基於反射係數之頻譜內插方法及其在歌聲合成的應用

A REFLECTION COEFFICIENT BASED SPECTRUM INTERPOLATION METHOD AND ITS APPLICATION TO SINGING VOICE SYNTHESIS

古鴻炎 (Hung-Yan Gu)　蔡哲彰 (Zhe-Zhang Tsai)

國立台灣科技大學 資訊工程系

e-mail: guhy@mail.ntust.edu.tw

摘要

本論文提出一種頻譜包絡(spectral envelope)內插的方法，來把音節間不連續的共振峰軌跡轉變成連續、平滑的軌跡。該方法實際上是透過反射係數的內插來實施，因此，我們也提出了反射係數與頻譜包絡之間相互作轉換的實施程序。為了訂定相關參數(如反射係數的階數)的數值，及驗證前述方法的有效性，我們把前述方法實際應用於作歌聲信號的合成，然後觀察合成歌聲的聲譜圖(spectrogram)，的確可讓原先不連續的共振峰軌跡轉變成平滑，此外我們也進行了合成歌聲的聽測實驗，實驗結果顯示，音節間共振峰的軌跡變成連續銜接後，確實可用以提升合成歌聲的流暢性。

ABSTRACT

A spectral envelope (SE) interpolation method is proposed in this paper to convert discontinuous formant trajectories (FTs) at the boundary between two adjacent syllables into continuous FTs. Interpolating two SEs to obtain an intermediate SE is achieved actually by interpolating their corresponding reflection coefficients (RC). Therefore, the procedures for transforming an SE into its corresponding RC and vice versus are also developed in this paper. As to the order of RC and to verifying the effectiveness of the proposed method, we have applied the proposed SE interpolation method to the synthesis of singing voice signals (SVS). Then, the spectrograms of the synthesized SVS are inspected. It is found that the originally discontinuous FTs indeed become continuous. In addition, we have conducted listening tests using some synthesized SVS. The results show that the fluency of SVS can be significantly improved when the FTs at the boundaries between adjacent syllables are converted and become continuous FTs.

1. 前言

在此”頻譜包絡”指的是振幅頻譜包絡(spectral magnitude envelope)，頻譜包絡的一個例子如圖 1 裡較為平滑的曲線，它是由一個/bi/音音框(frame)的 DFT (discrete Fourier transform)頻譜所估計出的。在語音處理的一些子領域中，會需要在兩條頻譜包絡曲線之間作內插，以求得兩者之間過渡的、符合連續性要求的頻譜包絡。例如在基於單元串接之語音合成系統裡，為了避免在串接點上發生頻譜不連續(spectral discontinuity)的問題，就需要作頻譜包絡的內插，再依據內插出的頻譜包絡去產生語音信號；雖說過去有一些方法被提出來減輕串接點上頻譜不連續的問題[1]，但那並未根本地解決問題。在歌聲合成的系統裡，由於可能的音高(pitch)、音長(duration)、音節之組合數量更為龐大，因此當從語料庫(corpus)選取合成單元來作串接時，串接點上幾乎都會發生音高軌跡(pitch contour)不連續、及頻譜的不連續，所以有需要在串接點的附近作頻譜包絡內插。一般來說，合成的語音或歌聲如果在單元邊界含有不連續的頻譜，則人耳聽到該聲音時會有不流暢、與不自然的感覺，即不像真人發音那麼地流暢、自然。另外，在語音辨識系統裡，一種可能的語者調適(speaker adaptation)方法是，將某兩語者的音素(phoneme)頻譜包絡作內插，用以逼近目前使用者的音素頻譜包絡。

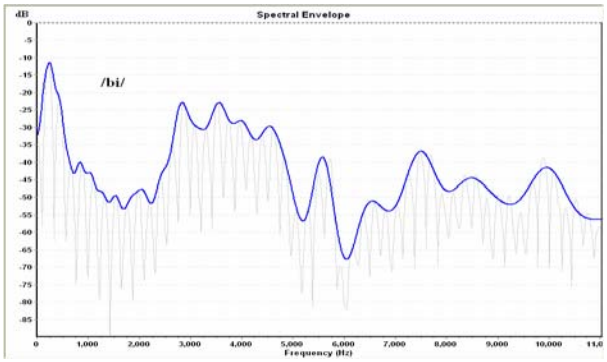
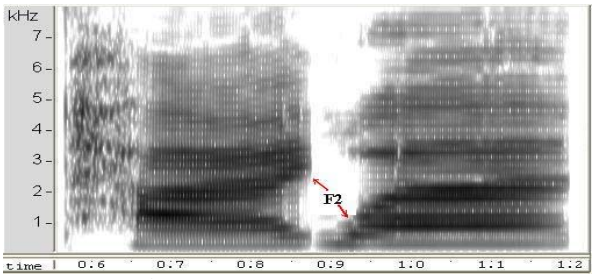


