基於HMM與GMM模型之歌聲合成與音色轉換
HMM and GMM Based Singing Voice Synthesis and Timbre Conversion

古鴻炎
台灣科大資訊工程系
guhy@mail.ntust.edu.tw

簡延庭
台灣科大資訊工程所
m9915069 @mail.ntust.edu.tw

摘要

本論文嘗試結合 HMM 頻譜模型與 GMM 音色轉換模型，以建造一個具有歌者音色轉換功能之國語歌聲合成系統。關於頻譜係數的分析，我們使用 STRAIGHT 來求取較為準確的頻譜包絡及音高資訊，然後將各音框的頻譜包絡換算成 DCC 係數；接著，以自行發展的程式來訓練 HMM 頻譜模型與 GMM 音色轉換模型。在合成階段，HMM 與 GMM 兩種模型都遇到頻譜過度平滑的問題，我們於是研究音段變異數之作法，使得過度平滑的頻譜得到改善。關於歌者音色的轉換，我們研究了三種轉換方法，分別是基本音色轉換法、使用 GMM 之相對振幅轉換法以及不含 GMM 之相對振幅轉換法。完成系統的製作後進行聽測實驗，其結果顯示，歌唱語料訓練的 HMM 模型比說話語料訓練的 HMM 模型較能夠合成出有共鳴感的歌聲；此外，音色轉換的聽測結果是，基本音色轉換法所轉換出的歌聲，在音色與聲音品質上，都比其它轉換方法的效果來得好。

關鍵詞：歌聲合成、音色轉換、HMM 頻譜模型。

1. 前言

大部分的人喜歡唱歌，也樂於聆聽好聽的歌聲，此外，許多人會想要聽聽自己心目中的名人的歌聲，但是可能沒有管道去獲得此位名人的歌聲(例如：張小燕)，或是這一位名人已經不在這世上了(例如：鄧麗君)。因此，本論文嘗試研究歌聲合成結合音色轉換的技術，以便合成出具有某一特定人音色的歌聲。

歌聲合成的方法有許多種類，包括：(a)基於樣本之作法(sample-based approach)、(b)統計式作法(statistical approach)、(c)弦波模型式作法。單元選擇就是一種基於樣本之作法，此種方法合成出的歌聲品質非常好，但是它需要龐大的語料庫，以便可以從語料庫挑選出適合的聲音來作串接。然而語料庫不可能包含各種可能的聲音變化，因此實際的作法是，先從語料庫中找到一個最相似的聲音，再作一些聲音特性的調整，以便讓此聲音符合需求[11]。目前市面上很有名的歌聲合成軟體 VOCALOID [21]，也應用了此一方法。

統計式作法的一個典型代表就是隱藏式馬可夫模型(hidden Markov model, HMM)，使用HMM來掌握頻譜演進的方式，再據以作歌聲合成[12]。它的好處是，不需要龐大的語料庫，對於聲音特性(例如：音高、音長)的修改是容易的，合成出來的聲音也比較穩定，不過其缺點是頻譜包絡(spectral envelope)有過度平滑(over smoothing)的問題，使得合成出的歌聲有悶悶的感覺。Tokuda等人就是使用HMM來研究歌聲合成，已把研究成果以網頁作呈現，可供測試使用[18]。

諧波加雜音模型(harmonic plus noise model, HNM)[20]屬於弦波模型的延伸，它除了考慮和諧的弦波部份外，還加上了高頻帶的噪音部分，使合成出來的聲音可以和原始聲音的音色更相似。過去廖皇量的研究[3]，就是以HNM為基礎，發展出一個能夠合成出清晰歌聲的合成方法。此外，林正甫的研究[1]，加入歌聲表情中的一個重要因素"抖音"，並研究其分析方法及建立模型，去控制HNM來合成出國語的歌聲信號，使得合成出來的歌聲信號，在自然度上得到大幅度的改進。

依據前面的說明，我們決定採取HMM作國語歌聲的合成，不過我們將以前人的成果作基礎[4, 6]，自行發展HMM的訓練程式及HMM歌聲合成的相關程式，而不使用HTK (hidden Markov model toolkit) [17]，以便將來建立一套可線上即時操作的歌聲合成系統。

在頻譜係數的分析上，我們使用STRAIGHT求得較準確的頻譜包絡及音高資訊，並修改前人所發展的頻譜係數估計程式[5]，使其依據STRAIGHT求得的頻譜包絡來換算出DCC (discrete cepstrum coefficient)係數。在語料分類上，我們不僅考慮了聲、韻母分類及文脈分類，也將語料依不同音高，再細分為高音、中音及低音等3種子類，以解決合成歌聲之音色不一致的問題。關於只有說話語料之目標歌者的音色轉換，我們研究三種轉換方法，分別是基本音色轉換法、使用GMM (Gaussian mixture model)之相對振幅轉換法及不含GMM之相對振幅轉換法。在歌聲信號合成方面，我們使用前人發展的HNM信號合成模組[5]來產生歌聲信號樣本。在合成階段，由於HMM模型及GMM音色轉換模型所產生的頻譜係數，都有頻譜過度平滑的問題，於是我們參考了全域變異數的觀念[19]，提出音段式變異數的作法對產生的頻譜係數作調整，以改善頻譜過度平滑的問題。

