基於加式弦波模型之國語歌聲合成方法
A Mandarin Singing-Voice Synthesis Method Based on Additive Sinusoidal Model

古鴻炎　　陳安瓚
Hung-Yan Gu　and　An-Siuen Chen

國立台灣科技大學資訊工程系
e-mail: guhy@mail.ntust.edu.tw

摘要

本文以加法式弦波模型為基礎，研究國語歌聲的合成方法。在分析階段，對於音源音節各音框的傅立葉頻譜作倒頻譜濾波，以量測出較準確的諧波參數。在合成階段，為了保持音色的一致性，及從一個音源音節去合成出不同的音高、音長的歌聲音節，我們研究了弦波模型內各諧波參數數值的決定方法；接著進一步改進模型的合成機制，而能夠合成出轉音與抖音的效果。此外，我們也進行了主觀的聽測評估實驗，來對本文的方法和先前提出的 TIPW 法作比較，結果顯示本文方法所合成出的國語歌聲，在自然度與清晰度上均有顯著的提升。

關鍵詞: 國語歌聲合成, 弦波模型, 音色, 轉音

1. 前言

歌唱是一項結合文藝(歌詞創作)與音樂(旋律創作)的文化活動，人們常藉歌唱以舒發情感，提振精神，及自娛娛人。工商發達後，生活步調變快，歌唱活動的需求與重要性也更為增加，如人們常去卡拉 OK 唱歌以舒解壓力。雖然大多數人都有唱歌的需求，但是仍有許多人因五音不全(音調或拍子不準)，而不敢唱歌來自娛、娛人，所以本研究希望發展國語歌聲合成的技術，來讓電腦幫助人們唱歌。此外，由於個人電腦的普及，人們也可以藉由電腦歌聲合成軟體來學習新的歌曲，學習歌譜上的老歌或民謠；而專業的作詞、作曲者，也可藉由歌聲合成軟體，來快速地聆聽、評估自己的作品。

歌聲合成的研究，相對於語音合成的研究，成果文獻是較少的，過去研究歌聲合成，被提出的方法大致可分類成時域、或頻域之處理方法。屬於時域的合成方法如 PSOLA (pitch synchronous overlap and add)[1, 2]、TIPW (time-proportioned interpolation of pitch waveform)[3, 4]、波表(wave table)合成法[5]。一般來說，時域上的處理方法，事先所作的參數分析(如基週頂點標記)，較為簡單、直接，且合成處理的計算量較少。前述的 PSOLA、TIPW 之方法，是起源於語音合成的研究，不過可用於作語音合成的方法，不一定就適合用於作歌聲合成，因為歌聲需求的音高變化(音域)是較大的。此外，波表合成法是一種常被用於作樂器聲音合成的方法，目前還未看到有關此法用於作歌聲合成的研究報告。

在頻域上作參數分析，再據以作樂音(樂器聲、人聲)信號合成的方法，過去在電腦音樂合成之研究領域[6, 7]，已發展出至少三類以上的合成方法，主要的類別為: (a)加法式合成(additive synthesis)，先產生出各個諧波後再相加[8]；(b)減法式合成(subtractive synthesis)，如 LPC 編碼之合成方法[9]；(c)頻率調變(frequency modulation)合成，過去有不少研究以此法來作樂器聲之合成[10]。除此之外，最近有一些研究成果使用了弦波模型(sinusoidal model)來作歌聲合成[11, 12]，基本上他們是採取頻域上加法式合成的觀念。

作歌聲合成時，一個基本的、必需顧慮的是音色(timbre)一致性的問題，即不管音高(基本頻率)變高或變低，音色都需保持為同一個人的[6]，一般常犯的錯誤是音色隨著音調高低在改變，就如同錄音機快速或慢速播放時發生的情形一樣。此外，國語歌曲裡常會聽到轉音(sweeping)之唱法，即一個字要唱 2 個甚至 3 個音符，比起一個字唱一個音符是較為困難的，前人所研究的弦波模型之合成方法，並未嘗試作轉音的合成，而在本論文裡，我們嘗試作轉音合成的研究，希望合成出像人一樣平順的轉音歌唱聲。在轉音之外，國語歌聲合成時也需要特別注意的是，音節前面無聲子音(如/s, ts/)部分的處理，因為前面提到的合成方法，都只著重於有聲(voiced)部分，然而無聲子音若處理不當，不只是子音本身變得不清楚，也會使拍子、節奏錯掉，因此我們考慮了無聲子音的合成問題。除此之外，真人歌聲裡可聽到的一個現象是抖音(vibrato)，抖音形成的原因是，音高變得不固定隨著時間作慢速(每秒 5 次左右)擺動[6, 7, 13]，依據此原理，本論文也嘗試了抖音的合成。