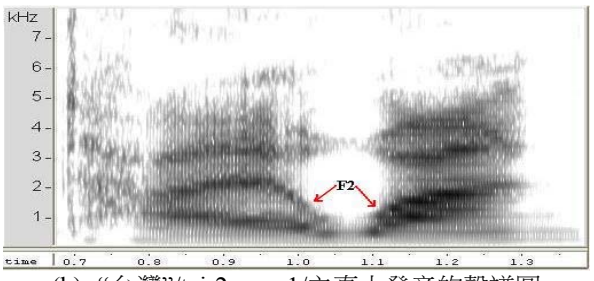
圖 1 /bi/發音的一個音框的頻譜包絡

關於語音信號的一個頻譜不連續的例子，如圖 2(a) 裡的合成語音 /tai-2 wan-1/(台灣) 的聲譜圖 (spectrogram)所顯示的情況，在音節/tai-2/的尾端，第二共振峰(formant) F2 的軌跡是逐漸地爬到高處，然而在音節/wan-1/的起點附近，其第二共振峰 F2 的軌跡則是逐漸地由低處往上爬，也就是在此二音節的邊界處，F2 的頻率值顯現一個劇烈改變的情形(突然由高處降到低處)，這就是頻譜的不連續；相對地，在圖 2(b)裡的真人發音/tai-2 wan-1/的聲譜圖裡，在音節/tai-2/的尾端，第二共振峰 F2 的軌跡會逐漸地往下走，以趨近音節/wan-1/的 F2 在音節起始時的較低的頻率值，而使得此二音節的 F2 在音節邊界處顯現出平順的銜接情形。

在作頻譜包絡的內插之前，參與的語音音框的頻譜包絡必需先被估計出來，關於頻譜包絡的估計，過去已有一些方法被提出[2, 3]，例如以 LPC (linear prediction coding) 全極 (all pole) 模型之頻率響應 (frequency response)來逼近頻譜包絡[4]；在 DFT 頻譜上以平滑化的倒頻譜 (cepstrum) 曲線作迭代改進之”True-Envelope”估計法[3, 5]；消除信號本身週期性之干擾、而具有極高準確性之 STRAIGHT 估計法[6]；以及我們最近嘗試作改進之離散倒頻譜(discrete cepstrum, DCep)估計法[7, 8]。雖然 STRAIGHT 估計法的準確性很高，而 True-Envelope 估計法的準確性也很好，但是兩者需求的計算量很大，目前要以個人電腦的計算能力來達成即時處理，是非常困難的，因此我們仍然決定採用 DCep(離散倒頻譜)之估計法。由於本篇論文主要想探討的是頻譜包絡的內插方法，所以頻譜包絡估計方法的相關議題，就不在多作討論了。



(a) “台灣”/tai-2 wan-1/之合成語音的聲譜圖



(b) “台灣”/tai-2 wan-1/之真人發音的聲譜圖

圖 2 頻譜連續與不連續的例子

2. 內插的方法、領域、與效果

談到內插的方法，一般人直覺上會想到的是線性內插 (linear interpolation)，如果使用線性內插就能達成所要求的效果，那麼就不必把事情複雜化。因此對於頻譜包絡內插的問題，首先考慮我們究竟要達成什麼效果? 其實從聲響音素學(acoustic phonetics)的知識[9]可知道，當連續發音兩種母音音素時，如連續發/a/與/i/，這兩個音素之間會有共發聲(coarticulation)的現象，而共發聲現象表現在聲譜(spectrogram)上，就是兩音素的共振峰軌跡(formant trajectory)在音素的邊界附近，會以平滑移動的方式相互靠近，一個例子如圖 3 所示，所以頻譜包絡內插要達成的效果就是，當逐漸改變內插比例值時，所得到的一序列頻譜包絡能夠讓共振峰頻率逐漸改變，而形成平滑移動之共振峰軌跡。在確定所需求的效果之後，接著就可考慮執行內插的領域(domain)、及內插的方法。

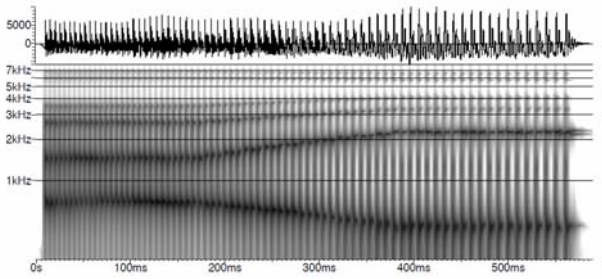


圖 3 共振峰軌跡以平滑移動方式相互靠近[10]

