用於國語歌聲合成之諧波加噪音模型的改進研究
Improving of Harmonic plus Noise Model for Mandarin Singing Voice Synthesis

古鴻炎　　廖皇量
Hung-Yan Gu and Huang-Liang Liao

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

摘要

本論文對*諧波加噪音模型*作改進與擴充，用以作國語歌聲信號的合成，期望能合成出更自然、清晰的歌聲信號。我們首先改進了原本模型的分析步驟的缺點，如 MVF 值的決定及基頻和諧波參數值的估計；然後擴增歌聲合成有關的處理方法，如相位同步和 ADSR，讓原本的模型更適合作歌聲合成；之後加入我們先前的研究經驗，來實作出一個即時的國語歌聲合成系統。用此系統合成出的歌聲，來進行聽測估評，初步結果顯示，我們的歌聲信號合成模型，可以顯著提升國語合成歌聲的自然度與清晰度。

ABSTRACT

In this paper, we study to improve and extend the harmonic plus noise model. The purpose is to synthesize more natural and clearer Mandarin singing voice. First, a few improvements are made to the original model, e.g. the determination of MVF value and the estimation of fundamental frequency and harmonic parameter values. Then, we extend the model to accommodate more processing, e.g. phase synchronization and ADSR, helpful to synthesize higher quality singing voice. In addition, some experiences from our previous studies are adopted and a real-time Mandarin singing voice synthesis system is built. Based on this system, perceptual tests are made to evaluate the performance of our system. Initial results show that our system can significantly improve the naturalness and clarity of the synthetic Mandarin singing voice.

Keywords: singing voice synthesis, harmonic plus noise model, phase synchronization, convolution noise.

1. 前言

由於個人電腦的普及，人們可以藉由電腦歌聲合成的功能來學習新的歌曲，學習歌譜上的老歌或民謠；而專業的作詞、作曲者，也可藉由電腦歌聲合成，來聆聽、評估自己的作品。

歌聲合成的研究，過去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 [1, 2]、TIPW [3, 4]、波表合成法[5]。一般來說，時域上的處理方法，事先需作的參數分析(如基週頂點標記)，較爲簡單、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，因爲歌聲需求的音長、音高變化(音域)是較大的。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，過去在電腦音樂之研究領域[6, 7]，已發展出至少三類以上的方法，主要的類別爲: (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。所謂的弦波模型，其一般化的公式爲：

$$s[n]=\sum_{k=0}^{K-1} A_k[n]\cdot cos\left(2\pi\cdot f_k\cdot\frac{n}{F_s}+\theta_k\right) \qquad (1)$$

其中 $s[n]$表示第 n 個樣本時刻上的信號樣本值，h 爲諧波編號，$A_k[n]$ 表示第 k 個諧波在樣本時刻 n 時的振幅(即爲時變的)，f_k 表示第 k 個諧波的頻率值，而 θ_k 則是第 k 個諧波的初始相位，F_s 是取樣頻率。

本論文採取的諧波加噪音模型(Harmonic plus Noise Model, HNM) [13]，它除了考慮信號中的諧波成分之外，還加入了信號中高頻部份的噪音，使得合成出的聲音的特性，可以更接近原始所錄的聲音。對於諧波部份，HNM 就是以加法式弦波模型來合成出信號；而噪音部份，HNM 將其當成是間隔爲 100Hz 的諧波信號，各諧波的振幅值，則由少量的倒頻譜參數代表的平滑頻譜上取樣得到，相位值以亂數產生。

當初 HNM 被提出時，是要用來作語音的分析與合成，如果將它用在歌聲的合成上，會發現它作信號分析的一些步驟，並不夠完善，因此我們對此提出一些改善的方法，例如：(a)基頻的偵測；(b)諧波參數的擷取；(c)頻譜上諧波與噪音兩部份的分界點的偵測，此分界點稱爲最大有聲頻率(Maximum Voiced Frequency, MVF)。