2. 模型訓練

訓練階段的主要處理流程如圖1所示。我們先將歌聲資料庫的歌曲切成一句一句的樂句，再作聲母、韻母之標音、切音；之後使用STRAIGHT對每一個切割出的音節作分析而得到各音框的頻譜包絡資料及音高資料；接著，我們修改蔡松峯發展的離散倒頻譜係數(DCC)之估計程式[5]，去估計出歌聲音節各音框的DCC係數，並計算一階差分係數。

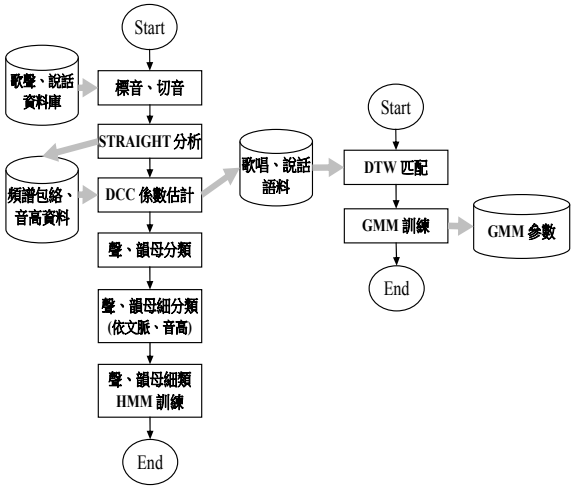


圖 1 訓練階段之主流程

關於HMM模型的訓練，我們先依據標音資訊，將各音節之一序列音框的DCC係數切分為聲母和韻母兩部分，並且把同一種聲、韻母的DCC收集在一個目錄裡。接著，再依據文脈及音高資訊作更細的聲、韻母分類，之後再對各個聲、韻母細類去訓練出一個對應的HMM。在GMM音色轉換模型訓練上，我們先對平行語料作DTW音框匹配，再進行GMM模型訓練。

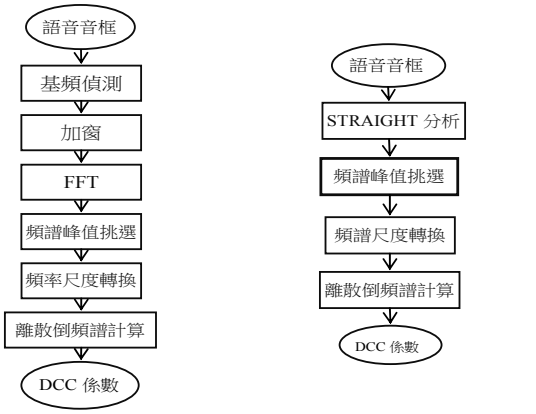
2.1 錄音、標音

歌聲語料錄音內容是選自校園民歌回顧[7]裡的44首歌曲，經過分割整理可得到787個樂句，音節總數為5,882個。此份歌詞我們請一位錄音者以說話及唱歌的方式，各錄製一份。當錄音完成後，再將其轉換成取樣率22,050Hz，解析度16bits/sample的音檔。

為了擷取出錄音樂句裡各音節的信號，我們先作標音的動作，就是在時間軸上標示出各個音節的左右邊界與音標。我們先使用HTK (HMM tool kit)來作音節forced alignment的處理，接著再使用Wavesurfer軟體來對音節邊界作人工之微調、更正，然後就可依標音資訊來進行音檔之切音。

2.2 STRAIGHT 分析及 DCC 係數估計

STRAIGHT分析法[9]是由日本學者H. Kawahara在1997年提出，他以channel vocoder技術為基礎，去作參數分析方法的改良。本論文採用了STRAIGHT的軟體來分析音檔，以求取各音框較為準確的頻譜包絡與基頻資料，然後修改前人發展的DCC估計的程式模組[5]，去求取各音框的DCC係數。修改的部分為：圖2(a)裡的基頻偵測、加窗及FFT等方塊以STRAIGHT分析作取代；依據STRAIGHT分析出的頻譜包絡資料，在挑選頻譜峰值的作法上作了調整。圖2(a)之原始DCC估計程式，頻譜包絡其實是逼近出來的，所以誤差應會較大。



(a)原始流程　　(b)結合STRAIGHT之流程

圖 2 修改前、後之 DCC 係數估計的流程

2.2.1 修正的頻譜峰值挑選方法

當處理的音框為有聲時，我們直接把MVF(maximum voiced frequency)設為5,500Hz，再依據MVF，把該音框的頻譜包絡切割成低頻的諧波(harmonic)部分和高頻的雜音(noise)部分。對於諧波部分，先將STRAIGHT分析出的基頻值F_0除以一個整數m，使得自定的虛擬基頻值$f_v = F_0 / m$介於40Hz至80Hz之間，然後在頻率範圍0Hz至5,500Hz內，記錄 f_v 的倍數頻率上之振幅值及其對應的頻率值。至於雜音部分，則直接去偵測頻譜峰點的位置，再記錄其振幅值及對應的頻率值，並更進一步對挑選出的峰點去檢查，如果一峰點與其左邊或右邊峰點的距離大於645Hz，我們會在兩個峰點間尋找其波谷並且把它選為頻譜代表點。至於無聲的音框，我們還是採取和有聲音框相同的頻譜峰點挑選方法，不過在這裡就把 f_v 直接設為50Hz。