執行內插的領域可以選擇在頻域，即直接使用頻譜包絡，或是選擇在其它頻譜係數的領域，例如: LPC 係數、LSF(line spectrum frequency) 係數、反射 (reflection)係數，DFT 倒頻譜係數、DCep 係數、…等等，至於選擇那一種係數較好? 應採取那一種內插方法來作搭配? 則不是立刻可以推想得知。不過，如果我們直接對兩條頻譜包絡曲線作線性內插，則可立刻推知其結果是，共振峰軌跡會以平行方式作淡進淡出 (cross fading)，一個例子如圖 4 所示。此外，由前人的論文也已經知道[10]，若選擇 LSF 係數來作線性內插，則會發生某些共振峰軌跡可以平滑地移轉、連接，而某些共振峰軌跡卻會中斷的不一致情形，例子如圖 5 裡經 LSF 係數內插/a/音框與/i/音框所得到的聲譜圖。

過去已經有人研究頻譜包絡內插的方法，並且可以達成前面所說的效果，Pfitzinger 的 DFW (dynamic frequency warping)為基礎的方法[10]、和 Ezzat 等人的聲訊流(audio flow)為基礎的方法[11]，都是直接在頻譜包絡上進行內插，但是使用了不同的內插方法。DFW 的觀念是，透過動態頻率校正來建立兩條包絡微分曲線之間的頻率軸對應關係；聲訊流的觀念，則是從 optical flow 的觀念衍生而來，應用時也是在包絡微

分曲線上進行聲訊流的計算。如果說還有另外一種頻譜包絡的內插方法，我們覺得基於統計模型、及最大似然(maximum likelihood)原則的方法也算是一種，例如最近常被使用於作語音合成之 HMM (hidden Markov model)模型及其音框頻譜係數的求解方法[12]，但是統計模型之參數需要收集大量的語料來作估計，這是它的缺點。

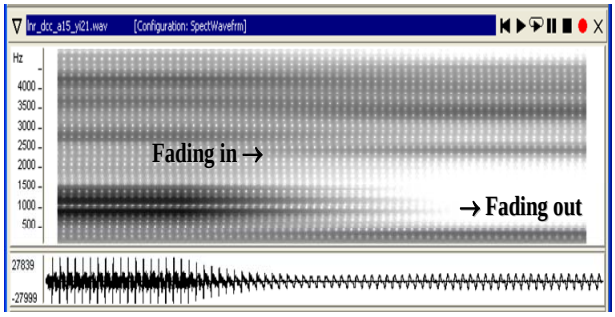


圖 4 共振峰軌跡淡進淡出之例子

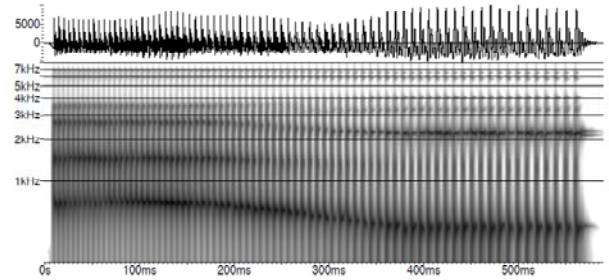


圖 5 共振峰軌跡平滑連接與中斷並存之聲譜圖[10]

3. 反射係數 與 頻譜包絡內插

尋找一種恰當的頻譜參數，再對它作簡單的線性內插處理，即可達成前述的頻譜包絡內插效果，一直是我們想要達成的目標。經過對幾種頻譜係數作實驗驗證，我們最後發現反射係數就是所要找的頻譜係數。反射係數具有的物理意義是，當把聲道(vocal track)模式化(modeling)成多節管(multi-tube)之聲響(acoustic)模型，如圖 6 所畫的，則第 k 個個反射係數 C_k的意義就是，第 k 與 k+1 節管子的截面積 B_k與 B_{k+1}的差值與和值的比值，即

$$C_k = \frac{(B_{k+1} - B_k)}{(B_{k+1} + B_k)} \quad . \tag{1}$$

從反射係數所具有的物理意義，我們認為這可用以解釋，為什麼對反射係數作簡單的線性內插，就可以達成所想要的頻譜包絡內插之效果。

3.1 頻譜包絡 至 反射係數之轉換

當從一個音框估計得到一條頻譜包絡曲線 S_i, i=0, 1, …,

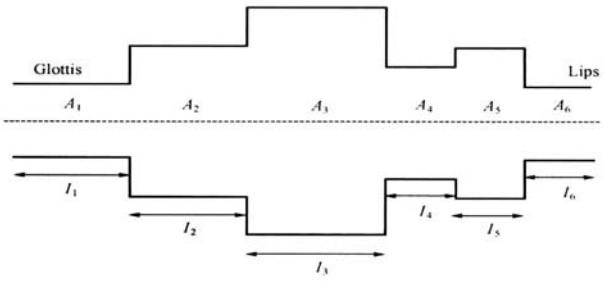


圖 6 聲道之多節管模型[13]

