基於歌聲表情分析之國語歌聲合成
Mandarin Singing Voice Synthesis Based on Expression Parameter Analyzing

古鴻炎 王如江
Hung-Yan Gu Ru-Jiang Wang

台灣科技大學資訊工程系
{guhy, M9315044}@mail.ntust.edu.tw

摘要

本論文一開始先探討、歸納和歌聲表情有關的聲學參數，接著錄製真人所唱的歌聲信號，以作表情參數的分析。為了應用所分析出的參數值，來合成出具有表情的國語歌聲，我們對先前研究的基於諧波加雜音模型的聲音信號合成系統作改進，以讓此系統可接受表情參數的控制。使用所合成出的歌聲去作聽測實驗，結果顯示表情參數的確可用以改善合成歌聲的品質，並且所合成的歌聲已可相當程度地模仿出真人歌聲的表情。

關鍵詞：歌聲合成、歌聲表情、諧波加雜音模型。

1. 前言

我們過去已有一些國語歌聲合成的研究經驗，例如以弦波模型(sinusoidal model)分析國語音節有聲部分(voiced part)的諧波參數，再據以作歌聲音節的合成[1]。由於弦波模型所合成的歌聲信號，高頻部分缺少雜音(noise)，音質顯得不自然，因此我們後來改變以諧波加噪音模型(harmonic plus noise model, HNM)為基礎，來研究國語歌聲的合成[2]，在信號的自然度和清晰度上已獲得大幅度的改進。

然而在先前的研究裡，我們尚未去作情感表達有關的信號合成的處理，因此所合成出的歌聲信號，聽起來機械味頗重而不自然。人類唱的歌曲所以會好聽，是因為歌聲裡含有豐富的情感表現。因此，對於歌聲裡的表情(expression)呈現，探究有那些聲學(acoustics)因素是和它緊密相關的，然後根據這些聲學因素去建造表情模型，以用來模仿人類歌唱的方式，可說是歌聲合成上重要的研究課題。

過去對音樂或歌聲表情的相關因素作探討的文獻並不多。在音樂的表情方面，Dixon提到作曲者或表演者可以從音符的特性上去展現表情[3]，音符的特性包含音高曲線(pitch contour)、拍子速度(tempo)、音長(duration)及動態的變化。在歌聲的表情方面，Meron和Hirose實作了抖音(vibrato)歌聲的合成[4]，先錄製具有抖音效果的候選單元，然後在合成處理時，將抖音振動的頻率、相位及振幅，調整成目標音符所需要的條件。由於抖音是歌聲表情的重要的影響因素，所以抖音的分析和合成，可以找到不少的文獻[5, 6, 7]。另外，國內雖然可以找到一些關於歌聲合成的研究文獻[8, 9, 10]，但是歌聲表情的聲學因素並不是他們研究的重點。

雖然作曲者的表情指令(例如樂譜上的漸強、漸弱符號)及歌曲的結構，也會影響到演唱者的表情、感情的表達，但如何從整首歌曲的角度來作分析，並不是本論文所要考慮的，所以我們將範圍縮小到各個音符(note)，音符是歌曲最基本的結構單位。在閱讀前人的研究文獻，以及自行觀察所錄製的歌聲信號和MIDI樂譜後，我們歸納出如下的歌聲表情之決定因素: (a)音符的音高曲線(pitch-contour), (b)音符的波形包絡(envelope), (c)音素的時長(duration)分配, (d)音符的時間落點, (e)音符演唱的飽滿程度, (f)音符的響度, (g)音符的強烈程度。

音高曲線: 音高曲線對於音符的表情有很大的影響，歌唱技巧如轉音(glissando)和抖音(vibrato)，它們的聲學上的主要作用，就是表現於音高曲線形狀的變化。許多人作抖音的分析研究，也是從音高曲線著手，據以分析出抖音的頻率、範圍 (extent)、音調(Intonation)等參數[5, 6, 7]。

波形包絡: 這裡指的是歌聲音符(國語音節)有聲部分波形外觀的曲線。波形包絡的曲線形狀一般來說會受到音節結構的影響，國語的音節結構整體來看是C_xVC_n，C_x可以是無聲(unvoived)子音、有聲子音、或是無子音，V可以是單母音、雙母音、或三母音(如/iau/)，而C_n可以是鼻音/n/、/ng/、或是無鼻音。在此用波形包絡來描述音節有聲部分的振幅大小的變化。