在 HNM 模型的原先的合成處理步驟之外，我們還擴增了數個用以提升歌聲合成品質的處理方法，包括：(a)加入低頻噪音，以改進 HNM 在信號低頻部分 modeling 能力的不足；(b)相位方面的處理，來讓合成音更能保持原始音的音色特性；(c)ADSR 式的音長分配，以使伸長或縮短過的音，可以維持自然性；(d)作諧波追蹤，來消除音框間諧波數量的不連續。

要合成出自然的歌聲信號，另一重要、必需考慮的層面是，韻律參數值的決定，如音符演唱的音量、音長、起始時間點。在此，我們將之前在歌聲合成方面的研究成果[8, 14]，和本論文研究的歌聲信號合成模型作整合，來製作出一個即時的國語歌聲合成系統。

2. HNM 分析方法之改進

2.1. 基頻偵測

爲了增加基頻偵測上的準確度，我們的修正作法是，先從音節中 Sustain 的部份取出一個音框，並且改成使用一種自相關 (auto-correlation) 函數搭配 AMDF (absolute magnitude difference function) 的偵測方法[15]，求出該音框的基頻值，當作整個音節中各個音框基頻值的參考。由於每一個音節的信號在分析時，會先經過音節分段的處理，所以可知道 Sustain 部份的時間位置。

接著，對於各個音框同樣以自相關函數搭配 AMDF 的方法，在時域上求出可能的初始基頻值；一個音框若是判斷爲有聲，就將該音框信號乘上 Blackman 窗，並且音框後面補上 0，使長度成爲 4096 點後，對音框作 FFT 分析得到頻譜；再依時域上求出的初始基頻，在頻譜上該基頻值的附近，搜尋頻譜振幅上的峰值；此峰值點與其前後頻率點的振幅值再作拋物線內插，來求得振幅的最大值，及其相對應的頻率值，也就是基頻。

2.2. 諧波參數偵測

求得一個音框的基頻值之後，再依照諧波的倍頻關係，來求得其餘各個諧波所在範圍的局部振幅最大值；在此爲求精確，我們一樣加入拋物線內插之處理，以求得振幅峰值，及其對應的頻率值與相位值。

此外，顧慮到合成階段進行音高調整時，能夠有較充分的頻譜資訊，來作較準確的頻譜曲線內插，因此一個音框的高次諧波的偵測，我們一直作到頻率範圍的 95%才停止，並非和原始 HNM 一樣，是依據 MVF。偵測後，將各諧波的振幅與相位值全都記錄到參數檔中。

2.3. MVF 設定

爲了避免各音節在錄音時音量的不一致，而使得音量較小的音節受到固定的 MVF 門檻值的限制，導致音量小的音節的諧波數太少，而使合成音顯現出失真，因此我們改以動態的方式來設定 MVF 門檻值。

動態式設定方式是，先從一個音節的各個音框裡找出各音框的最大諧波振幅，再找出跨音框的最大值，以這個值的 1/512 當作 MVF 的門檻值。然後將 MVF 門檻值和一個音框中各諧波的振幅值作比對，振幅值小於門檻者，視爲無聲的頻率，當發生連續 5 個諧波皆爲無聲時，最後一個有聲的諧波頻率即爲此音框的 MVF，而 MVF 之前所有的諧波頻率都視爲有聲的頻率。

3. 提升信號品質的方法

我們發現 HNM 對信號的 modeling 能力，有一些不足的地方，因此我們研究了以下數種方法，來提升合成出的歌聲信號的自然度。

3.1. Convolution noise

使用原始的 HNM 方法來合成出的聲音，若在頻譜上與原始音作比較，會發現在低頻的部份，合成音的頻譜波谷比原始音的波谷深，如圖 1所示，這是因爲在低頻的部份(頻率值小於 MVF)，HNM 是以純粹的諧波表示。爲了加強 HNM 在低頻部份 modeling 能力的不足，我們嘗試加入 Convolution Noise (CVNS)，希望讓低頻部份之頻譜，可以較近似於原始聲音的。