N-1, N 為 DFT 轉換的點數，本文中設 N=512，接著我們可對 S_i 取平方而得到功率頻譜(power spectrum)，然後作 N 點的離散反傅利葉轉換(inverse DFT)，以求得自相關(autocorrelation)係數 R_n [14, 15]，離散反傅利葉轉換的公式為

$$R_n = \frac{1}{N}\sum_{k=0}^{N-1}|S_k|^2 \cdot e^{j\frac{2\pi}{N}kn}, \quad n = 0,1,\cdots,N-1, \tag{2}$$

實作上則是以反快速傅利葉轉換(inverse FFT)來計算公式(2)。依據自相關係數 R_n，接下來可再使用 Levinson-Durbin (LD)演算法[9, 13, 14]來求取 p 階的 LPC 係數 α_k 與反射係數 C_k, k=1, 2, …, p，LD 演算法的詳細步驟如下:

步驟(a) 令 $E^{(0)} = R_0$, $m = 1$ (3)

步驟(b) $$C_m = \frac{1}{E^{(m-1)}}\left(R_m - \sum_{i=1}^{m-1}\alpha_i^{(m-1)} \cdot R_{m-i}\right) \tag{4}$$

步驟(c) $$\alpha_m^{(m)} = C_m \tag{5}$$

步驟(d) $$\alpha_i^{(m)} = \alpha_i^{(m-1)} - C_m \cdot \alpha_{m-i}^{(m-1)}, \quad i = 1,...,m-1 \tag{6}$$

步驟(e) $$E^{(m)} = E^{(m-1)}\left(1 - C_m \cdot C_m\right) \tag{7}$$

步驟(f) 如果 $m < p$，則令 $m = m$+1，再回到步驟(b)

最後令 $C_0 = R_0$，以保留能量資訊。

3.2 反射係數 至 頻譜包絡之轉換

當以線性內插(或其它內插方式)得到一組新的反射係數 C_i 之後，接著可執行前述的 LD 演算法，不過步驟(b)要略過，因為反射係數是已知的，如此就可求得 p 階的 LPC 係數$\alpha_i^{(p)}$，i=1,2, …, p。依據所求得的 LPC 係數，就可用以建立一個全極模型，然後計算此全極模型的頻率響應，就可求得對應的頻譜包絡，也就是藉由全極模型來定義一組反射係數對應的頻率響應，全極模型的正規化(normalized)頻率響應的計算公式為

$$Q_k = 1 \Big/ \left|1 - \sum_{n=1}^{p}\alpha_n^{(p)} \cdot e^{-j\frac{2\pi}{N}kn}\right| \quad , \; k = 0,1,\cdots,N-1 \tag{8}$$

理論上只要階數夠高(p 夠大)，頻譜包絡裡的零點(zero)效應仍可被極點(pole)所逼近[9]。

當使用公式(8)計算出頻譜包絡 Q_k 之後，會發現它和真實的頻譜包絡之間有能量上的偏移，因此需要對 Q_k 乘上一個能量調整值 u，$u = (E_{org} / E_{lpc})^{0.5}$，其中 E_{org} 表示原始頻譜包絡的能量(可由 C_0 取得)，E_{lpc} 則表示 LPC 頻譜包絡 Q_k 的能量。

4. 語音信號再合成之方法

令 SA_k, SB_k 表示兩條給定的頻譜包絡曲線，當對它們轉換出的反射係數作內插，我們可得到一序列時間點上的頻譜包絡曲線 $S_k^{(\tau)}$，k = 0, 1, …, N-1，τ = 0, 1, …, 40， τ表示時間軸上的第τ個內插出的音框。在此，我們藉由 cos 函數來達成曲線式的內插，也就是令

$$CT(S_k^{(\tau)}) = CT(SA_k)\cdot\tfrac{1}{2}[\cos(\pi\cdot\tfrac{\tau}{40})+1] + CT(SB_k)\cdot\tfrac{1}{2}[\cos(\pi\cdot\tfrac{\tau-40}{40})+1], \quad \tau=0, 1, ..., 40. \tag{9}$$