時長分配: 一個音節所演唱的時間長度，通常不會平均地分配給C_x、V、和C_n等組成音素，並且不同的時長分配方式，可被用來傳達不同的表情。實際上作觀察，可發現母音V通常會分配到較長的時間，因此我們再進一步把母音部分分割成A(attack)、S(sustain)、R(release)等三個片段(segment)，這是仿效電腦音樂合成裡的A、D(decay)、S、R之分割方式[11]。

時間落點: 音符起唱的時間點，會影響到歌聲的節奏，觀察國語歌曲的演唱過程，我們發現母音

之前的子音(包含有聲及無聲子音)通常會被提早唱出，這是因為母音是音節中的重音點。這樣的現象，在前人的論文裡已有提到[1][2]。

飽滿程度: 在此的定義是，實際唱出的音長相對於樂譜所規定音長的比率。如果音符都唱得很滿，整個樂句聽起來會比較黏，帶有懶散、哀傷、憂鬱的感覺；如果音符唱得較簡短，則聽起來會比較輕快，帶有愉快、歡樂的感覺。

響度: 在此所指的是，音節開頭無聲子音的振幅大小。

強烈程度(Strength): 較強烈的音符，通常其母音共振峰會比較明顯，氣音(breathiness)現象較少，而虛弱的音符則反之。

讓電腦能夠像真人一樣，以充滿表情和感情的方式，來演唱歌譜上的音符及歌詞，是我們的理想目標。因此，本論文想要探討表情參數的定義、和表情參數的分析方法，這可說是朝向理想邁進的起始步驟。

2. 表情參數分析

此節說明我們作表情參數分析的方法。不過，在作分析之前，必需先錄製真人所演唱的歌聲信號。當分析出各個歌聲音符的表情參數值後，這些參數值就可以存檔，然後用於控制歌聲信號的合成處理。

2.1 真人歌聲錄音

我們邀請了一位喜好唱歌並且可配合作業的女性，到本系的專業隔音室(Acoustic System RE-242)，來進行真人歌聲的錄製，信號樣本的規格是，取樣頻率22,050Hz，16bits之樣本寬度。

目前的流行歌曲大多有MIDI格式的電子樂譜，因此我們選用MIDI作為歌聲分析的資訊來源。在錄音時，我們將背景音樂搭配原MIDI所設定的音源，以較小的音量播給演唱者監聽(相對於主要旋律)，這可讓演唱者了解現在該演唱的段落，另外也可讓歌手清楚音樂的節拍及可演唱的音域。我們總共錄了6首流行歌曲的演唱。

2.2 歌聲音符切音

我們可根據MIDI樂譜檔裡的Note On、Delta Time及Note Off等資訊來取得各個MIDI音符的起始時間及音長，然後依據MIDI音符的起始時間及音長，來切出真人所演唱的各個歌聲音符(音節)。不過先決條件是，真人演唱時必須抓準MIDI音符的節拍。

輸入歌詞文字檔來取得各個歌詞音節的拼音時，在無轉音的情況，一個歌詞音節就分配一個MIDI音符，而在有轉音的情況，一個歌詞音節就要分配多個MIDI音符。這裡作轉音偵測的規則是，當前一個MIDI音符的Note-Off出現之前，就先出現下一個MIDI音符的Note-On，就表示這兩個MIDI音符之間有轉音的關係。由於轉音時會發生一個音節(歌聲音符)對應多個MIDI音符之情況，所以我們以音節為單位來分析、記錄表情參數值。

2.3 音高曲線和波形包絡分析

音高曲線和波形包絡的估計，都是以音框為單位來作分析，音框大小設為20ms，相鄰的音框則重疊四分之三。一序列音框分析出的參數值，就直接儲存。

在一個音框裡，我們估計音高的方法是，使用自相關(auto-correlation)函數$R(k)$和絕對振幅差(Absolute Magnitude Difference)函數$M(k)$來作組合[12]，也就是以$R(k)$和$M(k)$的比值來估計基頻值，詳細公式為

$$P = \arg\max_{P_{\min} \le k \le P_{\max}} \frac{R(k)}{M(k)+1} \qquad (1)$$

我們設定歌聲的音高在 60Hz 到 500Hz 之間，所以只對此頻率範圍內有可能的 k 值作計算，然後取出最大者作為此音框的基週估計值。

關於有聲、無聲的判斷，滿足以下任一條件者就判定為無聲音框：(a)音框能量太低者判定為靜音音框，當然是無聲；(b)最大的自相關函數值 $R(k)$ 小於音框能量的 1/4 者視為無聲；(c)最大 AMDF 值除以最小 AMDF 值小於 2.1 者視為無聲。