依據我們過去的研究經驗[3, 4, 8]，不管使用時域或頻域上的方法，我們所合成出的歌聲都仍然有缺點，例如時域方法所合成出的聲音，會比較呆滯而不夠靈活，而弦波模型方法合成出的聲音，則會

有類似金屬聲音的感覺，且當音高變得較高或較低時，音色也會產生一些改變。不過兩者比較起來，加法式弦波模型之合成方法，比較能夠合成出活靈活現的歌唱聲，因此在本論文裡，我們還是選擇以加法式弦波模型的方法作為基礎，再據以作改進，以加入更多的功能。所謂的弦波模型，其一般化的公式為:

$$s[n]=\sum_{h=0}^{H-1} A_h[n]\cdot cos\left(2\pi\cdot f_h\cdot\frac{n}{F_s}+\theta_h\right) \qquad (1)$$

其中 $s[n]$表示第 n 個樣本時刻上的信號樣本值，h 為諧波編號，$A_h[n]$ 表示第 h 個諧波在樣本時刻 n 時的振幅(即為時變的)，f_h 表示第 h 個諧波的頻率值，而θ_h 則是第 h 個諧波的初始相位，F_s 是取樣頻率。

使用弦波模型作國語歌聲合成，實作上要分成兩個階段來進行，即分析階段與合成階段，在分析階段我們首先錄製國語基本音節的發音到電腦裡，接著對各個音節的信號作分析，以決定音節前段無聲(無週期性)部分與後段有聲(有週期性)部分的分界點，然後把後段分割成音框，以求出各音框的弦波模型參數數值。分析得到的參數值存檔後，在合成階段就可取出來作即時的合成處理，其中第一項處理是，依據拍數、拍長來調整音節的時長(duration)，及無聲、有聲部分的時長分配；第二項處理是，依據音高頻率來調整各個諧波的振幅值，而於合成轉音或抖音的效果時，也要讓諧波頻率成為時變的；接下去再對無聲、有聲部分分別去合成出信號波形，而整體波形也要作振幅的調整，以控制音量。

2. 信號分析階段

為了提供合成階段所需之弦波模型參數，如公式(1)裡的 A_h, f_h, h = 1, 2, …, H, 我們使用 DAT (digital audio tape)數位錄音機，請一位音色較亮麗的女性來唸出國語的 409 個第一聲音節，各個音節獨立發音，然後轉存各音節的一份發音至電腦，在電腦上的音檔，取樣率為 22050Hz 且解析度是 16bits/sample。

接著我們作音節無聲、有聲部分的邊界點的標記，全部以手動的方式來對 409 個音節作標記，以避免程式自動標記可能引起的誤差。之後依據標記的有聲部分起始位置來設定長度為 512 點，且重疊一半(即 256 點)的音框序列，分割成音框是因為音節有聲部分大多不是單母音之情況，因此必須當作是時變的信號來分析。分析一個音框首先作快速傅立葉轉換(FFT)，作完 FFT 後可得到 256 個頻譜參數，然後對這些頻譜參數作處理，就可得到各個諧波的三個參數:振幅 A_h, 頻率 f_h, 相位 ϕ_h。

我們研究了一種找諧波頻率位置的方法，它分成兩個處理步驟：倒頻譜(cepstrum)定位、諧波擷取。倒頻譜定位是為了更精準的找出諧波頻率的位置，所謂的倒頻譜[14]，就是將 FFT 頻譜的振幅取對數後，再作一次快速傅立葉轉換，轉換得到的頻譜參數，低頻參數可視為原先頻譜的頻譜包絡部分，高頻參數可視為是原頻譜中的峰谷部分。因此我們將倒頻譜上的高頻參數部分清除為 0 值，再作快速傅立葉反轉換，以得到原頻譜的包絡曲線，如圖 1 中所示的較平滑的曲線。接著將原有的頻譜曲線扣掉倒頻譜濾波與還原後所得到的包絡曲線，就可以得到具有大約等高峰值之頻譜曲線，如圖 1 中下排所示的跳動曲線。由此可知，倒頻譜濾波處理可以讓頻譜峰點定位得更準確及減少諧波的偵測錯誤。