其中 $CT(\cdot)$表示頻譜包絡至反射係數之轉換。在公式(9)裡，所以採取曲線式內插而不用線性內插，是因為我們希望在邊界點上(即 k=0 或 k=40)，斜率是從 0 開始變化。

對於內插出的頻譜包絡序列 $S_k^{(\tau)}$，我們可藉由所合成出的語音信號，來讓人耳以聽覺判斷此頻譜包絡序列的流暢性，例如來自/a/音音框的 SA_k 與來自/i/音音框的 SB_k 經過內插後，再合成出的語音信號，是否會有雙母音/ai/的感受。此外，合成的語音信號作聲譜分析後，也可用以觀察共振峰軌跡的走勢。在此，我們決定採用 HNM (harmonic-plus-noise model)信號模型[16]，並且依據頻譜包絡來決定各個諧波(harmonic)與雜音(noise)的振幅，然後再據以產生出語音信號。

令 $F_0^{(\tau)}$表示分派給第τ個音框的基本頻率，則此音框第 n 個諧波的頻率就是 $n\cdot F_0^{(\tau)}$，而它的振幅則可依頻譜包絡曲線 $S_k^{(\tau)}$去決定，一個簡單的作法是，找出頻譜包絡曲線之頻率軸上最靠近 $n\cdot F_0^{(\tau)}$的相鄰兩點上的振幅值去作線性內插。另外，我們直接設定 MVF (maximum voiced frequency)為 6,500Hz，並且設定兩音框之間的間隔為 110 個樣本點，而取樣率則是 22,050。為了方便說明，令第 i 個音框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 個音框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 的諧波信號之樣本 $h(t)$時，我們先以如下公式作線性內插，

$$f(k,t) = f_k^i + \frac{t}{110}(f_k^{i+1} - f_k^i), \quad k = 1,2,...,L \tag{10a}$$

$$a(k,t) = a_k^i + \frac{t}{110}(a_k^{i+1} - a_k^i), \quad k = 1,2,...,L \tag{10b}$$

以求取時刻 t 時各諧波的頻率與振幅，其中 110 表示相鄰音框之間的樣本點數，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} 設為零值。然後，以如下公式計算 $h(t)$，

$$h(t) = \sum_{k=1}^{L} a(k,t)\cdot\cos\big(\phi(k,t)\big), \quad 0 \le t < 110 \tag{11}$$

$$\phi(k,t) = \phi(k,t-1) + 2\pi\cdot f(k,t)/22{,}050$$

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 $\phi(k,-1)$ ，我們令其等於前一音框裡的 $\phi(k,99)$以便維持相位的連續性，而當音框編號 i 為 0 時就以亂數來設定。

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[16]，就是把雜音當作是 MVF 之後頻率間隔固定為 100Hz、但振幅會隨時間改變之一些弦波的加總。先依 MVF 決定頻率 index 之下限 KL=MVF / 100, 而其上限明顯地是 KU=11,025 / 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，

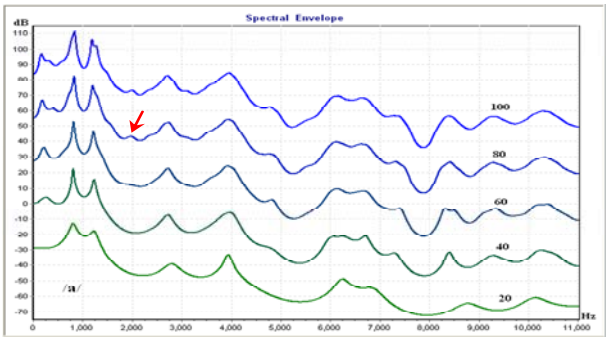
$$g(t) = \sum_{k=KL}^{KU} b(k,t)\cdot\cos\big(\psi(k,t)\big), \quad 0 \le t < 110 \tag{12}$$

$$\psi(k,t) = \psi(k,t-1) + 2\pi\cdot k\cdot 100/22{,}050$$

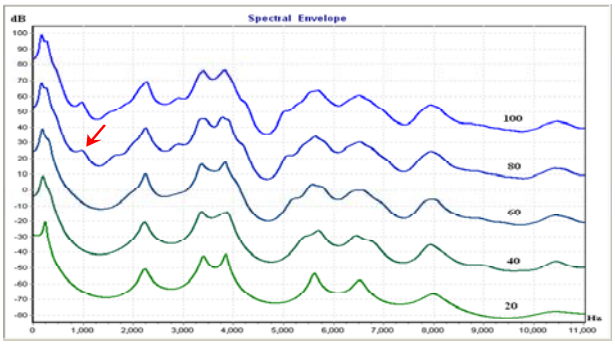
其中 $b(k,t)$ 表示時刻 t 時第 k 個弦波的振幅，其值也是以類似公式(10)之線性內插來求得，$\psi(k,t)$ 表示第 k 個弦波累積到時刻 t 時的相位量，其初始值也是以亂數來設定。最後，將 $h(t)$與 $g(t)$相加，即可得到時刻 t 的合成信號樣本。