2.2.2 頻譜逼近誤差

DCC係數的個數p必須先固定，然後才能解p個聯立方程式來求得DCC係數。為了準確地逼近大多數的頻譜包絡形狀，以避免音質下降及保持音色的一致，我們認為加大p值是必要的。那麼p值應設為多少？在此，我們以實驗的方式來探討離散倒頻譜階數和頻譜包絡之逼近誤差的關係，實驗裡使用的誤差量測公式如下：

$$Es = \frac{1}{Nr}\sum_{t=0}^{Nr-1}\left[\frac{1}{L}\sum_{k=1}^{L}\left|20\log_{10} a_k^t - 20\log_{10} S(t, f_k)\right|\right] \quad (1)$$

其中Nr表示歌聲音框的總數，a_k^t 表示第t個音框裡的

第k個頻譜峰點的振幅，$S(t, f_k)$表示第t個音框以離散倒頻譜所逼近出的頻譜包絡。隨著階數的增加Es值會明顯地下降，直到階數高於80時，Es值下降的幅度才趨於緩和，因此我們決定把p值設為80。

2.3 語料分類

在訓練階段我們作分類的方式有2種，一種是聲、韻母的分類，另一種是依文脈、音高作細分類。這2種分類動作在訓練階段是依序進行的，之後才分別對各類別去訓練出對應的HMM模型。

當把音節為單位的音檔做完DCC係數估計後，接下來把求得的一序列音框的DCC係數以聲、韻母為單位作分類、與切割。國語的聲母可分為21類，韻母則可分為36類，所以共可分為57類。

作完聲、韻母分類後，接著針對各個聲、韻母類別再作一次依文脈、音高的細分類。依文脈作分類，主要是考慮國語語句的發音是與前後音節有關係的，也就是前後文相依性(context dependency)，但是如果完全依聲、韻母的類別組合去作文脈細分類，則會發生類別組合數過於龐大的問題，因此我們退而求其次，在考慮一個聲、韻母的文脈組合時，改成考慮相鄰音素的發音口型關係，也就是依發音位置把同一位置的聲母做合併歸類的動作，而韻母則依串接點的發音口型作歸類。如此我們得到聲、韻母的發音歸類表如表1 所示，左邊為英文拼音，右邊為對應的注音符號。

表 1 聲、韻母本身之歸類

聲母	**bx**	b /ㄅ/		p /ㄆ/		f /ㄈ/	
	bv	m /ㄇ/					
	dx	d /ㄉ/		t /ㄊ/			
	dv	n /ㄋ/		l /ㄌ/			
	g	g /ㄍ/		k /ㄎ/		h /ㄏ/	
	zz	zz /ㄐ/		cc /ㄑ/		ss /ㄒ/	
	z	z /ㄗ/		c /ㄘ/		s /ㄙ/	
	zhx	zh /ㄓ/		ch /ㄔ/		sh /ㄕ/	
	zhv	r /ㄖ/					
韻母		o /ㄛ/	er /ㄦ/	e /ㄜ/	u /ㄨ/	yu /ㄩ/	i /ㄧ/
		a /ㄚ/	n(g) /ㄥ/	ii			

作完文脈分類之後，我們再依所唱音符的音高，將每一個文脈類別裡的成員細分為高音(high)、中音(medium)及低音(low)等三種子類別。為何要再做一次音高的分類？因為我們發現在不同音高範圍裡的同一音節所畫出來的頻譜包絡形狀會有不小的差異，並且各文脈類別作完HMM訓練後，再合成出來的歌聲，有不少音節的音色會明顯脫離真人所唱歌詞的音色。如何決定高音、中音及低音的音高範圍？在此我們先將各韻母類別的所有音框作音高值的統計，再觀察統計後的結果，最後決定分成如表2所示的三種音域範圍。

表 2 三種音域的音高範圍

low	F2 - D3 (87.307Hz~146.83Hz)
medium	D3 - A3 (146.83Hz~220Hz)
high	A3 - A3+150Hz (220Hz~370Hz)

決定三種音域的頻率範圍後，接著就可以把韻母及有聲聲母/m/、/n/、/l/和/r/的所有文脈類別再作一次音高分類。表3就是一個樂句分類的完整例子。

表 3 樂句"天上星星數不清"作分類的例子

字(音標)	音名(音高 Hz)	STEP 1 聲、韻母分類	STEP 2 依文脈分類	STEP 3 依音高細分類
天(tian)	G3 (196Hz)	t	t_sil_i	--
		ian	ian_dx_zhx	ian_dx_zhx_med
上(shang)	G3 (196Hz)	sh	sh_n(g)_a	--
		ang	ang_zhx_zz	ang_zhx_zz_med
星(sing)	A3 (220Hz)	s	ss_n(g)_i	--
		ing	ing_zz_zz	ing_zz_zz_high
星(sing)	G3 (196Hz)	s	ss_n(g)_i	--
		ing	ing_zz_zhx	ing_zz_zhx_med
數(shu)	E3 (164.8Hz)	sh	sh_n(g)_u	--
		u	u_zhx_bx	u_zhx_bx_med
不(bu)	G3 (196Hz)	b	b_u_u	--
		u	u_bx_zz	u_bx_zz_med
清(cing)	A3 (220Hz)	c	cc_u_i	--
		ing	ing_zz_sil	ing_zz_sil_high