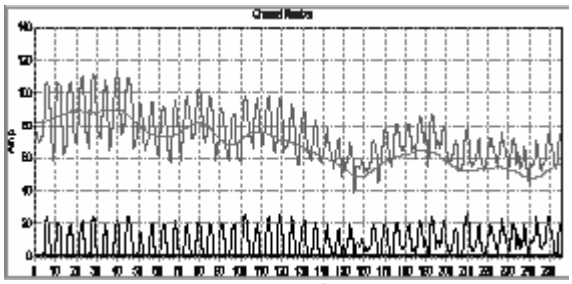

圖 1 倒頻譜濾波處理之示意圖

在諧波擷取步驟，首先要找出第一個諧波(基頻)的參數值，對於取樣頻率為 22,050Hz 的時域波形，作音框長度 512 點的快速傅利葉轉換，則相鄰二個頻譜點的頻率值會相差 11,025Hz / 256 = 43.066Hz，再者成年女生聲音的基頻約為 150~300 Hz 之間，所以只需找前面第 4~7 個頻譜點的振幅最大值即可。找到振幅最大的頻譜點後，再利用其前後各兩點的振幅值作 Lagrange 內插[15]，以求出 Lagrange 最大值(即最大振幅值) A_1 及其相對應的頻率值(即基頻)f_1 和相位值ϕ_1。在求出基頻之後，我們再依據諧波的倍頻關係，來求出其餘各個諧波所在頻率範圍的區域振幅最大值(也經由 Lagrange 內插)，並求出它所對應的頻率值與相位值。實作上，如果我們只是直接對頻譜點的振幅取區域最大值，則很可能將非諧波的頻率點誤判為諧波及錯失真正的諧波頻率值。

3. 歌聲合成階段

分析得到各音節內的邊界點及諧波參數之後，我們就可依據欲合成之音節及所對應音符的音高、音長數值，來作信號波形的合成處理。但是，考慮到轉音、抖音的合成處理，實作上公式(1)並不適合直接用來計算信號樣本的數值，因此我們首先把公式(1)改換成如下之形式[6, 7]:

$$s[n]=\sum_{h=0}^{H-1} A_h[n]\cdot cos\left(\omega_h[n]+\theta_h^0\right) \quad (2)$$

其中 $\omega_h[n]$ 表示第 h 個諧波從樣本時刻 0 到時刻 n 所累積起來的相位量，θ_h^0 表示從第一個音框分析得到的初始相位。早先在電腦音樂合成的研究領域，以公式(2)取代公式(1)，是為了把乘算換成加算以節省計算量，所以目的是不一樣的。

接者依據 $\omega_h[n]$ 的定義，可知 $\omega_h[n]$ 會滿足如下之遞迴關係:

$$\omega_h[n] = \omega_h[n-1] + \Delta\omega_h[n] \quad (3)$$

其中 $\Delta\omega_h[n]$ 表示第 h 個諧波在樣本時刻 n 時的相位增量，如果先不考慮**轉音**、**抖音**效果的合成，則 $\Delta\omega_h[n]$ 可簡化成 $\Delta\omega_h$，即和時刻 n 無關，而它的數值可由下列公式求得:

$$\Delta\omega_h = (2\pi\cdot f_h)/F_S \quad (4)$$

即由諧波頻率 f_h 來決定。所以，綜合公式(2)、(3)、(4)可知，在合成一個音符的歌聲音節時，如果音符的音高是固定的(無轉音與抖音)，則各個諧波頻率 f_h 和相位增量 $\Delta\omega_h$ 也都是固定值，而只有諧波振幅 $A_h[n]$ 的值需要隨時刻 n 作調整，以符合國語音節韻母的時變頻譜之特性；而當要合成出轉音或抖音的效果時，則 f_h 與 $\Delta\omega_h$ 也要和 $A_h[n]$ 一樣，變成是時變的。以下各子節就對音節的時長調整問題、音調高低調整的問題、轉音與抖音合成的問題、音量調整的問題，說明我們所研究的作法。