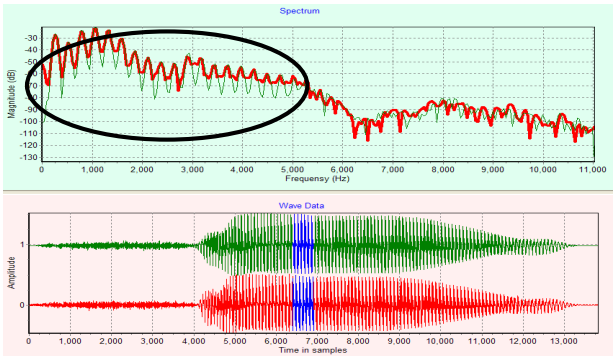

圖 1 原始音與合成音中對應音框之頻譜比較

3.1.1. Convolution noise 分析

先以 HNM 方法依原始聲音的音高、音長，產生出合成的聲音；接著依據 HNM 分析結果得知的有聲部份的位置資訊，分別對原始音、合成音兩個聲音有聲的部份取音框作 FFT 分析，我們取的音框大小是 512 點且重疊一半；分析出兩者的譜頻後再分別作倒頻譜分析，得到倒頻譜參數，最後將合成聲音的倒頻譜參數減去原始聲音的倒頻譜參數，得到 CVNS 參數。

3.1.2. Convolution noise 合成

合成時如果將 CVNS 直接加在合成音對應音框的倒頻譜參數上，會得到低頻過強的吵雜聲音，因此我們先將 CVNS 作兩項處理：(1)去除諧波影響，將近似基週長度和其倍數位置附近的峰值或谷值改設爲 0；(2)設定門檻值，以原先音框信號的各個倒頻譜參數的 0.1 倍作爲變動的最大限值。之後，將 CVNS 加至合成音對應音框的倒頻譜參數上，再反轉換回時域，即可得到含有低頻噪音的合成音信號。

3.2. 諧波相位之同步

我們將每個原始音音框分析得到的相位值都記錄成參數，然後在合成時加以利用，希望如此可以讓合成音保持更多原始音的音色特性。關於相位的處理，我們從下列三個層面來說明，亦即：(1)相位增量控制，(2)諧波之間的相位關係，(3)相對於基頻之相位延遲。

3.2.1. 相位增量控制

爲了方便作轉音和抖音等歌唱技巧的模擬，我們仍然以相位增量的觀念來計算每個取樣點的相位值，如下列公式：

$$\phi_k[n] = \phi_k[n-1] + \Delta\phi_k,\ \Delta\phi_k = \frac{2\pi \cdot f_k}{F_s} \qquad (2)$$

當不考慮原始音音框分析得到的相位值，而依固定的相位增量值，讓相位持續累加下去，則在音框邊界上同一個諧波的相位自然會連續。但是當考慮原始音音框的相位時，依相位增量累加得到的相位值，就不一定會符合下一個音框的起始相位。爲了解決這個情形，我們將累加的相位值與下一個音框作分析時的相位值之間的差，平均分配到音框裡的各個樣本點上，以兼顧相位連續性及配合分析的相位值。這裡，在取相位的差之前，要先作相位的展開，以使兩者的差在 π 之內，相位展開後的差值 d_k^i 的計算公式是

$$M_k^i = \left\lfloor \frac{1}{2\pi}\left(\phi_k^i + N\cdot\Delta\phi_k^i - \phi_k^{i+1}\right)\right\rfloor$$
$$d_k^i = \phi_k^{i+1} + 2\pi \cdot M_k^i - \phi_k^i \qquad (3)$$

其中 ϕ_k^i 表示第 i 個音框的第 k 個諧波在分析時得到的相位值，$\Delta\phi_k^i$ 表示第 i 個音框裡第 k 個諧波的相位增量，N 表示音框的長度。

3.2.2. 諧波之間的相位關係

當音高需要調整時，爲了維持音色的一致性，新音高的諧波頻率，要在原始音高的頻譜包絡上內插得到對應的振幅，相位也是如此，但是要在內插前先將相位展開成平滑的相位曲線，以避免在劇烈起伏的曲線上內插出錯誤的相位，圖 2就是相位展開處理前後之曲線比較。