註：ian_dx_zhx_med（本身類別_前串接的類別_後串接的類別_音高範圍）

2.4 HMM 模型訓練

考慮到一個音節的發音經過分類後，每一類別內的發音個數已經不多，因此參考前人的研究成果[13, 15]，我們決定一個狀態上只設定一個高斯混合(mixture)，並且對於高斯混合的機率計算，只採用對角化的共變異矩陣(covariance matrix)。在此HMM訓練的聲學單位為聲母或韻母，我們設定聲母HMM的狀態數為3個，而韻母HMM的狀態數為5個，而HMM的結構，則設定成由左至右(left-to-right)。

當一個聲、韻母的多個發音經過分類後，我們就逐一拿各類的發音來訓練出該類所對應的HMM模型，訓練的程序是，先求取HMM初始模型，再由初始模型開始作分段K中心法(segmental K-means)之反覆式訓練步驟[16]，終止條件是，當本次反覆之維特比(Viterbi)解碼後，各HMM狀態上所收集到的音框內容與上一次反覆所收集到的音框內容完全相同時，即結束訓練。此外，考慮一個音節作合成時，該音節的音長數值是由樂譜檔取得，但是我們必須依此音長值來調整音節內部各個狀態所佔的音長比例，因此在HMM模型訓練時，也需要加入狀態音長參數的訓練，也就是估計HMM第i個狀態被停留的平均音框數D_i及其變異數V_i。

2.5 GMM 音色轉換模型訓練

在訓練GMM模型之前，必需先對平行語料作DTW音框匹配。在我們的平行語料中，說話和唱歌方式的速度有很大的差異，因此我們參考前人的研究[2]，然後撰寫DTW匹配程式，以便找出來源語者(source)與目標語者(target)的對應語音單元之間的正確音框對應關係。我們共準備了四組平行語料來作DTW音框匹配及GMM模型訓練，如表4。

表 4　四組平行語料

組別	來源語者	目標語者
1	唸歌詞 787 句-簡延庭	唱歌詞 787 句-簡延庭
2	唱歌詞 787 句-簡延庭	唸歌詞 787 句-簡延庭
3	唸語料 375 句-朱楠群	唸語料 375 句-簡延庭
4	唸語料 375 句-簡延庭	唸語料 375 句-朱楠群

在此所建造的GMM模型是一種分段式的高斯混合模型(segmental GMM)，分段式高斯混合模型的細節可參考[4]，至於一個分段式GMM的混合組件數，在此我們設定每一種韻母的分段式GMM含有16個混合組件。當訓練一種語音單元的分段式GMM模型時，如何確定它已收斂？這裡採取的判斷方式是，當相接兩次迭代的log likelihood的差異值小於0.1時，就認為是收斂了。

3. 合成階段

合成階段的主要處理流程如圖3所示。首先對讀入的歌譜作分析，以取得歌詞音節及各個音符的音高資料及音長資料。接著，把各個歌詞音節分解成聲、韻母之單位，並且依據其文脈及音高資料去尋找適合的HMM模型。然後根據音長資料及HMM狀態駐留參數，去決定一個HMM裡各狀態應該產生出多少的音框個數，再套用係數產生方法去計算出各音框的DCC係數，然後作變異數調整。在得到各音框的DCC係數後，再進行基於GMM的音色轉換程序。接著，依據音符的資料及抖音參數去計算出各音框的音高數值。最後，依序把各音框的頻譜包絡、音高數值等資料送入HNM信號合成模組去合成出歌聲波形，HNM信號合成請參考[5]。

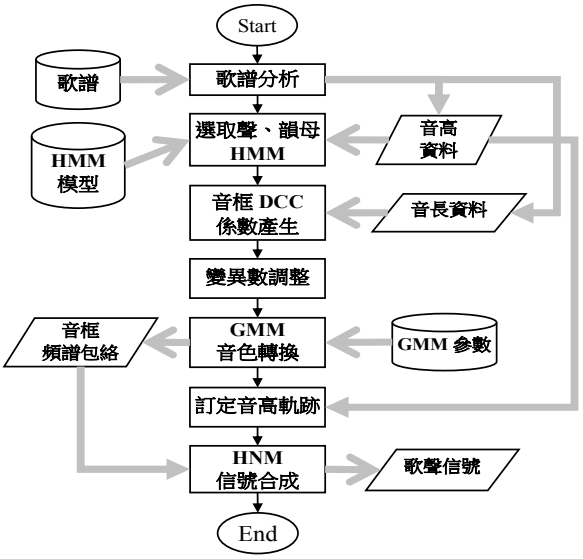


圖 3 合成階段之主流程

3.1 HMM 挑選

在合成階段，我們可依據一個歌詞音節的聲、韻母組成、音高及前後音節的文脈關係，去挑選適合的聲、韻母HMM模型，但是有時我們並不能挑到完全符合的HMM模型，主要原因是所錄的唱歌語料並沒有包含全部可能出現的組合，所以我們必需設定替代的HMM選取規則。

當沒有完全符合的HMM模型可作挑選時，考量到一個韻母的文脈主要是受到前聲母(或前韻母)及後聲母(或後韻母)的影響，而一個聲母的文脈主要是受到後韻母及前韻母的影響，並且考慮此音節的音高範圍，因此我們設定如下的搜尋替代HMM之規則(分別以聲、韻母來討論)：