關於波形包絡的估計，我們以計算各音框的rms振幅值來代表，如此在合成歌聲信號時，可以分析音框的振幅和合成音框振幅的比例，來調整合成波形的振幅大小。

2.4 端點和邊界點分析

我們先依據音節拼音開頭的字母將音節分為四類，亦即(I)爆破音開頭, 如/b,p,d,t,g,k/；(II)無聲子音開頭，如/c,f,h,s,z/；(III)有聲子音開頭，如/l,m,n,r/；(IV)直接母音開頭，如/a, i, u/。接著，依音節的類別，使用不同的聲學量測方式，去偵測音節的左邊端點，而右邊端點的偵測方式，都是以有聲(有週期性)變成無聲(無週期性)的邊界點，作為端點，這是因為國語音節的韻母部分都是有聲的。偵測程式處理完後，會將端點作繪圖顯示，然後允許操作者作手動更正。

關於(I)類音節，由於爆破音音素的前面，會出現一小段無聲的區間，因此我們可計算連續的音框的短時能量(short-time energy)[13]，再依短時能量的低點找出端點。關於(II)類音節，我們先計算連續各音框的過零率(zero-crossing rate)[13]，再訂定一個適當的門檻值，以找出無聲子音的範圍。

關於(III)類音節，由於以有聲子音開頭，可能造成信號波形跟前一個音節的韻母相連，因此我們應用有聲子音的頻譜變異數值(spectral variance)會比母音的小的特性，先計算連續各音框的頻譜變異數值，再去偵測頻譜變異數超過門檻值的音框，設

為音節的左端點。頻譜變異數的計算公式為

$$V_t = \frac{1}{N}\sum_{k=0}^{N-1}\left(|X_t[k]|^2 - M_t^2\right) \quad (2)$$

$$M_t = \frac{1}{N}\sum_{k=0}^{N-1}|X_t[k]| \quad (3)$$

其中 $X_t[k]$ 表示第t音框作快速Fourier轉換後的第k個頻譜係數。關於(IV)類音節，它的信號波形也可能跟前一個音節的韻母相連，因此我們使用和(III)類音節一樣的偵測方式。

分析出一個音節的左右端點之後，接著進一步去偵測音節母音部分兩邊的邊界點。對於有聲子音後面接母音的情況，或是母音後面接鼻音的情況，都可以先計算出各音框的頻譜變異數，再去作邊界點的偵測。就算程式自動偵測出的邊界點有誤差，仍可再以手動方式作更正。

接下來是進行母音部分ASR片段邊界點的偵測，由於目前尚無有效且通用的演算法，我們在此使用一個簡單方法作大略的標記，再以手動方式作調整。所用的方法是，依據母音部分的波形包絡(即各音框的rms能量值)，由左至右尋找，將rms值大於門檻值Ea的音框設定為起音(attack)和延音(sustain)的邊界，而再次找到rms值小於Ea的音框，就設定為延音和釋音(release)的邊界。

2.5 表情參數值求取

以前述方式作程式自動偵測及手動調整之後，就可確定一個歌聲音節的左右端點、母音部分的端點、及ASR片段的邊界點。之後，我們就可以計算出歌聲音節C_x、V、C_n等部分的時長；歌聲音節的時間落點，即C_x的起始時間；歌聲音節的飽滿程度和響度。在此，我們尚未作歌聲音節的強烈程度的分析與合成應用。

3. 歌聲信號合成

我們使用先前發展的HNM (harmonic plus noise model)信號分析與合成程式作基礎[2]，再加以改進，以用於作富含表情之歌聲信號的合成。首先要錄製國語的408種音節的發音，以用於分析出各個音節的HNM信號模型的參數值[14]，在此我們邀請了另外一位女性到隔音室錄音，錄音設備和信號規格就如2.1節所說的。建立各個音節的HNM模型後，接著我們需考慮如何利用之前分析出的表情參數來控制信號的合成，例如控制音高曲線、控制歌聲音符各片段的時間分配、及作波形包絡的調整。