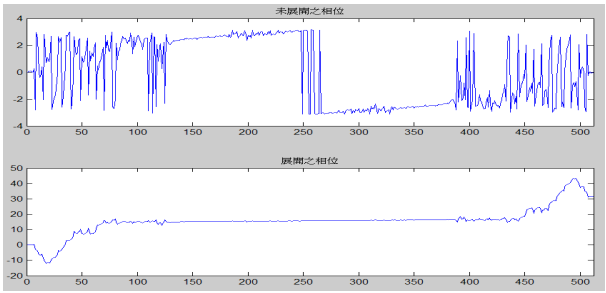

圖 2 未展開與展開之相位曲線比較

3.2.3. 相對於基頻之相位延遲

在切割音框作分析時，音框的長度不一定剛好是週期長度的整數倍，使得每個音框內，相位測量的基準點會在不同的相對位置上[16]，如圖 3裡的兩條縱實線，是相位測量的基準點，離音框的左邊界有不同的時間延遲。因此，各諧波分析得到的相位值，必須加上測量基準點的時間延遲所導入的相位延遲，計算公式爲

$$\phi_{i,k} = \theta_{i,k} + \Delta t_{i,1}\omega_{i,k},\ k = 1,2,\cdots,N \qquad (4)$$

其中 $\theta_{i,k}$ 表示第 i 音框的第 k 個諧波在音框起始處的相位，$\Delta t_{i,1}$ 表示由基頻所定義的測量基準點的時間延遲，$\omega_{i,k}$ 爲第 k 個諧波的角頻率，$\phi_{i,k}$ 表示測量基準點上的基頻同步後的相位。如此讓各諧波之間的相位比較，是在基頻的相同的相位位置上進行(即基週同步)，才能維持住諧波之間的真正相位關係。

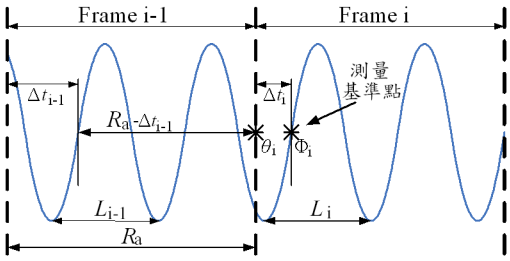

圖 3 相對於基頻之相位延遲

3.3. ADSR 式之音長伸縮

參考電腦音樂裡常用的 ADSR (attack, decay, sustain, release) 觀念[6]，僅在 Sustain 的部份對音長作伸長或縮短，其餘的部份則維持原始長度，以免線性比例調整音長至太短或太長時，合成音聽起來會感覺出不自然。

實際使用 ADSR 觀念於人聲時，在信號分析階段，我們先以人工將一個音節分成數個區段，接著記錄下這些區段的邊界點上的樣本點時間。在合成階段，根據分析時記錄下的區段邊界樣本點時間，分別施以不同的音長處理方式。所標示的邊界有下列三種：(a)起音(attack)部份，(b)釋音(release)部份，(c)有聲和無聲部份，即如圖 4中的區塊 0、1 和 2 所示。

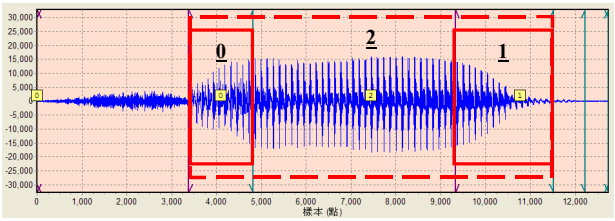

圖 4 音節分段示意圖

3.4. 諧波追蹤

在一序列的分析音框裡，MVF 的值突然起伏太大時，可能造成某個音框的諧波數較前後音框多或者是少的情形發生，當這種諧波數不連續太明顯時，就會在聽覺上感受到。因此我們進行諧波追蹤[17]，以消除這些短時間的諧波數量的不一致。