3.1 時長調整

起始時間點的調整: 從樂譜資料可知道各個音符的開始與結束演奏時間，所以可依各個音符的起始與結束時間來編排其對應音節的演唱時間。我們先前的實驗發現，對於非子音開頭的音節來說，直接把音符的起始時間當成音節的起始時間，所合成出的音節信號聽起來，拍子與配樂音符是契合的；但是對於以子音開頭的的音節來說，聽起來則感覺有拖拍的感覺，所謂拖拍就是音節聽起來沒有落在應該落下的時間點上，而落在應落的時間點之後。

要解決拖拍的問題，我們的作法是，建立一個子音(含有聲、無聲)之時間平移表，裡面記錄各個含有子音的音節的平移時間，再把音節按照對應的平移時間作起始時間與時間長度的修改。

換氣模擬: 當相鄰音符之間沒有停頓，即前一個音符的結束時間就是下一個音符的起始時間，此時如果直接依據音符的音長來作音節信號合成，會使歌詞聽起來就像是整串連在一起，十分地機械化，缺少人類歌唱時自然換氣、停頓的感覺，所以必須作換氣停頓的模擬。

我們模擬換氣的方法是，將每一個音符的音長均乘以一個長度調整參數β，β的大小則根據音符長度的歸類來設定，若是音符長度超過 1.3 秒，就將β值設為 0.94，若音符長度沒超過 1.3 秒，就將β值設為 0.85。

子音長度調整: 在我們的系統中，無聲(unvoiced)子音的信號波形並未以弦波模型來產生，而是在時域上對原信號波形直接作調整長度、及串接的處理。當子音為非送氣的爆破音時(如/ㄅ、ㄉ、ㄍ/)，長度不作調整；而當子音為其它的無聲子音時，則要作長度的調整，假設子音與母音的時長比例為 1:2，則我們希望合成的音節信號裡，時長比例亦為 1:2，但是在音節長度延長時，一些摩擦子音的長度若調整得太長，則聽起來會很不自然，所以我們訂定了一個子音時長延長的上限，即設為原始子音長度的 1.2 倍。至於母音部分的長度，就是音節的長度扣掉子音部分的長度了。

子音信號波形伸長或縮短的方法，例如李雪貞的論文[16]裡提到的是，將 hanning window 的長度設為欲合成子音長度的兩倍，然後在原始子音信號左右兩邊分別乘上 hanning window 函數，並移放於合成之子音信號兩端作疊加。若調整後的子音長度太短，則原始子音的中心部份振幅較強的信號，不會被用於疊加，造成子音強度變弱，而聽起來變得不清楚。所以我們研究了另一種子音信號調整的方法，不管伸長或縮短子音，都先將子音部分分割成音框序列，以音框為單位去原始子音中截取及作重組後，再作相鄰音框間的疊加，本方法的好處是，可以保留原始子音波形的振幅包絡，而使得調整過的子音和原始子音一樣清楚。

例如，原始子音長度為 2150 點，而合成子音之長度只需 1500 點，音框長度為 200 點，每 100 點作一次疊加的情況，我們先求得合成子音裡需要的音框數量為 1500 / 100 – 1= 14 個音框，再依數值 14 去計算原始子音裡，必須間隔多少點來取出一個音框，即計算出(2150-200) / (14-1) = 150，接著以間隔 150 點取一個音框的方式，從原始子音中取出 14 個音框來作疊加，如此就可得到合成子音的信號波形，疊加時所用的加權值仍然是以 hanning window 函數來計算。

3.2 音色、音高、與諧波振幅

要讓音高(基本頻率)升高或降低，不能夠採取 resampling 方式來處理[5, 10]，因為會造成音色隨著音調高低在改變，其原因是基頻和共振頻率(formants)同時受到同方向的改變(升高或下降)。其實要讓音色、音高兩項因素獨立開來，觀念上並不困難，因為從短時 FFT 頻譜上來看，音高是由第一個諧波的頻率值(基本頻率)或相鄰諧波間的頻率差