5. 歌聲頻譜內插與聽測實驗

前面 3.2 節提到，以內插得到的反射係數去建造對應的全極模型，再以全極模型的頻率響應作為導出的頻譜包絡。那麼全極模型的階數應設多大? 在此我們以實驗的方式來觀察不同階數對頻譜包絡曲線的影響，我們發現階數要到 80 時，來自/a/和/i/的兩個音框的頻譜包絡才會足夠細緻，如圖 7(a)與 7(b)所示，因此我們決定使用階數為 80 之全極模型。



(a) /a/音框的 5 條頻譜包絡



(b) /i/音框的 5 條頻譜包絡

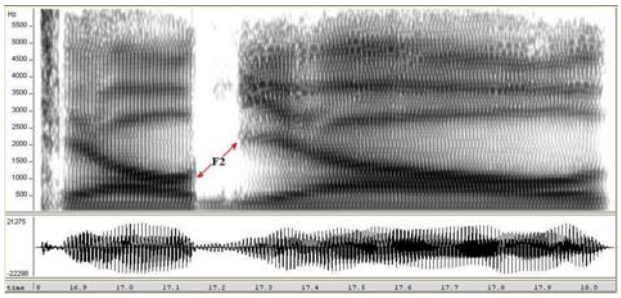
圖 7 全極模型階數對頻譜包絡曲線的影響

5.1 歌聲頻譜內插

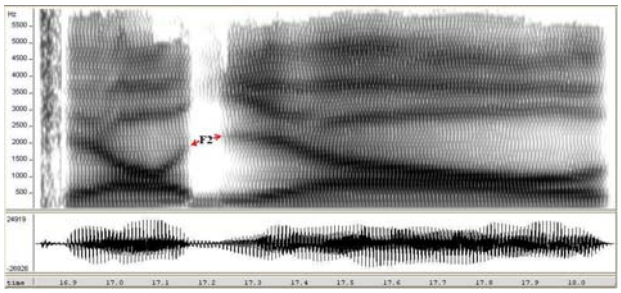
過去我們已經對國語歌聲的合成，作了一些研究[17, 18]。不過，先前我們都以分離(isolated)方式錄製國語的 408 個不同音節，然後對各個音節去分析所屬各音框的 HNM 參數。之後，再依據一個音節各音框分析出的 HNM 參數，去合成出一個歌詞音節的歌聲信號，並且把一首歌曲的合成，看作是一序列歌聲音節的信號的串接。

由於使用了前述的作法(把獨立合成出的歌聲音節拿來串接)，因此我們遇到了一個困難，那就是在一些會發生共發聲的音節邊界上，所合成的歌聲信號出現了頻譜的不連續，而使得合成的歌聲讓人聽起來覺得是顆顆粒粒地(相連音節是獨立無關聯地被串接)，不像真人所唱的歌聲那樣地流暢。主要會發生共發聲的情況是，前一音節的韻母以母音音素結束，並且後一音節的開頭沒有聲母，例如"跳躍"兩音節之間是/u/接/i/，它的頻譜不連續連接的情形可由圖 8(a)的聲譜圖看出，至於"太陽"的兩音節之間，雖也是母音接母音，但兩者是同樣的母音/i/，所以頻譜可以連續地連接。

為了解決前述因為共發聲而造成的頻譜不連續之問題，我們便考慮應用前述的基於反射係數之頻譜內插方法，來嘗試把不連續的頻譜轉變成連續的頻譜。首先，檢查兩兩相鄰之歌詞音節之間是否會發生共發聲的情況，如果不會發生共發聲，就可以一般的串接方式來處理，如果會發生共發聲，那麼就從前一音節最後 100ms 的附近找一個音框作為參考音框，且以該音框的頻譜包絡作為公式(9)裡的 SA_k，類似地也從後一音節前面 100ms 的附近找一個音框作為參考音框，且以該音框的頻譜包絡作為公式(9)裡的 SB_k，然後我們就可以使用公式(9)，來內插出兩音節邊界轉移區內的 40 個音框的頻譜包絡，在此相鄰音框之間的時間間隔設為 5ms。之後，我們就可把內插出的頻譜包絡，帶入公式(10)、(11)與(12)，來把一序列的信號樣本值計算出來。



