基於 HNM 之國語音節信號的合成方法

An HNM Based Method for Synthesizing Mandarin Syllable Signal

古鴻炎　　周彥佐
Hung-yan Gu　and　Yen-zuo Zhou

國立臺灣科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
{guhy, m9315058}@mail.ntust.edu.tw

摘要

本文提出一個基於 HNM (Harmonic-plus-noise model) 的國語音節信號的合成方法，使用此方法時，一種音節只需錄、存一遍發音，就可用以合成多種韻律特性的發音，並且不易查覺出信號品質的衰退。在這個方法裡，一個欲合成的音節的音長，首先被分割成它的組成音素的音長，依據原始和合成音節裡各音素的音長，可建造一個片斷線性的時間對映函數，如此合成音節時間軸上的一個控制點，就可經由對映至原始音節上找出和它對應的兩個音框。然後依據兩音框的 HNM 參數作時間上的內差，再進一步在音色一致性的條件下作基週軌跡調整的內差，來求得該控制點上的 HNM 參數。當各個控制點上的 HNM 參數值都決定之後，就可使用我們重新公式化的 HNM 合成公式，來計算出各個信號樣本的值。接著我們作聽測實驗來評估合成語音的清晰度，初步結果顯示，本文所提的 HNM 擴充的方法所合成出的信號，非常清晰且無迴音，明顯地比先前提出的 TIPW 法的好很多。

關鍵詞：語音合成，諧波加雜音模型，音色一致性

Keywords: Speech Synthesis, Harmonic-plus-noise Model, Timbre Consistency.

一、前言

自從 PSOLA 合成法被提出之後[1]，它已廣泛地被應用於作語音信號的合成，不過使用 PSOLA 所合成出的語音信號的品質，並不穩定，例如對一個錄製的語音單元的基周軌跡(pitch contour)或音長(duration)作較大幅度的改變時，則合成語音的品質會衰退很多[2]。雖然在採取語料庫(corpus based)爲基礎的研究方向[3, 4]時，韻律特性通常不需要作大幅度的改變，並且 PSOLA 是一個容易製作、運算量很少的信號合成方法，如此情況下 PSOLA 可說是一個不錯的選擇。但是，基於語料庫之研究方向的前提條件是，語料要錄得夠多，否則合成出的語音信號，在某些音節之間會出現音調銜接得不平順的問題，並且某些語句內會出現說話速度忽快忽慢的問題，那麼就仍然需要作較大幅度的韻律特性的調整了。此外我們考慮到，所研究的語音合成技術希望可以很經濟地(節省人力、時間)移轉到其它語言(如閩南語、客語)去使用，因此我們傾向於採取信號模型(signal model)爲基礎的研究方向，而不願意採取語料庫爲基礎的研究方向。

由於國語是一種有聲調的語言，且國語聲調之間的差異，主要表現於音節基週軌跡的高

度和形狀。因此當採取信號模型爲基礎的研究方向時，我們需要對一個合成單元的基週軌跡的高度及曲線形狀作大幅度的改變，如此 PSOLA 就不適合使用了，而必需另外尋求或研發適合的信號合成技術。最近我們發現 HNM 是一個不錯的基礎，可對它作改進而用來合成國語語音的信號。此外我們覺得語料庫爲基礎的研究方向裡，若語料錄得不夠多，則也可以考慮以 HNM 來取代 PSOLA。

HNM 是由 Y. Stylianou 所提出的一種語音信號的模型[5，6]，希望在作語音處理(編碼，合成)時，仍能保持信號的清晰度與自然度。HNM 可看成是弦波模型[7]的改進，它對於語音高頻部分的雜音(noise)信號成分，建立了較好的模型。HNM 的模型參數分析程序裡，提供了最大有聲頻率 MVF(maximum voiced frequency) 的一個偵測方法，依 MVF 值可將一個語音音框(frame)的頻譜(spectrum)分割成低頻、高頻之兩個部分，對於低頻部分的信號成分，採取以諧波成分(harmonic partials)的加總來模式化(modeling)，而對於高頻部分的信號成分，則採取以平滑的頻譜包絡(spectral envelope)來模式化，實際上是以少數的倒頻譜(cepstrum)係數來代表此頻譜包絡。

當應用 HNM 來合成國語語音時，我們發現有幾個議題，其解決方法並未能在 HNM 的文獻上找到，第一個議題是，如何讓合成的音節信號，保持音色(timbre)的一致性(consistency)，即音色一致性議題。由於我們只希望對各種國語音節錄製一次發音，然後透過修改一個音節的基週軌跡的高度及形狀，來合成出其它聲調的音節信號，因此當一個欲合成音節的基週軌跡被指定時，我們必需使用一種適當的方法來調整 HNM 各諧波成分的參數值，以同時滿足基週軌跡及音色一致性的要求。第二個議題是，對於一個放置於欲合成音節之時間軸上的一個控制點(control point) [8, 9]，如何決定此控制點上的 HNM 參數的數值，即參數值設定之議題。在合成一個音節的信號時，我們需要調整該音節原始錄音的音長以滿足韻律單元所指派的合成音節之音長，因此在合成音節時間軸上的一個控制點，當它被對映(mapping)至一個位於原始音節兩分析音框之間的時間點時，我們必需使用一種適當的內差方法來計算此控制點上的 HNM 參數值。此外，第三個議題是，如何校正(warp)合成音節的時間軸，以合成出較爲流暢(fluent)的音節及語句的信號，也就是時間軸校正的議題，此議題應是和語音合成較爲相關，而和 HNM 的相關性較少，當一個合成音節的音長需要伸長或縮短時，一個簡單的時間軸校正方法是線性校正(linear warping)，但此種作法通常會得到較差的流暢度。