距所決定，而音色則是由各諧波的振幅峰點所連接成的包絡(envelop)曲線的形狀所決定。因此，在調整各諧波的頻率位置時，其振幅高度也必需作調整，以保持包絡曲線為固定不變的形狀，如此音色就可維持不變。

在分析階段我們僅將各音框的各個諧波的三個參數 A_h, f_h, ϕ_h 記錄下來，因此若要決定一個新的頻率上的諧波的振幅，則需要依據先前記錄的諧波參數來作內插。由於線性內插的誤差比較大，合成出的音色仍可能受到一些改變，所以我們利用先前記錄的諧波振幅及頻率參數來作 Lagrange 內插，內插的公式為:

$$P(x)=\sum_{j=0}^{3}P_j(x)\ ,\ P_j(x)=y_j\prod_{\substack{k=0\\k\neq j}}^{3}\frac{x-x_k}{x_j-x_k} \quad (5)$$

其中 x 表示一個新的諧波頻率，$P(x)$表示 x 上內差出的振幅，x_0, x_1, x_2, x_3 表示 x 兩邊最接近 x 的 4 個原音節中的諧波頻率值，y_j 則是 x_j 上的振幅值。在保持包絡形狀不變的條件下，來內插出一個新頻率上的振幅值，其圖形說明如圖 2 所示。

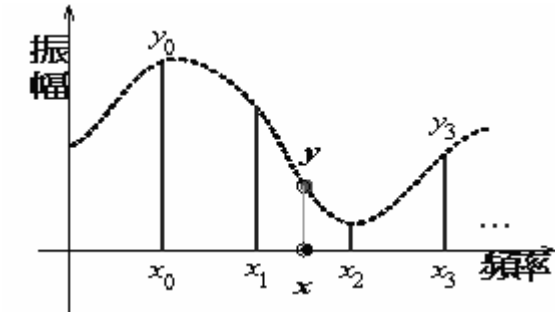

圖 2　振幅內差示意圖

在第一個音框上，要決定一個新的諧波頻率 x 的初始相位，我們也以 Lagrange 內插來先求出 x 對應的相位ϕ，然後依據公式(4)算出頻率 x 的相位增量，再倒退半個音框點數的相位量回去，即可求出 x 的初始相位θ。

3.3 轉音與抖音

在國語歌曲裡，經常會有一個音節含蓋多個音符的情況，例如在〝何日君在來〞這首歌中，〝好花不常開〞的〝不〞與〝常〞都含蓋了兩個音符，這樣的情況稱為轉音。圖 3 是一個歌唱聲轉音的聲紋圖(spectrogram)[14]，可以明顯的看出歌唱聲轉音時，各個諧波的頻率是平滑地自前一個音符的頻率轉變成下一個音符的頻率，如圖 3 上的橫向白色曲線所示。此外，在轉折區段中，頻率越高的諧波其曲線的斜率越大；並且在頻譜上，諧波振幅的包絡，在轉音區段內並不會受到轉音的影響。

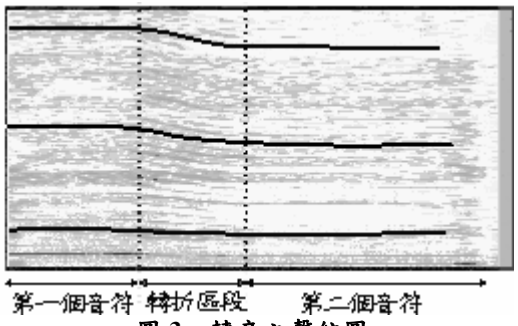

圖 3　轉音之聲紋圖

因此在轉音區段，我們必須機動地對各個諧波的頻率值及其對應的振幅值作調整。關於諧波頻率值的調整，考慮到聲紋圖上的變化必須保持平滑，所以不能夠將相位增量直接以線性內插來計算，我們的作法是，先將前後音符的各個諧波各自的相位增量都計算出來，進入轉折區段時，再把前後音符的對應諧波作相位增量的曲線式內插，在此我們使用了 cos 函數來作曲線內插。假設某一個音節含蓋了兩個音符，設其中第 i 個音符的第 h 個諧波頻率為 f_h^i，則第 1 與第 2 個音符的第 h 個諧波頻率值的關係可能會有兩種情況，即 $f_h^1 > f_h^2$ 或 $f_h^1 < f_h^2$，因此曲線內插公式必須滿足此二種情況。假設 f_h^1 與 f_h^2 各別對應的相位增量為 $\Delta\omega_h^1$ 與 $\Delta\omega_h^2$，且轉折區段的長度為 M 個樣本點，則在轉折區段中第 m 個樣本點的相位增量 $\Delta\omega_h[m]$，我們使用如下公式來計算:

$$\Delta\omega_h[m]=\left(\frac{\Delta\omega_h^1+\Delta\omega_h^2}{2}\right)+\left(\frac{\Delta\omega_h^1-\Delta\omega_h^2}{2}\right)\cdot cos\left(\pi\cdot\frac{m}{M}\right) \quad (6)$$

使用此公式求出的相位增量值隨著時間改變的一個例子如圖 4 所示。

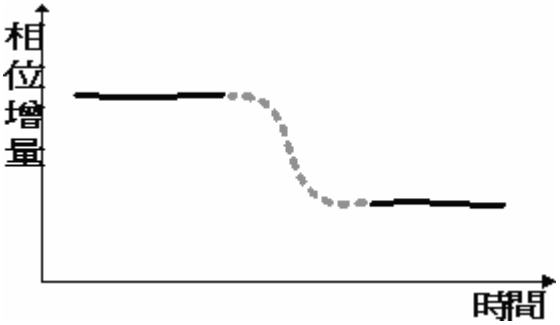

圖 4　曲線式內差之相位增量變化

在轉折區裡，各諧波的頻率值都會逐漸改變，所以各諧波的振幅值也必須跟隨頻率值的改變而改變，以保持音色的一致性。但是若僅將轉折區段的起始與結束點的諧波振幅值計算出來，後然直接作線性內插來計算各個樣本點上的諧波振幅值，則轉折區段的頻譜包絡會變形，而使信號音色聽起來變成不同人的聲音。因此，對於第 h 個諧波，其振

幅在第 m 個樣本點上的值 $A_h[m]$的求取，我們必須先依公式(6)算出時變之相位增量 $\Delta\omega_h[m]$，再用公式(4)反向求出對應的頻率值 $f_h[m]$，接著帶此頻率值到公式(5)內作 Lagrange 內插，以便求出時變的、可讓音色保持一致之諧波振幅值 $A_h[m]$。

人類歌聲還有一項重要特性，就是抖音，我們可從抖音信號的聲紋圖上，明顯地看到各諧波的紋路會如弦波一般地跳動，也就是一個諧波的頻率值會隨著時間快速地變化。前面我們以公式(6)來變化相位增量的值，可以使聲紋圖上的諧波頻率值以曲線狀緩慢改變，同理我們也可用此作法來模擬抖音，只是要把變化的速度加快。設第 h 個諧波的相位增量為 $\Delta\omega_h$ ，且抖音頻率為 f_v（即每秒抖動 f_v 次），則抖音區段中第 m 個樣本點上的相位增量 $\Delta\omega_h[m]$ 的計算公式為:

$$\Delta\omega_h[m] = \Delta\omega_h \cdot \left(1 + \lambda \cdot sin\left(2\pi \cdot f_v \cdot \frac{h \cdot m}{F_s}\right)\right) \qquad (7)$$

其中 f_v 的值通常介於 4.5~7Hz 之間[13]，而 λ 的值可設為 0.2 左右。

3.4 波形包絡與音量調整

在合成出一個音節的時域波型後，如果觀察波形的包絡形狀，時常可發現包絡形狀會和原音節的不一樣，如圖 5 所示，這是因為合成音節的基頻和原音節的基頻不一樣，且音節內各個諧波之間的相對振幅高低，在經過如圖 2 所示之內差處理後，已經變得很不一樣了。此時，如果不對時域波形的包絡形狀作調整，則合成音節的音量會隨著時間在變動，而讓人感覺到不平順。所以在合成好信號波形之後，還要依據原音節的波形包絡，來對合成音節的包絡形狀作調整。