3.1 HNM簡介

當對一個有聲的語音信號作頻譜分析後，信號所具有的週期性特性，可以在頻譜圖上很容易地觀察到：頻譜中曲線的峰值會出現在基頻和其倍數的頻率值上。但是這些峰值的出現並不會含蓋整個頻率域的範圍(0Hz到取樣率的一半)，而是會分佈在某個頻率值之前。以圖1為例，圖中上半部的頻譜圖，是由下半部波形上標示的音框分析得到。頻譜上振幅峰值的分佈，大約在5000Hz之前峰值間的距離較為相等，以小方塊標示，而5000Hz之後峰值間的距離變得較不固定，甚至是峰值不明顯。HNM就是依據信號的這種特性，將信號分解為諧波$h(t)$和雜音$n(t)$兩個部份來表示的模型。

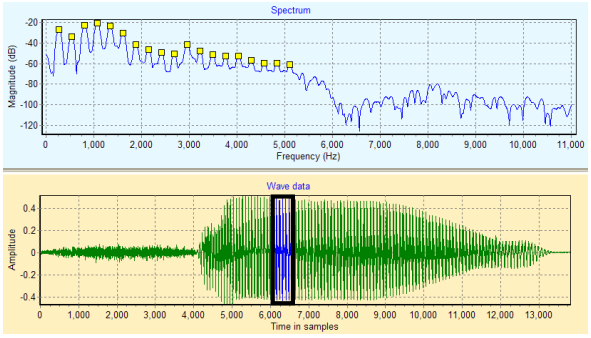

圖1　音節/ㄕㄚ/的一個母音音框及其頻譜圖

對於語音信號中一個有聲(voiced)的音框，HNM會先依信號的頻譜計算出最大有聲頻率(Maximum Voiced Frequency, MVF) $F_m(t)$，最大有聲頻率$F_m(t)$是一個隨著音框在改變的值，依據$F_m(t)$，頻率值小於MVF的頻譜部份，在HNM中視為諧波部份，而頻率值大於MVF的頻譜部份，則視為雜音部份。如此，一個音框的語音信號$s(t)$就視為是由諧波$h(t)$和雜音$n(t)$兩部份的信號值作相加而得到，即 $s(t) = h(t) + n(t)$。諧波部份$h(t)$，則可以看成是由頻率值有倍數關係的數個弦波所組成，其公式如下：

$$h(t) = \sum_{k=1}^{K(t)} a_k(t)\cos\left(\phi_k(t)\right) \quad (4)$$

其中 $a_k(t)$ 和 $\phi_k(t)$ 表示在時間為t時，第k個弦波的振幅和相位，$K(t)$ 則代表時刻t時諧波部分所包含的弦波個數。對於有聲的信號音框的頻譜，頻率值大於MVF的部份，HNM視為是雜音，而對於一個無聲(unvoiced)的音框，則將MVF之值視為0，即整個頻譜都是屬於雜音部份。

HNM對於雜音的一種模式化方式，是把雜音看作是頻率間隔一直是固定值的諧波信號，而去求各諧波上的振幅值，實作上將頻率間隔固定為100Hz，而且只需振幅之參數值，相位值則不記錄。求出100Hz及其各倍頻諧波上的振幅值之後，對這些振幅值作倒頻譜(cepstrum)分析，然後以少量的倒頻譜係數，來代表平滑過的雜音頻譜。

3.2 片段線性時間對映

作HNM信號合成時，我們採取片段線性的時間軸對映(mapping)方式，來對應單獨錄音的音源音節上的片段(segment)，至欲合成的歌聲音節上的片段，以提升合成歌聲的流暢性。而這也意味著，我們必需先對音源音節、以及對真人所唱的歌聲音

節，去標記出音節裡各區段的邊界點，亦即標記有聲子音的端點、母音部分ASR片段的的邊界點，以程式作自動標記的方法，已在2.4節裡說明。

在此令單獨錄音的音源音節的開頭有聲子音的長度為x1、母音起音、延音、釋音長度分別為x2、x3、x4，而結尾有聲子音的長度為x5；此外令真人所唱的歌聲音節所分析出的數值是，開頭有聲子音的長度為y1、母音起音、延音、釋音長度分別為y2、y3、y4，而結尾有聲子音的長度為y5。依據(x1, y1), (x2, y2), …, (x5, y5)的時間長度數值，就可以建立時間軸的片段線性對映關係，如圖2所示。先前我們研究HNM的信號合成處理，y1、y2、y3、y4、y5之長度值，是由程式自動決定，而在本論文中，為了模仿真人所演唱的歌聲音符，所以它們的數值是從表情分析用的歌聲音符作分析而取得的。