本文研究了前述三項議題的解決方法，然後再據以建造出一個 HNM 改進、擴充的國語音節信號合成系統，此系統的主要處理流程如圖 1 所示。當要合成一個音節的信號時，很明顯地此音節的各個韻律參數值已經由韻律單元訂定、指派好了，因此圖 1 裡的第一個方塊首先作的是，將合成音節的音長規劃、分割成此音節的組成音素(phoneme)的時長(duration)，接著依據相連音素的時長來建造一個片斷線性(piece-wise linear)的時間校正函數，以便將合成音時間軸上的時間點對映至原始音的時間軸上；在圖 1 裡的第二個方塊，先均勻地在合成音的時間軸上佈放控制點，然後對各個控制點求取該點上的 HNM 參數值；接著在圖 1 裡的後面三個方塊，將信號分類成三種形態分別去作合成處理，對於短時間的無聲(unvoiced)聲母(syllable initial)，其信號片斷直接由原始音裡複製到合成音裡，對於長時間的無聲聲母，其信號則當作是 HNM 的雜音信號成分來作合成，至於音節的有聲(voiced)部分，包括有聲子音及母音，則先分別合成出諧波和雜音成分，再作相加。

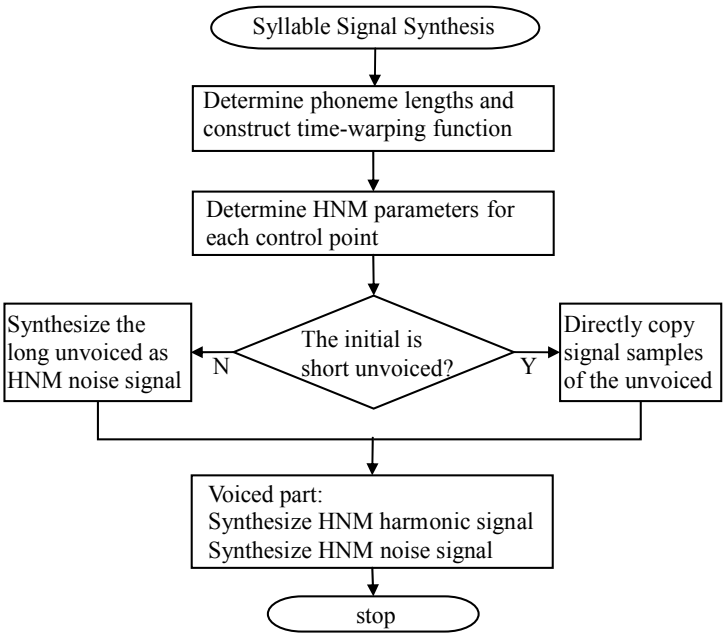

圖 1、基於 HNM 之音節信號合成方法的主要處理流程

二、音素時長規劃及時間校正函數

國語音節的結構可以看成是 C_xVC_n，其中 C_x 可以是有聲子音、無聲子音、或是空音(null)，而 C_n 可以是鼻音/n/或/ng/、或是空音，V 則可以單母音、雙母音、或是三母音。當 C_x 是一個無聲子音時，它可再被分類成短無聲(如/b/)或長無聲(如/p/)，然後進行圖 1 菱形塊右邊或左邊方塊的處理；此外當 C_x 是一個有聲子音(如/m/)或空音時，則我們把 C_x 或 V(當 C_x 是 null 時)的起始部分的信號當作是短無聲來看待及作處理(也就是走到圖 1 菱形塊右邊的方塊去)，這樣的對待方式是合理的，因爲以有聲音素開頭的音節，其起始部分的一、二個信號音框，在作 HNM 的參數分析時，也經常被判斷成非週期性的。

當一個音節是以廣義的短無聲子音開頭時，如/bau/及/man/，則此子音的時長規劃，就是直接依據原始錄音裡此子音的長度。相反地，如果音節是以長無聲子音開頭時，則此子音的時長規劃是，將原始錄音裡此子音的長度乘上一個比例值 *Fu*，*Fu* 的值基本上設定成合成音節的音長除以原始音節的音長，不過當 *Fu* 大於 1.4 時就改設爲 1.4，而當 *Fu* 小於 0.6 時就改設爲 0.6，這是因爲音節音長的伸長或縮短，主要是在母音部分進行，而非等比例的方式。在規劃了音節起始的無聲子音的時長 *Du* 之後，音節後面有聲部分的長度 *Dv*，很明顯地就是音節音長減去 *Du*。