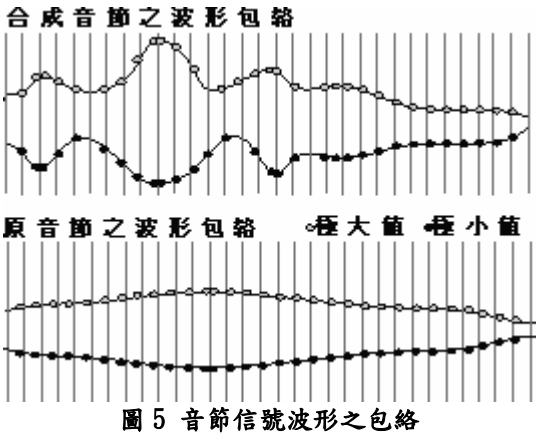

圖 5 音節信號波形之包絡

我們的調整方法是，將原音節與合成音節都平均切割成相同個數之區段，區段個數由原音節長度除以它的 3 倍週期長度來計算；然後對原音節波形與合成音節波形的每一個區段分別找出極大振幅值與極小振幅值，如圖 5 裡白點與黑點的所示。設原音節中第 n 個區段裡的極大與極小振幅值分別是 Max_n^o 、 Min_n^o ，而合成音節中第 n 個區段裡的是 Max_n^s 、 Min_n^s ，則我們訂定合成音節第 n 區段的振幅調整比率為

$$r_n = \left(Max_n^o - Min_n^o\right) / \left(Max_n^s - Min_n^s\right) \qquad (8)$$

另外，人在說話或唱歌的時候，嘴型大小與信號振幅大小會成正比的關係，例如音節/a/的嘴型很大，對應的信號振幅也很大，而音節/i/的嘴型很小，對應的信號振幅也很小，可是兩者的音量聽起來卻是相當的。所以接下來還要依據合成音節的韻母，來對合成音節的波形作振幅調整，至於各韻母的振幅調整量，可根據語音合成方面的研究報告[16, 17]，事先建立一個韻母與振幅衰減量的對照表來作查詢。

4. 聽測評估

由於合成出的歌曲的好壞尚不能直接作量化評分，所以我們以聽測實驗來評估。這裡考慮了三種國語歌曲合成的方式:(a)TIPW 合成法、(b)本文的加式弦波合成法、(c)本文方法合成出的歌聲再加上 MIDI 伴奏。所使用的歌曲為兒歌"只要我長大"，評估的項目有三項，分別是清晰度、自然度、及喜好排序。清晰度主要是評估合成出來的歌聲信號聽起來是否清楚無雜訊以及咬字的清晰程度，自然度則是評估合成歌聲與人聲的接近程度，喜好排序則是將三種合成方式(a)，(b)，(c)，依照喜好程度由高至低排列。評分的標準依照優、佳、可、差、劣五個等級分成 5、4、3、2、1 等五個級分。

我們總共找了 15 位測試者來試聽，測試者的性別為 6 位女性及 9 位男性；年齡層 20 歲至 30 歲共 14 位，30 歲以上 1 位；職業為學生者共 11 位，上班族 4 位。聽測的方法是，隨機排列三種合成方式所合成出的歌聲，讓測試者試聽及評分，評分的平均值結果如表 1 所示。由表 1 的結果顯示，弦

表 1 歌曲"只要我長大"聽測實驗之結果

	清晰度	自然度	喜好排序
TIPW 無伴奏	**3.73**	**3**	**3**
弦波模型 無伴奏	**4.13**	**4**	**1.87**
弦波模型 MIDI 伴奏	**4.46**	**4.46**	**1.13**

波模型所合成出的歌聲比起 TIPW 法合成出的歌聲，其自然度提升了 1 個等級，而清晰度也有些許

的提升，自然度提升的原因，可能是因為以弦波模型來作歌聲合成，並且加入抖音及轉音，可以比較接近真人唱歌時的感覺，因此聽起來較為自然；而清晰度的提升，可能是因為使用本論文的加式弦波模型作合成，聲音較為清亮，且對於咬字的清楚程度也有幫助。