在作完有聲部份所有音框的 HNM 分析後，我們對這些音框作諧波追蹤，讓相鄰音框的諧波數不會改變得太劇烈。當前後兩個音框中的諧波數不一致時，在接下來的 50ms 中，若找得到可以相接的諧波頻率，則將這 50ms 裡的各個音框，加上此諧波，否則就將這個諧波刪除。在作諧波追蹤時，我們將 MVF 之前的諧波頻率都視爲有聲，再者因爲音框間的基頻值並不會變化太大，所以大致上相鄰音框間的諧波都會一個對一的相連，只有在靠近 MVF 的諧波頻率，可能因爲 MVF 差異太大而產生不相連的情況，因此這裡作的諧波追蹤，可以看成是對 MVF 的變化加以平滑化。

4. 國語歌聲合成

國語歌聲合成系統的製作，我們首先請一位女性在錄音室中唸出國語 409 個音節的第一聲發音，然後轉存成電腦音檔，取樣率爲 22050Hz，解析度是 16bits /sample。在電腦裡，一種音節，只有一個對應的音檔。有了原始音檔，接著我們依前述的 HNM 改進方法，分別製作分析與合成兩階段的程式。

4.1. 分析階段

音節信號分析的流程如圖 5所示。由於考慮 ADSR 式之音長伸縮作法，首先要對各音節作分段標記的處理，我們以手工對 409 個音節的起音、釋音，及有聲部份的邊界作標記，每個音節共有 3 組即 6 個邊界點，記錄在 WAVE 格式檔案的檔頭之後。

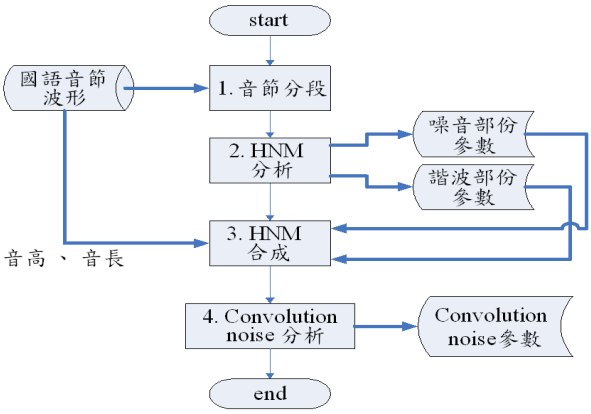

圖 5 音節信號的分析流程

然後執行 HNM 分析程式，此處使用的是如前節所述的修改過的 HNM 分析作法。對於標記爲有聲的部份，我們設定音框長度爲 512 點，且音相鄰框重疊 2/3，有聲部份的參數包括諧波參數及噪音參數兩部份。對於標記爲無聲部份的音框，則只需記錄噪音參數，所以作 FFT 轉換後，對得到的頻譜再作倒頻譜轉換，取出 30 階的倒頻譜係數作爲噪音參數。

另外，對於較短的無聲子音，或是有聲子音、母音開始的音節，由於這類音節在信號起始處通常有較不規律的波形變化，但是 HNM 是以音框爲分析單位，這些短暫的波形變化會無法被準確分析出來，所以我們將這樣的信號樣本值直接記錄下來，合成時再將這些信號樣本選擇適當的位置，與合成音相接。最後分析出 CVNS 參數，記錄至參數檔案中。

4.2. 合成階段

歌聲音節的合成流程如圖 6所示，分成如下各子節的步驟來說明。

4.2.1. 歌詞檔處理

歌詞檔的範例如表 1所示，包含下列幾個部份：曲名、每分鐘拍數以及歌詞序列，其中拍數可以控制整首歌的演唱速度。歌詞又可分爲幾個部份：(a)序號：依歌詞數逐漸累加；(b)歌詞字：依據歌詞字到字典檔中找出對應的拼音及參數檔，破音字時，可在歌詞字後放相同發音之另一字，以作爲辨別；(c)音高：以音符爲單位來記錄，當一個歌詞字爲「|」時，我們視爲轉音，即此時的音符音高要用前一音符的歌詞來唱，若音高值爲 0 時，則表示爲休止符；(d)拍數：根據演唱速度，