(a) 頻譜不連續連接



(b) 頻譜連續連接

圖 8 "跳躍"/tiau iau/兩音節之間的頻譜連接情形

以前面提到的歌詞"跳躍"/tiau iau/為例，當使用前一段所說的方法來作頻譜內插及信號樣本值計算，則圖 8(a)的聲譜圖就可以轉變成如圖 8(b)所示的聲譜圖。比較圖 8(a)與圖 8(b)可發現，圖 8(b)裡/tiau/音節的尾端，F2 頻率值已明顯被提高了，而使得兩音節的 F2 軌跡變得比較平順地銜接。另外，以"青春舞曲"這首歌中的歌詞"一去無影蹤"為例，若歌詞"去無影"的音節各自去合成出信號後再串接起來，則其聲譜圖會如圖(9)所示，由圖(9)可發現兩兩音節之間(即/cyu/與/u/之間，/u/與/ing/之間)，相鄰音節的共振峰軌跡是各走各的；但是，當使用本論文提出的頻譜內插方法之後，圖(9)的聲譜圖就會轉變成如圖(10)所示的聲譜圖，圖(10)中兩兩音節之間，共振峰軌跡很明顯地表現出相互趨近的情形，而使得共振峰軌跡變成比較平順地銜接。

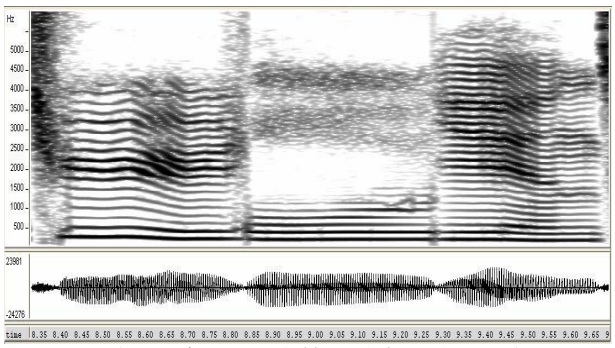


圖 9 歌詞"去無影"的音節信號各自合成之聲譜圖

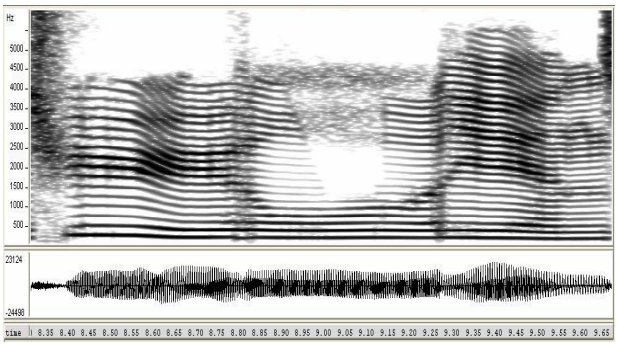


圖 10 音節邊界先經頻譜內插再合成信號之聲譜圖

5.2 聽覺測試

從 5.1 節的說明，可知本論文的頻譜內插方法，的確能夠把音節邊界不連續的頻譜，經由頻譜內插和信號再合成，而轉變成平順銜接的頻譜。不過，本論文所得到的一種共振峰平順轉移的方式，是否就能夠讓合成出的歌聲聽起來變得比較流暢? 為了驗證此一疑問，我們便決定進行主觀的聽測實驗。首先我們準備了四首合成的歌聲檔案，在此分別以 AO, AS, BO, BS 為代號，AO 與 BO 的"O"表示使用原始的合成方法，即直接串接而不作頻譜內插，AS 與 BS 的"S"則表示使用本論文的頻譜內插方法；另外，代號開頭的"A"與"B"，在此以"A"表示"青春舞曲"這首歌，而以"B"表示"噢!蘇珊娜"這首歌，這四首歌的音檔可從網頁 http://guhy.csie.ntust.edu.tw/vibrato/SingIntrP.html 去下載試聽。當進行聽測時，我們先連續播放兩首歌 X 與 Y，然後請受測者分辨 Y 的流暢性是否比 X 的好，受測者可要求再重複聽一次，至於 X 與 Y，在此是以隨機方式指派 AO 與 AS 至 X 與 Y，或者隨機指派 BO 與 BS 至 X 與 Y。