接著，考慮音節有聲部分裡各音素的時長規劃，這裡以音節/man/爲例來說明，令/man/的原始錄音裡，音素/m/、/a/、/n/分別佔據的長度是 *Rm*、*Ra* 和 *Rn* 毫秒(ms)，且有聲部分的總長度是 *Rv* = *Rm* + *Ra* + *Rn*，另外令合成音節裡，這三個對應的音素的時長值分別是，*Dm*、*Da*、*Dn*，且令 *Dv* = *Dm* + *Da* + *Dn*，則我們規劃 *Dm*、*Da*、*Dn* 數值的作法就如下列的程序：

```
r = 0.6;
while ( r >= 0.1 ) {
   Dm = (Rm/Rv) * r * Dv;
   Dn = (Rn/Rv) * r * Dv;
```

```
    Da = Dv - Dm - Dn;
    if (Da > Dv*0.4) break;
    r = r - 0.05;
}
Db = Dm + Dn;
if (Dm > 0 && Dm/Db < 0.35) { Dm = 0.35*Db;  Dn=Db-Dm; }
if (Dn > 0 && Dn/Db < 0.35) { Dn = 0.35*Db;  Dm=Db-Dn; }
```

如果一個音節的結構是和/san/或/an/一樣的，也就是缺少有聲開頭的子音，則上列程序裡 *Rm* 和 *Dm* 的值可直接設爲 0；相同地如果音節的結構是和/ma/或/sa/一樣，即沒有結尾的鼻音，則上列程序裡 *Rn* 和 *Dn* 的值也可直接設爲 0。當 *Dm*、*Da*、*Dn* 的值設定好之後，就可以將合成音裡的有聲音素依序對應至原始音裡的有聲音素，而建造出如圖 2 所示的片斷線性之時間校正函數。

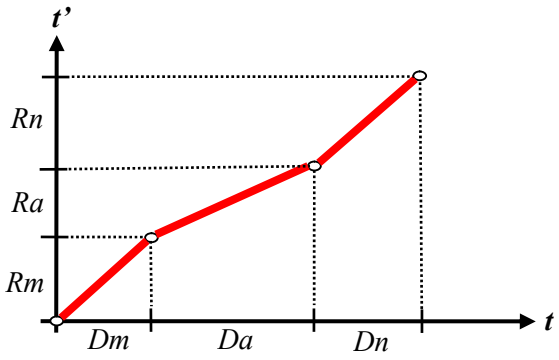

圖 2、片段線性之時間校正函數

上面的音素時長之規劃程序，其演算法是基於一個觀察，即有聲子音長度對於母音長度的比值 (*Rm* + *Rn*) / *Rv*，在句子裡發音節/man/的，會比單獨/man/音節發音的小許多。目前我們僅以簡單的程序來模擬此現象，並且時間校正函數也簡單地設定爲如圖 2 的片段線性之型式，未來我們將採取另一種較爲系統化的方法，來研究合成音節和原始音節之間的時間對映問題。不過，我們認爲即使只用簡單的時間校正函數，仍然可讓流暢度獲得明顯的改進。

三、控制點及其 HNM 參數的設定

3.1 控制點之佈放

我們錄製國語音節信號的取樣率是 22,050Hz，對於信號作 HNM 的參數分析時，設定音框的長度爲 512 個樣本點(23.2ms)，且音框每次前進 256 個樣本點。另一方面在作音節信號的合成時，我們採取了電腦音樂合成裡常用的觀念--控制點[8, 9]，這裡分別使用〝音框〞和〝控制點〞兩名詞，藉以指出在合成音節有聲部分的時間軸上所佈放的控制點，它上面的 HNM 參數，並不是從某一音框所分析出的 HNM 參數直接複製過來，而是拿兩個對應音框的 HNM 參數來作內差而求得的，不過在合成長時間的無聲子音時，我們就只有簡單地把一個音框分析出的 HNM 參數指派給一個對應的控制點。這兩種 HNM 參數的設定方式，以分別對付有聲部分和長時間無聲部分，就如圖 3 裡的圖形所表示的。

從圖 3 也可看出，在合成音的無聲子音部分，所佈放的控制點的數量，其實就是原始音裡無聲子音部分的音框的數量，因此合成音的無聲部分的時長調整，只是簡單地以線性

方式作伸長或縮短。然而在合成音的有聲部分，相鄰的控制點永遠是間隔 100 個信號樣本(4.5ms) ，因此控制點的數量是由所指派的有聲部分之時長來決定，至於選擇以 100 個樣本(而不用較大的數值)作間隔，是因爲我們希望對頻譜的演進(progressing)作較精細的控制。