可以換算出一個音符要唱多久；(e)強弱：用來控制各音符的音量大小。

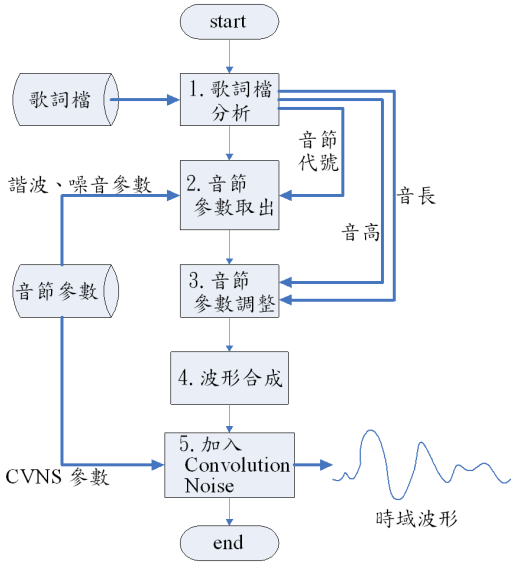

圖 6 歌聲音節的合成流程

表 1 歌詞檔範例

青春舞曲(曲名)		180(拍數/分)		
01	太	B3	1	1
02	陽	A3	1	1
03	下	F3#	1	1
04	山	G3	1	1
05	明	B3	1	1
06	朝找	A3	1	1

（歌詞序列）

4.2.2. 韻律參數決定

爲了避免拖拍的問題[8, 18]，我們希望音節中有聲的部份可以在拍子的起始處就被聽見，所以要對音節的演唱起始時間作調整。由於我們的使用的音節已標示過有聲與無聲處的邊界，只要將此處移至拍子起始點即可。

爲了模擬人在唱歌時換氣的動作，我們將歌詞的音長保留一部份作爲靜音：音符音長大於 1.3 秒時，保留音長的 25%；音長小於 1.3 秒時，保留音長的 17%作爲靜音。

在音量方面，因爲人在說話或唱歌的時候，嘴型大小與音量的大小是成正比的，所以音量要根據音節中韻母的不同來調整[14, 18]。

4.2.3. 音節波形合成

一個音符的音高由歌詞檔讀出後，就可決定出新的諧波頻率值，爲了維持音色的一致性，對於新的諧波頻率的振幅、相位數值，要在原始的頻譜包絡上，找出新諧波頻率前後相鄰的四個原始諧波的參數，作拉格蘭日內插來得到。

時間長度分配上，依 ADSR 觀念設計音長伸縮規則如下：(a)子音的長度伸縮的倍率限制在 0.6～1.2 倍之間 ；(b)起音和釋音的長度保持和原音節相同；(3)限制尾音的長度要小於延持部份長度的 1/4。如此，可避免 Sustain 以外的部份變化過度而不自然。

實際合成時，對於音長有變化的部份，我們均勻分成間隔 200 個樣本點的控制點，控制點上的參數依時間比例由原音節中的音框參數作拉格蘭日內插得到，控制點之間的樣本點上的參數，則由兩邊控制點的參數作線性內插得到，以加快信號合成的速度。最後再作 CVNS 參數的處理，以產生含有適度低頻噪音的合成音。

若音節的音長太短，可能發生 Sustain(ST)部份的長度過短，使得 Attack(AT)和 Release(RL)兩部份幾乎是直接相連，如果 AT 的結束音量和 RL 的起始音量相差太多時，聽起來就會有振幅不連續之 click。因此當合成音的音長過短時，我們會對 AT 和 RL 兩部份分別偵測各自的振幅最大值，然後將 AT 的音量調整成和 RL 相近。

4.2.4. 轉音和抖音

關於轉音和抖音的處理，由於我們是以相位增量來控制每個信號樣本點的相位變化，所以可以直接參考之前學長的作法來產生出轉音和抖音的效果[14]。

但是在 HNM 中作轉音模擬時，會因爲轉音前後音框的 MVF 不變，但因前後音符的基頻不同，而發生較低音高的音框其諧波數量會比較多，爲了不使多出來的諧波對聽覺造成影響，因此我們將諧波數較多的一方在轉音區中向另一方延伸，直到轉音區的邊界，再將其振幅逐漸降爲 0，如此所得的頻譜圖形如圖 7所示。