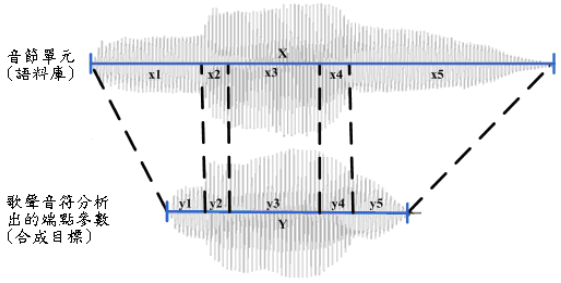

圖2 音源音節與歌聲音節之間的區段對應

3.3 控制點及其HNM參數值

為了降低信號合成時求取HNM模型參數值的計算量，我們採取設置“控制點”[11]的方式去作音高(基頻)及HNM諧波參數(頻率、振幅、相位)和雜音參數的求取動作，也就是在合成歌聲的時間軸上，均勻地每間格100個樣本點(4.54ms)設置一個控制點，當碰到控制點時，才比較精細地去計算HNM模型參數值，而在其它的樣本點上，就只用簡單的線性內差方式，依據左右相鄰的兩個控制點去計算HNM參數值。

佈放控制點之後，對於各個控制點，就可依其所在的時間位置Ty，作如圖2所示的片段線性之對映，找到音源音節時間軸上的對應位置Tx，Tx是以音框為單位計數，依Tx可得知兩個和Ty對應的分析音框，即編號為$\lfloor Tx \rfloor$和$\lfloor Tx \rfloor+1$之音框。然後再依據這兩個音框所分析出的HNM參數值去作內差，就可求得控制點Ty上的HNM模型參數值，較詳細的內插作法，可參考我們先前的研究論文[2]。

3.4 控制點音高調整

作HNM信號合成時，每個控制點可分別去設定它的音高，因此我們可依據真人歌聲音節所分析得到的音高曲線來設定各控制點上的音高。真人歌聲音符作表情分析時，記錄的是每個音框的基頻，並且所用的音框大小為20ms(441個樣本點)，而合成時所用的控制點間隔為100個樣本點，因此我們必需將分析得到的音高曲線作拉格朗日(Lagrange)內插，以求出各個控制點上的音高，拉格朗日內差內插的公式為：

$$P(x)=\sum_{j=1}^{4}P_j(x)\ ,\ P_j(x)=y_j\prod_{\substack{k=1\\k\neq j}}^{4}\frac{x-x_k}{x_j-x_k} \qquad (5)$$

其中x為一個控制點的時間位置經過線性時間比例對應到真人所唱音高曲線的時間軸上的時間值，依x可找出x兩邊附近的時間點x_j，並取得該時間點上的頻率值y_j，然後就可依公式(5)來求得控制點上的音高$P(x)$。

以HNM作信號合成時，如果只是去改變諧波的頻率值，而不對諧波的振幅值作更動，即如圖3所示的情況，則幅峰(formant)頻率值也會受到相同倍率的改變，其結果是音色也會受到影響而不能保持一致性，在圖3的例子是，大人的音色變成了小孩的音色(因為幅峰頻率值變大)。

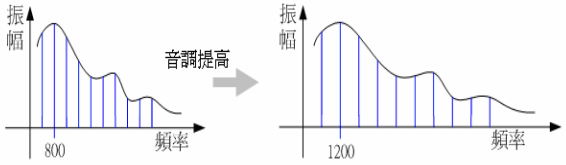

圖3 諧波頻率改變而振幅不變之頻譜曲線變化

因此，調整一個控制點的音高時，必需要讓頻譜包絡保持不變[11]，也就是頻率值被調到$\tilde{F}_k$的第k個諧波的振幅$\tilde{A}_k$，它的值必需從原始音高之諧波振幅值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F_1，F_2，F_3，…中找出最靠近$\tilde{F}_k$且比$\tilde{F}_k$小的頻率值，令找出的是F_j，接著，就以F_{j-1}，F_j，F_{j+1}，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(如公式(5)之形式)，以求出頻率值$\tilde{F}_k$上所應對應的振幅值$\tilde{A}_k$。