(替代規則1)：韻母：將待搜尋韻母後面串接的音素類別，改以尋找發音口型近似的音素類別，如果還是搜尋不到近似發音的HMM模型，就作替代規則2的搜尋。聲母：將待搜尋聲母的前面串接的音素類別不作比對，如此去作HMM模型搜尋，並且挑選訓練語料數較多的HMM模型。如果還是搜尋不到近似發音的HMM模型，接著就作替代規則2的搜尋。

(替代規則2)：韻母：如果此音節具有聲母部分，就把韻母前串接的音素類別，替換成發音口型近似的音素類別，再作搜尋。如果此音節不具有聲母，則把韻母前串接的音素類別捨去不作考慮，去搜尋HMM模型並挑選訓練語料數最多的HMM模型。聲母：將待搜尋聲母的後面串接的音素類別不作比對，如此去作HMM模型搜尋，並且挑選訓練語料數最多的HMM模型。如果還是搜尋不到近似發音的HMM模型，接著就進行替代規則3的搜尋。

(替代規則3)：不管是聲母或是韻母，如果替代規則1和2都無法找到近似文脈的HMM，在此就只好單純的搜尋訓練語料數較多的HMM模型。

3.2 HMM 狀態駐留時長

在進行音框DCC係數產生之前，需先依歌譜音符來決定一個歌聲音節的時間長度，然後才能據以設定所選取之HMM各狀態應駐留的音框數。對於一個音節的聲母部分的時間長度，我們依所選取的聲母HMM模型，把各個狀態的時長平均值加總起來，得到 con_i ， con_i 表示第i音節的聲母時長音框數。至於韻母部分的時間長度，在此就以歌聲音節的長度減去聲母長度來作計算，得到 vow_i ， vow_i 表示第i音節的韻母時長音框數。

得知聲、韻母各分配的音框數後，接著考慮聲、韻母HMM各狀態應駐留的音框數。若用HMM原始訓練出的狀態平均時長，則合成歌聲會發生咬字怪異的現象，所以我們參考電腦音樂裡常用的ADSR觀念[8]，將一個歌詞韻母的發音，也類似地分成起音、延音、與釋音三個片段來描述，即ASR分段方式，如此音長之縮短或延長主要是在延音片段進行。在實作上，以公式(2)、(3)和(4)來求算每一個狀態所應分配的時間比例，用以逼近ASR三片段所佔的時間比例。

$$stateRatio(q_k) = calRatio(q_{k+1}) - calRatio(q_k) \qquad (2)$$

$$calRatio(q_k)=\left[\frac{2\times q_k}{stateNum}-1\right]^{1/d}\times 50\%+50\%,\quad 0\le k\le stateNum \qquad (3)$$

$$d = 1+\text{韻母時長(sec)} \qquad (4)$$

其中 $stateRatio(q_k)$ 表示狀態 q_k 所佔的時長比例，$calRatio(q_k)$ 表示 q_k 狀態所對應的時長比例百分比值，*stateNum*為狀態總數，韻母時長(sec)為此音節韻母以秒為單位的時間長度。

3.3 音框 DCC 係數產生

我們採用了兩種音框DCC係數的產生方法，第一種產生方法是最大似然法(簡稱MLE)，它是由Tokuda等人所提出[14, 15]，實作上我們使用HTS_engine API [10]裡的程式模組，並將其移植到我們的歌聲合成系統；第二種產生方法是加權式線性內插法(簡稱WLE)，實作上是使用前人發展的程式模組[6]。

3.4 頻譜過度平滑之改進

由於HMM訓練及GMM訓練時，都會作取平均的動作，使得DCC係數還原後的頻譜包絡曲線出現過度平滑的現象，導致合成出的聲音有悶悶的感覺。為了解決頻譜過度平滑的問題，我們查考過去的文獻得知，Toda與Tokuda等人在HMM語音合成方面的研究[19]，提出了全域變異數(global variance, GV)的方法來改善頻譜過度平滑的問題，因此我們也有了自己的想法。首先，我們依所錄音的44首歌曲的歌譜，使用所製作的HMM歌聲合成系統來產生出每一首歌曲之每一音節的DCC係數，並將所有音節的DCC係數切割成聲、韻母部分，再依據韻母分類分別去收集，接著計算每一類韻母的DCC係數的平均值 $\overline{c_l^{Syn}}$ 及標準差 σ_l^{Syn}。同樣地，我們也將真人所唱的44首歌曲裡的音節DCC係數，依韻母分類分別作收集，去計算每一類韻母的DCC係數平均值 $\overline{c_l^{Natrl}}$ 及標準差 σ_l^{Natrl} 。

一般來說合成歌聲韻母DCC的平均值與真人所唱韻母DCC的平均值差異不大。但是從圖4則可看出合成歌聲韻母/a/的DCC前四維的標準差與真人所唱韻母/a/的DCC標準差之間有不小的差異。因此我們的想法是，先求取訓練歌曲的合成音和真人發音之各韻母的DCC平均值及標準差，在此稱這些參數為音段式變異數，然後在合成階段利用音段式變異數及公式(5)去調整DCC係數。

$$\hat{c}_i(k)=\left(c_i(k)-\overline{c_l^{Syn}(k)}\right)\times\frac{\sigma_l^{Natrl}(k)}{\sigma_l^{Syn}(k)}+\overline{c_l^{Natrl}(k)},\quad k=1,\ldots,80 \qquad (5)$$