就弦波模型來看，加上伴奏比未加伴奏的歌曲，清晰度與自然度分別上升了 0.33 與 0.46 級分，這顯示有伴奏的樂曲對於人類聽覺而言，可以使合成歌曲聽起來更逼近真人所唱的感覺。至於在三種合成方式中，大部分的測試者的喜好依次為弦波模型有伴奏、弦波模型無伴奏、TIPW 法；只有 2 位的喜好次序為弦波模型無伴奏、弦波模型有伴奏、TIPW 法。所以使用弦波模型作歌聲合成後，再加上 MIDI 伴奏的演奏方式，較為多數人所喜愛。

5. 結論

我們以加法式弦波模型為基礎，對國語歌聲的合成處理進行研究，得到的成果如: (a)發展了一種較準確的諧波參數分析的作法；(b)提出諧波之振幅參數的求值作法，而能夠在保持音色一致的前提下，合成出國語音節的歌唱聲信號；(c)改進模型的信號合成機制，而能夠提供轉音、抖音等效果的合成；(d)提出音節時長與音量的調整作法。對於本文研究的合成方法，我們也進行了聽測實驗，結果顯示本方法所合成出的歌唱聲，在自然度及清晰度上，都比先前提出的 TIPW 法有明顯的改進。

未來我們可朝如下的方向續研究、改進: (a)音節之間仍有分離感而不夠自然，原因是使用線性比例來延長母音區段，未來可改成以 ADSR 方式(電腦樂器聲合成常用的方式)[6, 7]，來作音長延長之處理；(b)目前只作到單聲部歌唱之音節信號合成，因此可再研究多聲部歌唱的信號合成方法，讓合成之歌聲有更豐富的表現。

參考文獻

[1] Hamon, Christian, Eric Moulines, and Francis Charpentier, “A Diphone synthesis System Based On Time-Domain Prosodic Modifications of speech”, IEEE ICASSP, pp. 238-241, 1989.

[2] 林政源，國語歌曲的歌聲合成，碩士論文，國立清華大學資訊工程研究所，2001。

[3] Gu, H. Y. and W. L. Shiu, "A Mandarin-syllable Signal Synthesis Method with Increased Flexibility in Duration, Tone and Timbre Control", Proc. National Science Council, R.O.C., Part A: Physical Science and Engineering, Vol. 22, No. 3, pp. 385-395, 1998.

[4] 古鴻炎，http://guhy.ee.ntust.edu.tw/syn-sound.html (可試聽合成的歌唱聲)，國立台灣科技大學資訊工程系.

[5] Russ, M., Sound Synthesis and Sampling, Boston: Focal Press, 1996.

[6] Dodge, C. and T. A. Jerse, Computer Music: Synthesis, Composition, and Performance, 2'nd ed., New York: Schirmer Books, 1997.

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

[8] 盛思豪，即時歌唱聲合成系統與音樂合成系統之整合，碩士論文，國立台灣科技大學電機研究所，2002。

[9] 邵芳雯，國語歌曲之合成，碩士論文，國立交通大學電信研究所，1994。

[10] Roads, C., The Computer Music Tutorial, MIT Press, 1996.

[11] Macon , M.W., L. Jensen-Link, J. Oliverio, M. A. Clements, and E. B. George, "A Singing Voice Synthesis System Based on Sinusoidal Modeling", IEEE ICASSP-97, Vol. 1, pp. 435 -438, 1997.

[12] O'Brien, D. and A. I. C. Monaghan, “Concatenative Synthesis Based on a Harmonic Model”, IEEE trans. Speech and Audio Processing, Vol. 9(1), pp. 11-20, Jan. 2001.

[13] Meron, Y. and K. Hirose, "Synthesis of Vibrato Singing", IEEE ICASSP '00, Vol. 2, pp. II745 -II748, 2000.

[14] O’Shaughnessy, D., Speech Communications: Human and Machine, 2nd ed., IEEE Press, 2000.

[15] Stoer, J. and R. Bulirsch, Introduction to Numerical Analysis, 2nd ed., New York: Springer-Verlag, 1993.

[16] 李雪貞，客語語音合成之初步研究，碩士論文，國立台灣科技大學資訊工程所，2001。

[17] 潘能煌，中文文句翻語音系統之音量音調韻律研究，碩士論文，國立中興大學應數系，1997。