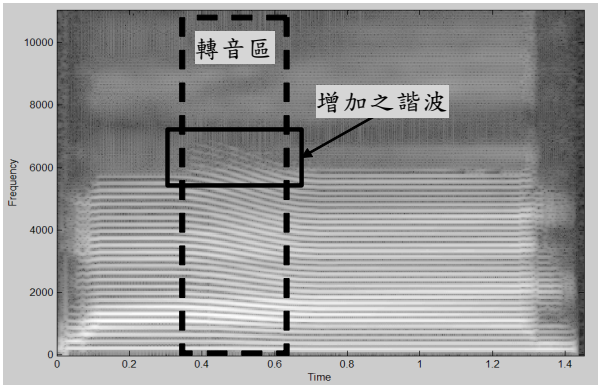

圖 7 轉音區中的諧波數量調整

5. 聽測實驗

國語歌聲合成軟體完成之後，接著可進行主觀的聽測實驗，來對合成出的歌聲信號的好壞作一個評估。其實藉由實際聽測，作者本身才發現到好幾個因素是必需考慮、加入的，例如 2.3 節之 MVF 動態設定，若不加入此因素，則某些音節的合成音信號會是失真的；若不加入 3.3 節之 ADSR 式音長伸縮，則某些合成的轉音音節會很不自然；若不作 3.4 節之諧波追蹤，則某些合成的音節會附隨有微弱的怪音成分。除此之外， 2.1 節的基頻偵測和 2.2 節的諧波參數偵測，是爲了追求準確性，愈準確的作法，理論上應是愈好。至於 3.1 節的 CVNS，聽測上作者本身已覺得不容易分辨其產生的差異，更不用說是一般的受測者。因此，我們選擇 3.2 節的相位同步和 3.3 節的 ADSR 音長作爲聽測的因素，選擇 ADSR 音長，是因爲它對自然度具有明顯的影響。

詳細說來，我們分成五種合成方式來互相作比較：(1)相位未與原始音同步，只作累加；不使用 ADSR，依線性比例調整音長。(2)相位未與原始音同步，只作累加；使用 ADSR，在延持部份調整音長。(3) 相位與原始音同步；不使用 ADSR，依線性比例調整音長。(4) 相位與原始音同步；且使用 ADSR，在延持部份調整音長。(5) 以學長之前研究的方法合成出歌聲，但不加入 MIDI 伴奏[14]；學長的方法就是弦波模型，並且 MVF 不管任何音節都固定成定值(6,000Hz)，沒有作 ADSR 音長伸縮、相位同步及 convolution noise 等處理。這五種合成方式所產生的歌聲信號，我們將會放在網頁上[4]，以供有興趣者來試聽。

聽測的曲目共有兩首，分別是節奏較快的「青春舞曲」和節奏較慢的「康定情歌」，評估的項目是自然度與清晰度兩項。自然度在於評估合成歌聲與人類歌聲的接近程度；清晰度在於評估合成出來的歌聲訊號聽起來是否清楚無雜訊，以及咬字的清晰程度。評分的範圍由最高 5 分到最低 1 分，可以打至小數點下一位。

評估的作法是由聽測者分別聽兩首歌曲，每首歌曲都有五種方式合成的音檔，我們將其中第二種方式(即無相位同步但有作 ADSR 式音長縮放)的合成歌聲，當作自然度與清晰度皆爲 3 分的參考音。另外四種方式合成出的歌聲，則以隨機的順序排列讓聽測者試聽，聽測者可以自由選取歌曲中的一段範圍反覆試聽，並與參考音作比較，以比較後的優劣程度來作評分。

試聽者一共有 15 位，評分的平均值如表 2所示。在自然度方面，因爲「青春舞曲」的節奏較快，所以有、無考慮 ADSR 的方式差異不大，相差約 0.1 到 0.2 分，而有作相位同步的方式，其自然度整體上都較不作相位同步的自然度來得高。在清晰度方面因爲基於相同的合成模型，所以有、無作 ADSR 的影響不大，但有作相位調整者較清晰，可能是受自然度的影響，其中有作 ADSR 式音長分配及相位同步的得分最高，與之前學長的合成方式比較，相差約 1 分。