3.5 波形包絡調整

波形包絡的調整，就是調整信號樣本的振幅，振幅的調整分成有聲部分和無聲部分分別考慮。在有聲信號的部分，我們先對合成好的波形計算各音框的rms能量，再依能量比例來作振幅的調整，能量比例的定義是$R_t=E_t/\hat{E}_t$，其中R_t表示第t個音框的調整比例，$\hat{E}_t$為合成波形第t個音框的rms能量，E_t為真人歌聲第t個音框所分析出的能量。至於音框內信號樣本的振幅調整公式為

$$\hat{x}_t[m]=x_t[m]\cdot R_t,\quad m=I_t-\sigma,\ I_t-\sigma+1,\ \ldots,\ I_t+\sigma-1 \qquad (6)$$

其中$\hat{x}_t[m]$表示調整過後的信號樣本值，$x_t[m]$表示調整前的樣本值，I_t為第t個音框的編號(音框中心點)。由於相鄰音框之間有四分之三的重疊，因此必需設定m的範圍，從$I_t-\sigma$到$I_t+\sigma-1$，其中σ為音

框大小的八分之一，即55個樣本點。

在無聲信號的部分，信號主要是雜音成分，因此我們只根據信號的最大振幅值來調整。由於表情參數中包含了歌聲音符無聲部分的最大振幅值，因此可先找出合成波形無聲部分的最大振幅值，然後將無聲部分的信號樣本乘上振福比例就是作無聲部分的振幅調整。

3.6 音符串接

在合成出各個歌聲音符的波形後，接著可依據表情參數所記錄的各音符的時間落點，把音符串接成一個樂句，進而連結成一個段落的歌聲。

例如圖4是合成完畢但還未串接的獨立音符，存放在記憶體中，時間軸上P1、P2、P3及P4代表MIDI音符所記載的各音符的時間起點位置。而在圖5中，各音符已經經過串接的處理，其中note1及note2的起始點分別對應到音符時間點P1和P2上；note2的尾端和note3前端有重疊的部分，那是因為兩個音節之間有連音現象，因此要將前後音符的音長拉長，來讓有聲部分重疊以合成出連音現象；note3前端有作時間的拉長，因此會超前音符時間點P3；而note4的起始時間也超前於音符時間點P4，那是因為note4是由無聲子音開頭，所以將note4的母音部分起始位置對應到P4，而呈現超前唱的情形。

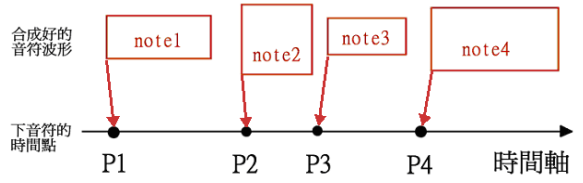

圖4 串接前的獨立音符

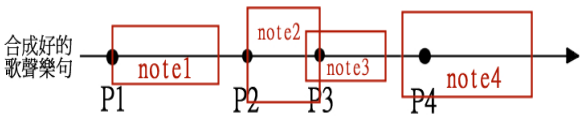

圖5 串接後得到的樂句

4. 聽測實驗

聽測實驗所用的音檔是由三種合成方式產生，合成方式如表5-1所列，方式I使用了第2節所說的表情參數來作歌聲的合成；方式II則未使用表情參數而只使用MIDI歌譜及歌詞來作歌聲合成；方式III則是直接播放真人演唱的歌聲。

表1 合成方式

方式	I	II	III
使用表情參數	V	X	真人演唱歌聲

聽測實驗要測試的是，合成歌聲的表情逼近真人歌聲表情的程度，實驗前先向聽測者說明，待評分的歌聲是模仿真人演唱的合成歌聲，模仿的歌聲表情因素包括: 節奏，抖音，轉音，輕重音等。第一個被播放的音檔為真人演唱的歌聲，令其評分為滿分5分，也就是表1裡的方式III；第二個被播放的音檔為未使用歌聲表情的合成歌聲，令其評分為0分，也就是表1裡的方式II。依據前述的評分標準，再讓聽測者試聽方式I所合成的歌聲，聽測者可依據模仿的歌聲表情的好壞程度，給予0到5分之間的評分。