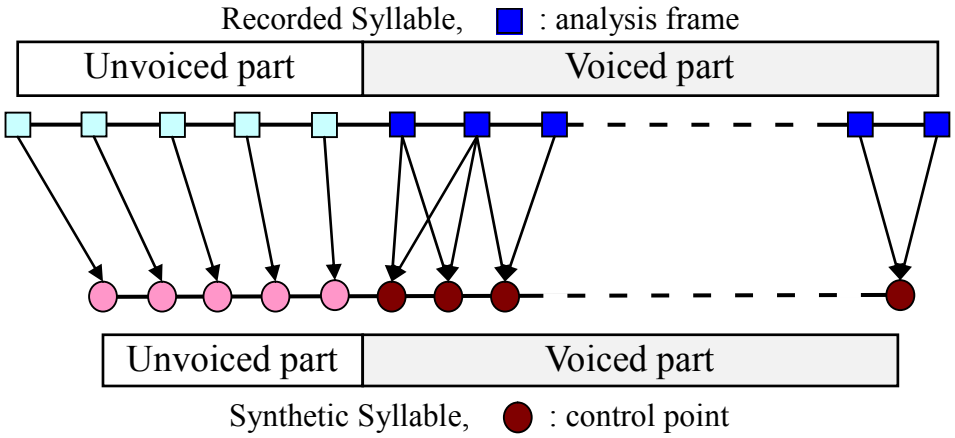

圖 3、控制點至分析音框之對映

3.2 音高(pitch)未變之 HNM 參數

要爲一個位於合成音有聲部分的控制點決定它的 HNM 參數值，第一步是依據如圖 2 所示的時間校正函數，將此控制點所在的時間位置 *ts*，對映至原始音時間軸上一個以音框爲單位的時間位置 *tr*；然後依據第⌊ *tr* ⌋和⌊ *tr* ⌋+1 音框兩者所分析出的 HNM 參數來作內差，以求得此控制點上的 HNM 參數，在此我們提出以線性內差的方式來求取各諧波的振幅、頻率、相位等三項參數，詳細的公式爲

$$\begin{aligned} \overline{A}_i &= (1-w)\cdot A_i^n + w\cdot A_i^{n+1} \\ n &= \lfloor t_r \rfloor, \quad w = t_r - n \end{aligned} \tag{1}$$

$$\overline{F}_i = (1-w)\cdot F_i^n + w\cdot F_i^{n+1} \tag{2}$$

$$\overline{\theta}_i = w\cdot(\hat{\theta}_i^{n+1} - \theta_i^n) + \theta_i^n \tag{3}$$

其中 A_i^n、F_i^n、θ_i^n 分別代表第 *n* 個分析音框裡第 *i* 個諧波的振幅、頻率與相位，而 $\overline{A}_i$、$\overline{F}_i$、$\overline{\theta}_i$ 則分別代表此控制點上第 *i* 個諧波的振幅、頻率與相位；此外，$\hat{\theta}_i^{n+1}$ 表示 θ_i^{n+1} 對於 θ_i^n 作反包裹(unwrap)運算後的相位值，也就是 $\hat{\theta}_i^{n+1} = puw(\theta_i^{n+1}, \theta_i^n)$ ，相位 θ_i^{n+1} 必需作反包裹運算以保證相位差值會落於-π到π之間，在此我們修正過的相位反包裹運算的作法是

$$\hat{\theta}_i^{n+1} = puw(\theta_i^{n+1}, \theta_i^n) = \theta_i^{n+1} - M\cdot 2\pi \tag{4}$$

$$M = \left\lfloor \frac{1}{2\pi}\left(\theta_i^{n+1} - \theta_i^n + \theta_c\right) \right\rfloor, \quad \theta_c = \begin{cases} \pi, & \text{if } \theta_i^{n+1} \geq \theta_i^n \\ -\pi, & \text{otherwise} \end{cases}$$

另外，由於一個音框作 HNM 參數分析後，會得到 10 個代表雜音成分的倒頻譜係數，因此，我們就拿兩個被對映到的相鄰音框所分析出的倒頻譜係數來作線性內差，而求得此控制點上的 10 個倒頻譜係數。

3.3 音高改變之 HNM 參數

在一個控制點上計算出原始音高(pitch)的諧波參數 $\overline{A}_i$、$\overline{F}_i$、$\overline{\theta}_i$ 之後，下一步要作的是，計算音高改變後的諧波參數 $\widetilde{A}_k$、$\widetilde{F}_k$、$\widetilde{\theta}_k$。由於諧波頻率值 $\overline{F}_i$ 所定義的音高，其實是在音節錄音時就決定了，因此我們必需作音高的調整，以使連續的控制點各自的音高所串成的音高軌跡，能夠滿足韻律單元所指派的基週軌跡要求。在此假設諧波頻率值 $\overline{F}_i$ 所定義的音高爲 100Hz，而韻律單元要求的音高是 150Hz，那麼一個簡單的調整作法是，令 $\widetilde{F}_k = \overline{F}_k \cdot 150/100$，而 $\widetilde{A}_k = \overline{A}_k$ 且 $\widetilde{\theta}_k = \overline{\theta}_k$，這就如圖 4 所畫的情況，由此圖可看出，音高的確可由 100Hz 調整到 150Hz，但是共振頻率(formant frequency)值也被調高了 1.5 倍，例如圖 4 的第一共振頻率 240Hz 被調高成爲 360Hz，這樣的共振頻率的改變，其後果是音色也被改變了，如果各個控制點上的頻率調整倍率高高低低不一致，則音色也會變來變去地不一致。

要在音色保持一致的前提下，調整一個控制點的音高，我們必需遵守的原則是，要讓頻譜包絡保持不變[8]，也就是頻率值被調到 $\widetilde{F}_k$ 的第 k 個諧波的振幅 $\widetilde{A}_k$，它的值必需從原始音高之諧波振幅值 $\overline{A}_i$ 所構建的頻譜包絡曲線中去內差出來，較詳細的作法是，先從舊的諧波頻率序列 $\overline{F}_1$，$\overline{F}_2$，$\overline{F}_3$，…中找出最靠近 $\widetilde{F}_k$ 且比 $\widetilde{F}_k$ 小的頻率值，令找出的是 $\overline{F}_j$，接著，就以 $\overline{F}_{j-1}$，$\overline{F}_j$，$\overline{F}_{j+1}$，$\overline{F}_{j+2}$ 四個原始音高之諧波頻率值和它們對應的振幅值，來作階數 3 之 Lagrange 內差，以求出頻率值 $\widetilde{F}_k$ 上所應對應的振幅值 $\widetilde{A}_k$，也就是依我們提出的公式(5)作計算，

$$\widetilde{A}_k = \sum_{m=j-1}^{j+2} \overline{A}_m \cdot \prod_{\substack{h=j-1 \\ h \neq m}}^{j+2} \frac{\widetilde{F}_k - \overline{F}_h}{\overline{F}_m - \overline{F}_h} \tag{5}$$