其中 $\hat{c}_i(k)$ 表示第*i*個音框之第*k*維調整後的 DCC係數，*l*表示第*l*類韻母，$\overline{c_l^{Syn}(k)}$ 及 $\sigma_l^{Syn}(k)$ 表示合成音的音段式變異數，$\overline{c_l^{Natrl}(k)}$ 及 $\sigma_l^{Natrl}(k)$ 表示真人歌聲的音段式變異數。當使用音段式變異數來調整3.3節產生出的DCC係數後，去合成出歌聲來聽，悶悶的感覺的確可改善很多。

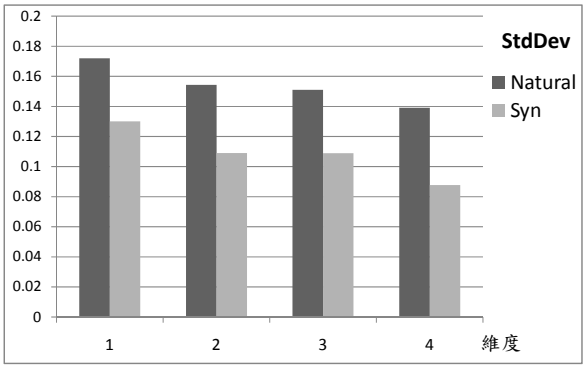


圖 4 韻母/a/ DCC 係數前四維的標準差

4. 音色轉換方法

4.1 基本音色轉換法

第一種轉換方法稱為基本音色轉換法，基本音色轉換法的流程如圖5，使用表4第四組平行語料所訓練出的GMM模型，轉換是以歌詞音節為單位，在頻譜對映前需先挑選出正確的韻母GMM，然後再作GMM頻譜對映，以得到轉換後的DCC係數。當使用簡延庭歌聲訓練的HMM模型去產生出DCC係數後，接著就可以選取對應韻母的GMM模型來作音色轉換，以合成出具有朱楠群音色的歌聲(朱楠群只有說話的語料可用)。

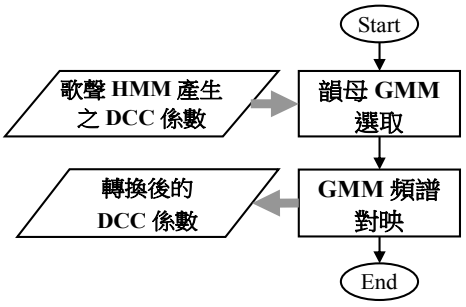


圖 5 基本音色轉換法之流程

實作上我們發現，如果在使用HMM模型產生出一序列音框的DCC係數之後才作GMM頻譜對映，則會發生頻譜包絡在一些時刻出現不連續的現象，導致合成出的歌聲信號音質不佳，因此我們嘗試在音框DCC係數產生之前，就先對HMM模型各狀態的平均DCC向量作GMM頻譜對映，然後才用以產生該歌詞音節的一序列音框DCC係數，這樣的作法可以有效改進頻譜包絡不連續的問題，而使合成出的歌聲信號品質獲得提升。

4.2 使用 GMM 之相對振幅轉換法

第二種轉換方法稱為**使用GMM之相對振幅轉換法**，其處理流程如圖6，這個方法首先把來源語

者(簡延庭，以A為代號)的歌聲頻譜DCC轉成來源語者的說話頻譜DCC，即(A歌聲) => (A語音)；接著把來源語者的說話頻譜DCC轉成目標語者(朱楠群，以B為代號)的說話頻譜DCC，即(A語音) => (B語音)；然後依據DCC轉出的頻譜包絡 $X(f)$ 與 $Y(f)$ (即2.2.2節裡的 $S(t,f_k)$)，去計算頻譜振幅差值 $\Delta dB(f)$ ；之後，把來源語者的歌聲頻譜DCC係數轉成頻譜包絡 $Z(f)$，再把 $Z(f)$ 加上頻譜振幅差值 $\Delta dB(f)$，如此就可以得到音色轉換後的頻譜包絡 $V(f)$，接著就可以 $V(f)$ 來合成出具有朱楠群音色的歌聲(朱楠群只有說話的語料可用)。

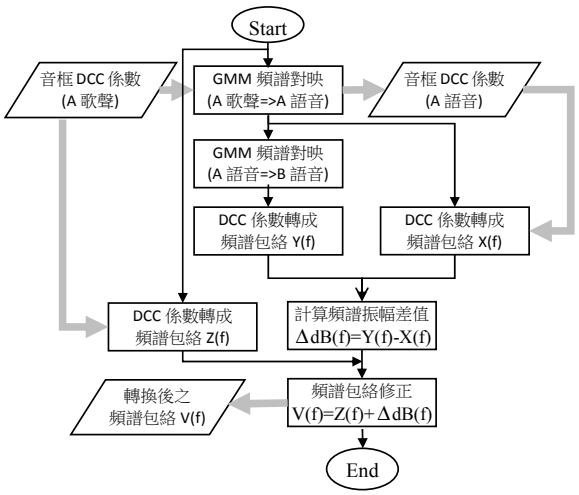


圖 6 使用 GMM 之相對振幅轉換法的流程

4.3 不含 GMM 之相對振幅轉換法

第三種轉換方法稱為**不含GMM之相對振幅轉換法**。由於圖6的相對振幅轉換法使用了GMM頻譜對映來得到A語音及B語音的音框DCC係數，以至於最後合成出的歌聲音質悶悶的。因此我們另外研究一種不含GMM頻譜對映的處理流程，如圖7之流程，它直接從A(簡延庭)語音及B(朱楠群)語音訓練的HMM去計算頻譜振幅差值 $\Delta dB(f)$ ，然後把 $\Delta dB(f)$ 加上A歌聲HMM產生的頻譜包絡 $Z(f)$，如此以求得音色轉換後的頻譜包絡 $V(f)$ ，之後就可以 $V(f)$ 來合成出具有朱楠群音色的歌聲。