我們共邀請了 15 位受測者，其中 12 位並未從事歌聲合成之研究。流暢性的評分方式是，當 Y 比 X 明顯流暢(或明顯不流暢)時，給 2 分(或-2 分)，當 Y 比 X 稍微流暢(或稍微不流暢)時，給 1 分(或-1 分)，而當 Y 和 X 的流暢性分辨不出時，則給 0 分。進行聽測之後，我們將 15 位受測者所給的評分作整理和計算，結果得到如表 1 所示的平均分數和標準差。從表 1 兩首歌的平均分數 0.6 和 1.0 可知，使用本論文的頻譜內插方法來對相鄰的共發聲音節之間作頻譜的平滑化處理，的確可讓合成的歌聲信號變得比較流暢；至於歌曲 A(噢!蘇珊娜)的平均分數 1.0 比歌曲 B(青春舞曲)的平均分數 0.6 來得高一些，我們認為它的原因是，歌曲 A 的拍速(tempo)比歌曲 B 的快，而拍速快的歌曲所合成出的歌聲信號(即 AS 與 AO)，比較難以分辨流暢性的差異。

表 1 聽覺測試之平均分數與標準差

	AS vs AO (青春舞曲)	BS vs BO (噢!蘇珊娜)
平均分數	0.600	1.000
標準差	0.611	0.632

6. 結語

關於頻譜包絡的內插，若要達成共振峰軌跡的平滑移動，並不是各種的頻譜參數經由內插都能做到，然而透過反射係數來作內插，我們由歌聲合成的實驗顯示，的確可讓共振峰以平滑的軌跡來移動。至於反射係數與頻譜包絡之間的相互轉換方法，本論文提出了一個實際可行的實施程序，並且已經由歌聲合成的實驗作了驗證。

將音節間原本不連續的共振峰軌跡，經由反射係數內插而合成出具有連續的共振峰軌跡的歌聲，然後使用合成的歌聲進行聽覺測試之實驗，結果聽測實驗得到的平均評分顯示，經反射係數內插所合成的歌聲，的確可提升合成歌聲的流暢性。

雖然經由公式(9)作反射係數的內插，可以將原本不連續的共振峰軌跡，修改成平滑移動的共振峰軌跡，但是如此得到的共振峰軌跡，並不一定就是真人歌聲中的共振峰所行走的軌跡，因此所合成的歌聲，其流暢性應是會比真人歌聲的差。所以，作反射係數內插的函數(如公式(9))，未來仍可進一步作探討。此外，未來我們也應再做更多的實驗，以比較本論文的頻譜包絡內插方法和他人提出的方法。

7. 參考文獻

[1] D. T. Chappell and J. H. L. Hansen, "A comparison of spectral smoothing methods for segment concatenation based speech synthesis," *Speech Communication*, vol. 36, pp. 343–374, 2002.

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

[3] A. Robel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," *Int. Conf. on Digital Audio Effects*, Madrid, Spain, pp. 1-6, 2005.

[4] J. D. Markel and A. H. Gray Jr., *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

[5] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method," *Electron. and Commun. in Japan*, vol. 62-A, no. 4, pp. 10-17, 1979. (in Japanese)

[6] H. Kawahara, I. Masuda-katsuse, and A. De Cheveign, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds", *Speech Communication*, vol. 27, pp. 187-207, 1999.

[7] O. Cappé and E. Moulines, “Regularization techniques for discrete cepstrum estimation,” *IEEE Signal Processing Letters*, vol. 3, no. 4, pp. 100–102, 1996.
[8] H. Y. Gu and S. F. Tsai, "A Discrete-cepstrum Based Spectrum-envelope Estimation Scheme and Its Example Application of Voice Transformation", *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 14, no. 4, pp. 363-382, 2009.
[9] D. O`Shaughnessy, *Speech Communication: Human and Machine*, 2nd ed., IEEE Press, Piscataway, NJ, 2000.
[10] H. R. Pfitzinger, “DFW-based spectral smoothing for concatenative speech synthesis,” *Int. Conf. Spoken Language Processing*, Jeju, Korea, pp. 1397-1400, 2004.
[11] T. Ezzat, E. Meyers, J. Glass, and T. Poggio, "Morphing spectral envelopes using audio flow," *INTERSPEECH 2005*, Lisbon, Portugal, pp. 2545-2548, 2005.
[12] K. Tokuda, H. Zen, and A. W. Black, “An HMM-based speech synthesis system applied to English,” *IEEE Workshop on Speech Synthesis*, Santa Monica, CA, pp. 227-230, 2002.
[13] 王小川，*語音訊號處理*(修訂二版)，全華圖書公司，台北，2009。
[14] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.
[15] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1999.
[16] 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.
[17] 古鴻炎、廖皇量，「用於國語歌聲合成之諧波加噪音模型的改進研究」，*國際電腦音樂與音訊技術研討會* (WOCMAT 2006)，台北，session 2: 音訊處理 I，(2006)。
[18] 古鴻炎、林正甫，「國語歌聲抖音參數之分析」，*國際電腦音樂與音訊技術研討會* (WOCMAT 2007)，新竹，Session III: Audio Signal Processing，(2007)。