一個說明上述觀念(即在保持頻譜包絡不變的前提下調整音高)的圖形如圖 5 所示，在此圖裡，各諧波的頻率被調高了 1.25 倍，但舊諧波(直的實線)和新諧波(直的虛線)所構建的頻譜包絡曲線是重疊在一起的，如此就可確保音色的一致。

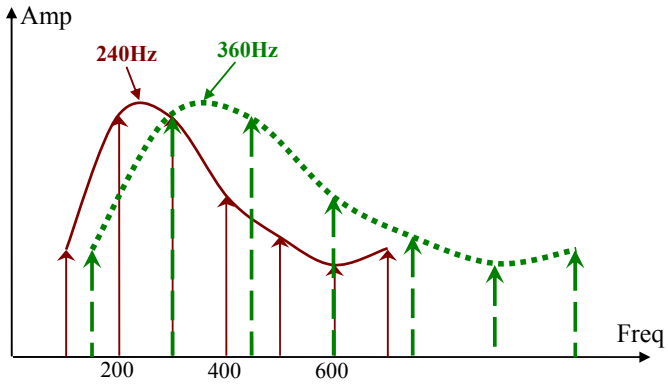

圖 4、音高和頻譜包絡一起被調升頻率值

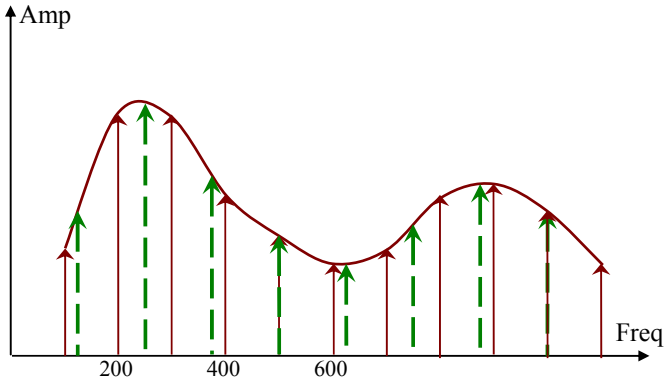

圖 5、音高調升頻率值而頻譜包絡不變

接著考慮頻率爲 $\tilde{F}_k$ 之新諧波的相位參數 $\tilde{\theta}_k$ ，在此我們一樣以頻率爲 $\bar{F}_{j-1}$， $\bar{F}_j$， $\bar{F}_{j+1}$，$\bar{F}_{j+2}$ 之四個舊諧波的相位值，來作 Lagrange 內差而計算出 $\tilde{\theta}_k$ 的值。不過，舊諧波的相位值在作內差之前必需先經過相位反包裹的運算，以避免相位值發生不連續的情況，詳細作法是，計算 $\hat{\theta}_{j-1} = \bar{\theta}_{j-1}$ ， $\hat{\theta}_j = puw(\bar{\theta}_j, \hat{\theta}_{j-1})$ ， $\hat{\theta}_{j+1} = puw(\bar{\theta}_{j+1}, \hat{\theta}_j)$ ， $\hat{\theta}_{j+2} = puw(\bar{\theta}_{j+2}, \hat{\theta}_{j+1})$ ，再拿計算出的相位值去作內差。

四、信號波形合成

在圖 3 裡合成音節的有聲部分，合成的信號 $S(t)$是由諧波信號 $H(t)$和雜音信號 $N(t)$相加而得到，也就是 $S(t) = H(t) + N(t)$。但是詳細說來，$H(t)$其實表示多個諧波信號成分的加總，而且 $N(t)$也是表示多個雜音信號成分的加總，在以下的兩個子節就分別說明 $H(t)$和 $N(t)$的合成方法。

4.1 諧波信號合成

對於位於第 n 和第 n+1 個控制點之間的諧波信號 $H(t)$，可以如下我們修正過的公式來計算它的樣本值，

$$H(t) = \sum_{k=0}^{L} a_k^n(t) \cos\left(\phi_k^n(t)\right) \quad , \quad t = 0,1,...,99 \, , \tag{6}$$

$$a_k^n(t) = \tilde{A}_k^n + \frac{t}{100}(\tilde{A}_k^{n+1} - \tilde{A}_k^n) \, , \tag{7}$$

$$\phi_k^n(t) = \phi_k^n(t-1) + 2\pi f_k^n(t)/22{,}050 \quad , \quad \phi_k^n(0) = \hat{\theta}_k^n \, , \tag{8}$$

$$f_k^n(t) = \tilde{F}_k^n + \frac{t}{100}(\tilde{F}_k^{n+1} - \tilde{F}_k^n) \, , \tag{9}$$