由於「康定情歌」的節奏較慢，因此有作 ADSR 的效果較爲顯著，尤其是在音長較長或有轉音的音節，故在自然度上有、無考慮 ADSR 的方式，差異增加至 0.3 到 0.4 分，有作相位同步的方式，其自然度整體上仍然較不作相位同步的自然度來得高。在清晰度方面，表現也與「青春舞曲」的情況類似，其中有作 ADSR 式音長分配及相位同步的仍然是得分最高，與之前學長的合成方式比較，同樣相差約 1 分。

表 2 聽測實驗之結果

合成方式	青春舞曲		康定情歌		合成選項	
	自然度	清晰度	自然度	清晰度	ADSR	相位
1	2.9	3.1	2.7	3.0	X	X
2	3.0	3.0	3.0	3.0	V	X
3	3.4	3.4	3.2	3.2	X	V
4	**3.6**	**3.5**	**3.6**	**3.4**	**V**	**V**
5	2.7	2.5	2.6	2.2	--	--

6. 結論

我們以 HNM 爲基礎，對國語歌聲信號的合成模型與方法進行研究，得到的成果如：(a)改善 HNM 模型原本的基頻偵測、諧波參數計算、及 MVF 的決定方式，以求得更爲準確的參數值；(b)加入 Convolution Noise 之處理，以加強 HNM 在低頻部份 modeling 能力的不足；(c)加入相位同步的處理，以讓合成音能夠保持住更多原始音的音色特性；(d)使用 ADSR 觀念作音符時間長度的伸縮，以保持合成的歌聲音節的自然性；(e)加入諧波追蹤的處理步驟，以解決音框間諧波數量的跳動、不連續情況。

製作出國語歌聲合成系統後，我們進行了聽測實驗，來評估本論文的歌聲信號合成方法。從實驗的結果可知，有作 ADSR 式音長分配及相位同步的合成方式，在自然度和清晰度上都得到最高的分數，顯示本論文合成出的國語歌聲信號，其品質的確能夠獲得顯著的改進。

未來可繼續研究、改進的方向，例如：(a)對於一些在短時間內的不規律訊號，可考慮以暫態(transient)模型來分析，及以參數化的形式來表示；(b)目前所錄的音，音節間沒有前後文關係，當合成快節奏的歌曲時，會感覺歌詞被斷開，因此可考慮錄製連續的發音來分析出參數，應可改善此種情況；(c) 歌聲中的表情(expression)、感情是音樂中相當重要的部份，如果能將表情的模擬也加入歌聲合成模型之中，一定能使合成出的歌聲更加自然、動聽。

7. 參考文獻

[1] Hamon, C., E. Moulines, and F. 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，新竹，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.csie.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，台北，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，新竹，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] Yannis Stylianou, *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*, Ph.D. Dissertation, Ecole Nationale Supèrieure des Télécommunications, Paris, France, 1996.

[14] 陳安璿，*整合 MIDI 伴奏之歌唱聲合成系統*，碩士論文，國立台灣科技大學資訊工程研究所，台北，2004。

[15] 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」，*第十六屆自然語言及語音處理研討會*，台北，2004。

[16] Andersen, T. H. and K. Jensen, “Phase modeling of instrument sounds based on psycho acoustic experiments,” *In Proceedings of the MOSART Workshop on Current Research Directions in Computer Music*, pp.170-173, 2001.

[17] Lagrange, M., S. Marchand, and J.-B. Rault, “Using Linear Prediction to Enhance the Tracking of Partials,” *In Proceedings of the IEEE International Conference on Speech and Signal Processing*, 2004.

[18] 古鴻炎、陳安璿、廖皇量，「整合 MIDI 伴奏之國語歌聲合成系統」，2005 電腦音樂與音訊技術研討會(台北)，WOCMAT 2005 Session B。