、吳俊欣，"仿趙氏音高尺度之基週軌跡正規化方法及其應用"，第十六屆自然語言與語音處理研討會(ROCLING XVI)，台北，第325-334 頁, 2004。

[8] 王小川，語音訊號處理(修訂二版)，全華圖書公司，台北，2009。

[9] Y. S tylianou, "Modeling speech bas ed on harmonic p lus n oise m odels", in *Nonlinear Speech Modeling and Applications*, eds. G . Chollet *et al*., S pringer-Verlag, Berlin, pp. 244-260, 2005.

[10] Y. Stylianou, *Harmonic plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification,* Ph.D. thesis, Ecole Nationale Supèrieure des Télécommunications, Paris, France, 1996.

[11] Y. Shiga and S. King, "Estim ating deta iled spectral envelope s using articulatory clustering", *Int. Conference on Spoken Language Processing* (ICSLP2004), Jeju, Korea, October 2004.

[12] F. R. Moore, *Elements of Computer Music*, Prentice-Hall, Englewood Cliffs, NJ, 1990.

[13] M. Dolson, "The phase vocoder: A tutorial", *Computer Music Journal*, Vol. 10, No. 4, pp. 14–27, 1986.