其中 L 表示諧波成分(弦波)的個數，100 是相鄰控制點之間間隔的樣本數，22,050 是取樣率，$a_k^n(t)$ 表示第 k 個諧波在時刻 t (從第 n 個控制點算起之第 t 個樣本) 的振幅，$\phi_k^n(t)$ 表示第 k 個諧波累積到時刻 t 時的相位， $f_k^n(t)$ 表示第 k 個諧波的時變頻率，而

$\hat{\theta}_k^n = puw(\tilde{\theta}_k^n, \hat{\theta}_k^{n-1})$ ，也就是$\tilde{\theta}_k^n$對$\hat{\theta}_k^{n-1}$作反包裹運算後的相位。

當使用公式(6)來合成信號樣本時，累積的相位值$\phi_k^n(t)$通常會在邊界的時間點(即 t=0 或 t=100 時)上發生不連續的現象，亦即$\phi_k^n(100) \neq \phi_k^{n+1}(0)$，這會導致信號波形的不連續，而讓人聽到喀噠聲(click)。爲了避免此種相位的不連續，我們可先計算出在邊界時間點 t=100 時，兩者相差的相位量ξ_k^n，然後把差異的相位量平均地分配給相鄰控制點之間的 100 個樣本點，如此隨時間累積的相位就會變得連續了。在此我們所提用以計算ξ_k^n的公式是，

$$\xi_k^n = puw\left(\phi_k^n(100), \phi_k^{n+1}(0)\right) - \phi_k^{n+1}(0) \tag{10}$$

其中的相位反包裹運算 $puw(x,y)$就如公式(4)所定義的，而$\phi_k^n(100)$可直接以如下我們所提的公式來計算出，

$$\phi_k^n(100) = \phi_k^n(0) + \frac{\pi}{22,050}(101\tilde{F}_k^{n+1} + 99\tilde{F}_k^n) \tag{11}$$

此公式是從公式(8)和(9)作反覆地疊代而得到。當把相位差異量ξ_k^n作平均分配之後，公式(6)就可以改寫成爲

$$H'(t) = \sum_{k=0}^{L} a_k^n(t) \cos\left(\phi_k^n(t) - \frac{t}{100} \cdot \xi_k^n\right) \quad , \quad t = 0,1,...,99 \tag{12}$$

關於 L 的值，令 L_n 表示第 n 個控制點上的諧波成分的個數，一般來說 L_n 和 L_{n+1} 的值可能會不相等，此時我們就令公式(6)和(12)中的 L 值爲 L_n 和 L_{n+1} 的較大者。當 L_n 比 L_{n+1} 小時，我們就需爲第 n 個控制點擴增出 L_{n+1} - L_n 個諧波成分及定義這些諧波的參數值，以便能夠套用前述的信號合成之公式，在此我們基於信號波形連續性的考慮，就簡單地定義 $\tilde{A}_k^n$=0, $\tilde{F}_k^n = \tilde{F}_k^{n+1}$, $\tilde{\theta}_k^n = \tilde{\theta}_k^{n+1}$, k = 1+L_n, 2+L_n, …, L_{n+1}。

4.2 雜音信號合成

對於位於控制點 n 和 n+1 之間的雜音信號 $N(t)$，在合成處理上我們以固定頻率間距的多個弦波信號成分的加總來計算 $N(t)$的樣本值[5]。令 G_k 表示第 k 個弦波的頻率，由於 G_k 值不會隨著時間 t 改變，因此不需區分是在那一個控制點上的，在此令 G_k 代表的值是 100·k (Hz)，這樣 G_k 的下標 k 的起始值就不是 1 了，且各控制點上的起始值也可能會不一樣，因此我們以 K_S^n 表示第 n 個控制點上下標 k 的起始值，它其實可由控制點的 MVF 值來計算得到，即 K_S^n = ⌈MVF(n) / 100⌉；此外，下標 k 的終值是一個固定值，K_e=⌊11,025 / 100⌋，因爲 G_k 不可大於取樣率的一半。

另外，在第 n 個和第 n+1 個控制點之間，第 k 個弦波的振幅我們以$b_k^n(t)$表示，這表示它的值會隨著時間 t 在改變，實際上它是依據時間邊界點(即第 n 和第 n+1 控制點)上的振幅參數 B_k^n 來作線性調整的，至於 B_k^n 參數的求取方式是，先將第 n 個控制點上的 10 個倒頻譜係數補上一序列的零，使成爲 2048 個數值的序列，然後作反向 DFT(discrete Fourier trandform)轉換，取指數，而得到頻譜係數，X_j, j=0, 1, …, 2047，接著依 G_k 找出