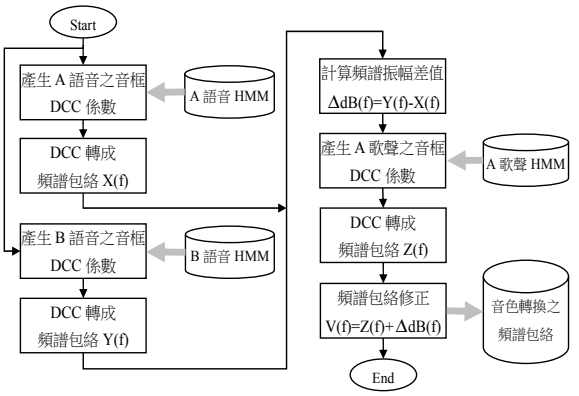


圖 7 不含 GMM 之相對振幅轉換法的流程

5. 聽測實驗

為了比較歌聲HMM模型和說話HMM模型所合成歌聲之共鳴度差異，以及比較不同音色轉換方法所轉出歌聲之音色相似度和聲音品質，因此我們進行聽測實驗，共分成三個聽測項目，分別是共鳴感聽測、音色相似度聽測及聲音品質聽測，將於各子節逐一介紹。在以下聽測實驗裡，我們共邀請21位受測者，其中11位有語音處理領域的背景，其餘10位則無；而聽測實驗所用的合成歌聲音檔，可從網頁 http://guhy.csie.ntust.edu.tw/SongHMM/ 去下載試聽。

5.1 共鳴感聽測

進行共鳴感聽測時，給每一位受測者聽二首歌曲，每一首歌曲有兩個合成音檔，其中一個使用歌唱聲所訓練之HMM模型作合成，另一個則使用說話聲所訓練之HMM模型作合成，每一個音檔都使用了音段式變異數來作調整。共鳴感聽測的評分範圍為1~5分，分數越高，代表此歌聲越具共鳴感；反之，分數越低，代表此歌聲越不具共鳴感。共鳴感聽測後，我們計算平均評分和標準差，所得到的數值如表5。

表 5 共鳴感聽測之平均評分與標準差

編號	歌曲名稱	HMM 模型	平均評分	標準差
1(A)	青春舞曲	說話	3.0476	0.8047
1(B)		唱歌	**3.6667**	**1.0165**
2(A)	康定情歌	說話	2.9047	0.9436
2(B)		唱歌	**3.0952**	**1.3381**

觀察表5裡的平均評分，可發現“青春舞曲”與“康定情歌”的分數，使用歌唱HMM模型所合成歌聲之共鳴感都是比使用說話HMM模型的高，不過在節奏較慢的“康定情歌”裡，兩種不同HMM模型所合成歌聲之共鳴感差距不大，因此，在合成節奏較快的歌曲時，使用歌唱HMM模型確實可明顯提升共鳴感。

5.2 音色相似度聽測

在此共有三種音色轉換方法，以ABX聽測方式作音色比較，三種方法分別是基本音色轉換法、使用GMM之相對振幅轉換法和不含GMM之相對振幅轉換法。聽測方式為，請受測者判斷音檔X的音色比較接近音檔A，還是音檔B，並依照評分標準評分，評分範圍為1~5分，分數越低代表越接近目標語者，反之分數越高代表越接近來源語者。6組ABX聽測後，計算各組的平均評分和標準差，結果得到的數值如表6所示。

從表6觀察不同的音色轉換方法的平均評分可發現，不管是青春舞曲還是康定情歌，基本音色轉換法的平均評分最靠近目標音色，而使用GMM之

相對振幅轉換法的平均評分則是有點靠近目標音色，但是又有點分不出來是誰的音色，至於不含GMM之相對振幅轉換法的平均評分則很接近3分，這表示大部分的受測者分不出來這是誰的音色。

表 6　音色聽測之平均評分與標準差

組別	歌曲名稱	音色轉換方法	平均評分	標準差
1	青春舞曲	基本音色轉換法	2.3333	1.0165
2	康定情歌	基本音色轉換法	2.5238	0.9284
3	青春舞曲	相對振幅轉換法(使用 GMM)	2.7619	0.9952
4	康定情歌	相對振幅轉換法(使用 GMM)	2.9524	0.8047
5	青春舞曲	相對振幅轉換法(不使用 GMM)	3.0952	1.1792
6	康定情歌	相對振幅轉換法(不使用 GMM)	3	0.7071

5.3 聲音品質聽測

在聲音品質的聽測實驗中，每一首歌曲我們準備4個音色轉換過的音檔及一首直接使用朱楠群說話HMM模型合成的音檔，詳細資訊如下表。在此

表 7　聲音品質聽測的合成歌聲資訊

代號	HMM 模型	音色轉換方法	音框係數產生方法
A	簡延庭唱歌	基本音色轉換法	最大似然法
B	簡延庭唱歌	相對振幅轉換法(使用 GMM)	最大似然法
C	簡延庭唱歌	相對振幅轉換法(不使用 GMM)	最大似然法
D	朱楠群說話	無	最大似然法
E	簡延庭唱歌	相對振幅轉換法(使用 GMM)	加權式線性內插法