我們邀請了15位聽測者，聽測者來自系上實驗室的同學及一般民眾，實驗室同學對於聽測實驗較有經驗，而一般民眾的能力則無法確定。我們準備了兩首歌曲來作表情逼近度的評估，一首是快歌而另一首是慢歌，歌詞內容如表2所列，聽測後計算出的平均分數分別是3.53分(慢歌)和3.47分(快歌)。這樣的評分，可說明我們所合成出歌聲，其歌聲表情的展現，在一般聽測者的判斷中，已經是相當程度地靠近真人歌聲的表情。

有興趣的讀者可連線至http://guhy.csie.ntust.edu.tw/~asriver/achievement.htm，來試聽、比較真人所演唱的歌聲和本研究所合成的歌聲；另外，也歡迎有興趣者連線至http://guhy.csie.ntust.edu.tw/syn–sound.html，來試聽我們先前的研究所合成出的歌聲。

表2 聽測歌曲之歌詞

張惠妹 – 姊妹 副歌部份(中板)	王菲 – 執迷不悔 橋段部份(快歌)
我想你一定知道，你是我的姊妹，你是我的背貝，喔耶，不管相隔多遠。	我不是你們想的如此完美，我承認有時也會辨不清真僞，並非我不願意走出迷堆，只是這一次，這次是自己而不是誰。

5. 結語

過去的歌聲合成的研究，我們並未找到針對歌聲表情的聲學因素作詳細探討的文獻，因此本論文嘗試去歸納歌聲表情有關的幾個重要的聲學因素，也就是(a)音符的音高曲線, (b)音符的波形包絡, (c)音素的時長分配, (d)音符的時間落點, (e)音符演唱的飽滿程度, (f)音符的響度, (g)音符的強烈程度等等。

接著我們錄製真人演唱的歌曲，去分析出各音符的歌聲表情的參數，然後依據所分析出的表情參數，去改進HNM為基礎的信號合成方法，以讓它可以接收表情參數的控制，再實作成一個可以模仿真人歌聲表情的國語歌聲合成系統。

聽測實驗的結果驗證了，使用表情參數來控制歌聲的合成，可以合成出更好聽、品質更好的歌聲，並且近似真人歌聲表情程度的評分，可達到滿分5分之3.5分。雖然距離真人演唱歌聲的品質還有一段差距，但是基於分析出的表情參數來合成歌聲

的構想已經可以實現了。

我們的理想是，能夠讓電腦像真人一樣有表情、有感情地演唱歌譜上的音符及歌詞，因此表情參數的分析可說是一個朝向理想邁進的起始步驟。未來可對真人歌手錄製大量的演唱歌曲，去分析其歌聲的表情資訊，然後據以建立歌聲表情的模型，如使用類神經網路之模型，如此就可以用模型來自動產生出表情參數值，而使電腦可以自主地唱出近似真人表情的歌聲。

參考文獻

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

[2] 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」，*WOCMAT 2006 國際電腦音樂與音訊技術研討會*(台北)，session 2 (音訊處理I)，2006。

[3] S. Dixon, "On the Analysis of Musical Expression in Audio Signals", in *Storage and Retrieval for Media Databases*, 2003.

[4] Y. Meron and K. Hirose, "Synthesis of Vibrato Singing", *IEEE ICASSP*, 2000.

[5] E. Prame, ”Measurements of the vibrato rate of ten singers”, *J. Acoust. Soc. Am.*, Vol. 96**,** pp. 1979-1984, 1994.

[6] E. Prame, ”Vibrato extent and intonation in professional western lyric singing”, *J. Acoust. Soc. Am.*, Vol. 102, pp. 616-621, 1997.

[7] I. Arroabarren, et al., ”Measurement of vibrato in lyric singers”, *IEEE instrumentation and measurement technology conference*, pp. 1529-1534, 2001.

[8] 歐婉菁，*合成歌聲*，碩士論文，國立台灣大學資訊工程研究所，2003。

[9] Tzu-Ying Lin, *A Corpus-based Singing Voice Synthesis System for Mandarin Chinese*, Master thesis, Dept. of Computer Science, National Tsing Hua University, 2004.

[10] 詹詩涵，*基於音高調節之歌聲合成系統*，碩士論文，國立清華大學資訊系統與應用研究所，2005。

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

[12] 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。

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

[14] Y. Stylianou, "Applying the Harmonic plus Noise Model in Concatenative Speech Synthesis", *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 1, pp. 21-29, 2001.