兩個相鄰的 X_j，其下標 j 代表的頻率值會包含 G_k，然後對這兩個 X_j 作線性內差，來求出 B_k^n 的值。

當相鄰的兩個控制點上的弦波起始編號 K_s^n 和弦波振幅值 B_k^n 都算出之後，接著就可用這些參數值來合成這兩個控制點之間的雜音信號，我們修正後的計算公式爲，

$$N(t)=\sum_{k=K_s}^{K_e} b_k^n(t)\cos\left(\gamma_k^n + t\cdot 2\pi G_k/22{,}050\right) \quad , \quad t=0,1,...,99\,, \tag{13}$$

$$b_k^n(t)=B_k^n+\frac{t}{100}(B_k^{n+1}-B_k^n)\,, \tag{14}$$

$$\gamma_k^n=\gamma_k^{n-1}+100\cdot 2\pi G_k/22{,}050\,, \tag{15}$$

其中 K_s 設定成 K_s^n 和 K_s^{n+1} 中的較小者，γ_k^n 表示第 n 個控制點上第 k 個弦波的初始相位值。

關於圖 1 和 3 裡長時間無聲子音信號的合成，公式(13)、(14)和(15)仍然可被使用，不過，公式(13)裡的下標 k，它的起始值就要改成 1 了，這相當於設定 MVF 之值爲 0Hz。

五、信號合成實驗和聽覺測試

由於數年前我們曾提出一個改進的 PSOLA 變種之合成方法，稱爲 TIPW[10]，因此我們想要比較 TIPW 法和本文研究的 HNM 擴充之合成法，兩者在信號清晰度上的表現，如果一個合成的語音信號聽起來較不吵雜(noisy)、較無迴音(reverberant)，則它可說是具有較高的清晰度。這裡我們主要關心的是合成信號的清晰度，因此在文句分析和韻律參數產生方面就使用相同的程式模組[11, 12]。至於語音單元，兩種合成法一樣都使用國語音節爲單元，並且都不作單元選擇，因爲每一種國語音節都只存了一遍平調的發音。當使用一台 CPU 爲 Pentium 2.6GHz 的電腦來作語音信號的合成處理時，這兩種合成方法都可以被即時地執行，不過執行速度是有差異的，HNM 擴充之合成法的執行速度只有 3 倍的即時速度，即合成 30 秒的語音信號需花費 10 秒的 CPU 時間，而 TIPW 合成法的速度可以快到 20 倍的即時速度，即合成 30 秒的語音僅需花費 1.5 秒的 CPU 時間。

首先以觀察聲紋圖(spectrogram)的方式來比較這兩種合成方法。兩方法分別去合成出國語短句〝旋轉力〞/syuen-2 zhuan-3 li-4/的語音信號，然後以聲紋分析軟體(在此使用 wavesurfer)作分析來得到聲紋圖，圖 6 的聲紋是對 HNM 擴充法所合成的信號作分析而得到，圖 7 的則是對 TIPW 法合成的信號作分析而得到。從圖 6 和 7 我們可觀察倒，圖 7 裡的諧波紋路顯得比圖 6 裡的較爲零碎、較多斷裂的地方，並且圖 6 裡的諧波條紋較爲平滑，而不像圖 7 裡的顯得有一些毛燥、扭曲，因此 HNM 擴充法合成出的信號應會比 TIPW 法的清晰。

此外，我們選了一篇短文來讓這兩種方法去合成出語音信號，並且存成波形檔案，短文是一篇小學生的作文，有 132 個音節。接著我們將這兩個波形檔以隨機次序播放給 15 位參加聽覺測試者聆聽，然後請他們對前、後播放的檔案作清晰度的比較，評分的規則是，兩者無法區分時給 0 分，如果後者(前者)比前者(後者)稍好一些，則給 1 分(-1 分)，

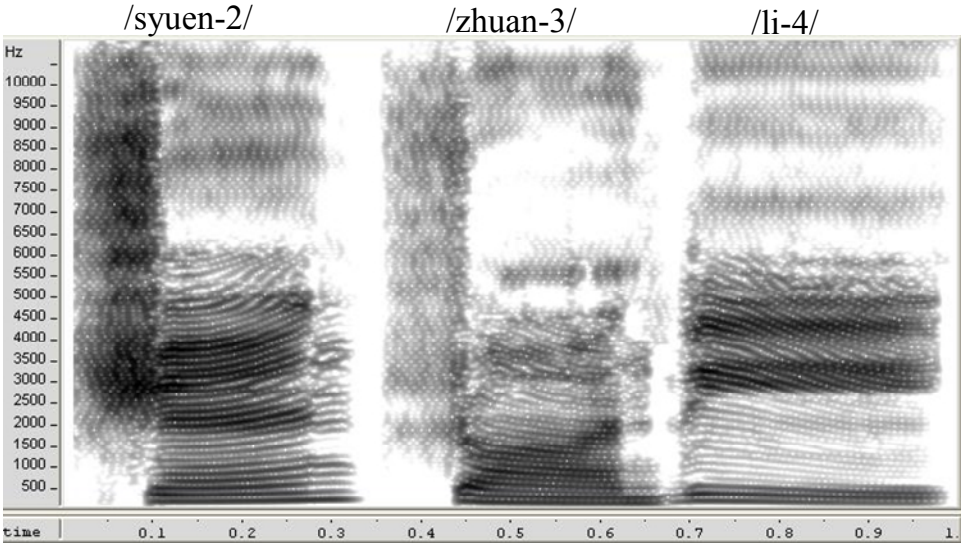

圖 6、HNM 擴充法所合成信號的聲紋圖

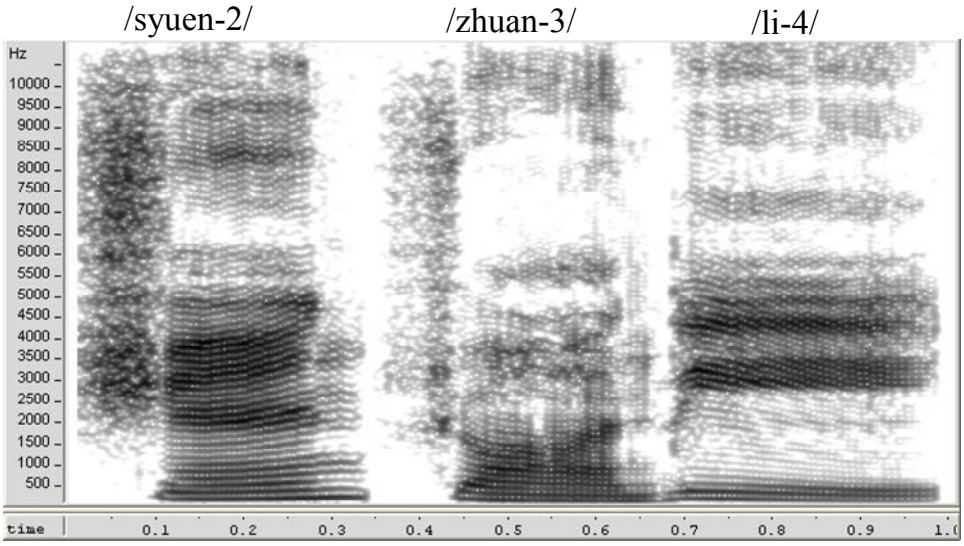

圖 7、TIPW 法所合成信號的聲紋圖

而如果是明顯地好或好很多則給 2 分(-2 分)，結果我們得到的平均分數是 1.53 分，也就是 HNM 擴充法會合成出比較清晰的語音。另外，爲了讓有興趣者能夠試聽這兩種方法所合成出的語音信號，我們設定了一個網頁以供人瀏覽，其網址是 http://guhy.csie.ntust.edu.tw/hmtts/hnm-demo.html。

六、結語

本文以 HNM 爲基礎，考慮了三個議題，即(1)如何保持音色的一致性，(2)如何決定控制點上的 HNM 參數值，(3)如何校正合成音節的時間軸。由於每一種國語音節(不區分聲調)，都只錄、存一次發音而已，所以必需作基週軌跡的調整，這就引發了音色一致性的問題；再者，合成音節的音長可能和原始音節的音長相差很多，如此在合成音節時

間軸上均勻佈放的控制點，各點上的 HNM 參數就會有數值訂定的問題；此外，要提升合成音節的流暢度，就會牽涉到合成音節時間軸的校正問題。本文對前述的三個議題，分別提出了可行的解決方法，在此稱它們整體爲 HNM 擴充之音節信號合成法。

爲了檢驗所提出的方法的效能，我們已把它製作成可即時執行之軟體，然後將它合成出的語音波形，和另一種 TIPW 法所合成的，去作聲紋比較和聽覺測試，初步結果顯示，本文提出的合成方法的確可明顯地提升合成語音的品質(較清晰、無迴音)。

雖然本文所提的 HNM 擴充之合成法，已可明顯地提升合成語音的清晰度，但是所合成的、供聽測用的語音信號檔案，聽起來仍然令人覺得有很強的機器味道，其主要原因應是韻律參數值的產生模型不佳所造成，例如音節的音量和音長數值只是依據幾個簡單規則來作決定而已，因此未來我們將再研究韻律模型改進的問題。

參考文獻

[1] Moulines, E. and E Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones”, *Speech Communication*, Vol. 9, pp. 453-467, Dec. 1990.

[2] Dutoit, T., *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

[3] Chou, Fu-chiang, *Corpus-based Technologies for Chinese text-to-Speech Synthesis*, Ph.D. Dissertation, Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, 1999.

[4] 張唐瑜, *以大量詞彙作爲合成單元的中文文轉音系統*, 碩士論文, 國立中興大學資訊科學研究所, 2004.

[5] Stylianou, Yannis, *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.

[6] Stylianou, Yannis, “Modeling Speech Based on Harmonic Plus Noise Models”, *Nonlinear Speech Modeling and Applications*, Springer-Verlag, Germany, 2005.

[7] Quatieri, T. F., *Discrete-Time Speech Signal Processing*, Prentice-Hall, NJ, USA, 2002.

[8] Dodge, C. and T. A. Jerse, *Computer Music: Synthesis, Composition, and Performance*, second edition, Schirmer Books, New York, 1997.

[9] Moore, F. R., *Elements of Computer Music*, Prentice-Hall, 1990.

[10] Gu, H. Y. and W. L. Shiu, “A Mandarin-Syllable Signal Synthesis Method with Increased Flexibility in Duration, Tone and Timbre Control”, *Proceedings of the National Science Council ROC(A)*, Vol. 22, No. 3, pp. 385-395, 1998.

[11] Gu, H. Y. and C. C. Yang, “A Sentence-Pitch-Contour Generation Method Using VQ/HMM for Mandarin Text-to-speech”, *International Symposium on Chinese Spoken Language Processing*, Beijing, China, pp. 125-128, 2000.

[12] Gu, H. Y., Y. Z. Zhou, and H. L. Liau, “A System Framework for Integrated Synthesis of Mandarin, Min-nan, and Hakka Speech”, to appear in *International Journal of Computational Linguistics and Chinese Language Processing*.