聽測的方式為，請受測者判斷音檔X的品質比較好亦或音檔Y的品質比較好，並依照評分標準評分，評分範圍為1~5分，分數越低代表音檔X的品質越好，反之分數越高代表音檔Y的品質越好。聲音品質聽測實驗之後，我們計算平均評分和標準差，得到的數值如表8。

表 8　聲音品質聽測之平均評分與標準差

青春舞曲				
-	A vs B	B vs C	C vs D	B vs E
平均評分	2.1429	3.5238	4.8571	2.8571
標準差	1.1084	0.8729	0.3586	0.6547
康定情歌				
-	A vs B	B vs C	C vs D	B vs E
平均評分	2.7619	2.7619	4.9048	3.381
標準差	0.9437	0.9952	0.3008	0.8646

依據表8中的平均評分比較A與B兩種不同音色轉換方法所產生音檔的音質，不管是"青春舞曲"還是"康定情歌"，都是A音檔的聲音品質比B音檔的好一點。接著，比較B與C兩種音色轉換方法所產生音檔的音質，發現兩首歌曲的結果有不一致的評分，不過它們的分數都相當靠近3分(分不出來)，所以我們認為B與C兩種音色轉換方法的音質，受測者似乎分不清楚誰比較好。關於C與D兩音檔的音質比較，由於D音檔是直接使用朱楠群說話的HMM模型來合成，未經過音色轉換的處理，所以評分的結果顯示D的音質遠優於C，由此可知，各種的音色轉換方法都會造成音質退化。

另外，我們使用同樣的音色轉換方法，也就是使用GMM之相對振幅轉換法，來比較不同的音框係數產生方法所產生的音檔，測驗是否不同的音框係數產生方法會造成音質上的差異，結果如表8裡B和E比較的平均評分所顯示的，雖然兩首歌的平均評分有不一致情形，但是兩首歌的平均評分都相當靠近3分，因此兩種音框係數產生方法所合成的音檔，受測者似乎分不清楚誰的音質比較好。

6. 結論

本論文研究一種結合HMM頻譜模型及HNM信號模型之國語歌聲合成系統；此外我們把HMM頻譜模型、GMM音色轉換模型兩者作結合，以發展一個具有歌者音色轉換功能的歌聲信號合成系統。

在頻譜係數的分析上，使用STRAIGHT來求得較準確的頻譜包絡及音高資訊，並修改前人所發展的頻譜係數估計程式，使其依據STRAIGHT求得的頻譜包絡來換算出DCC係數。在語料分類上，我們不僅考慮了聲、韻母分類及文脈分類，也將語料依不同音高，再細分為高音、中音及低音等3種子類，以解決合成歌聲之音色不一致的問題。在HMM模型及GMM音色轉換模型的訓練方面，都有頻譜過度平滑的問題，於是我們提出音段式變異數的作法，以改善頻譜過度平滑的問題，但其反效果是會造成合成音裡出現喀嚓聲(click)。

關於只有說話語料之目標歌者的音色轉換，我們研究了三種轉換方法，分別是基本音色轉換法、使用GMM之相對振幅轉換法及不含GMM之相對振幅轉換法。

聽測實驗的結果是，歌唱語料訓練的 HMM模型比說話語料訓練的 HMM模型較能夠合成出有共鳴感的歌聲；此外，音色轉換的聽測結果是，基本音色轉換法所轉換出的歌聲，在音色與聲音品質上，都比其它轉換方法的效果來得好。

未來還可以改進的地方，在語料的分類上，由於所錄製的語料不夠平衡，造成很多HMM模型的訓練發音只有一個成員，這可能造成訓練出HMM模型較不穩定；此外在語料的分類上，應可以再考慮依音長來作分類，因為同一發音及音高，不同音長的頻譜特性仍是不太一樣的。在歌者音色的轉換上，不管是音色相似度，還是聲音品質，都還有進步的空間，未來可以嘗試不同的音色轉換方法，以改進音色相似度及聲音品質。

致謝

感謝國科會計畫之經費支援，國科會計畫編號NSC 100-2221-E-011-157。

參考文獻

[1] 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)。

[2] 古鴻炎、吳昌益，「基於 ANN 之頻譜演進模型及其於國語語音合成之應用」，*第二十屆自然語言與語音處理研討會* (ROCLING 2008)，台北，第 66-77 頁，(2008)。

[3] 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)。

[4] 古鴻炎、蔡松峰，「使用分段式 GMM 及自動 GMM 挑選之語音轉換方法」，*第 23 屆自然語言與語音處理研討會*，台北，第 216-226 頁，(2011)。

[5] 古鴻炎、蔡松峰，「基於離散倒頻譜之頻譜包絡估計架構及其於語音轉換之應用」，*第 21 屆自然語言與語音處理研討會* (ROCLING 2009)，台中，第 151-164 頁，(2009)。

[6] 古鴻炎、賴名彥、蔡松峰，「結合 HMM 頻譜模型與 ANN 韻律模型之國語語音合成系統」，*第 22 屆自然語言與語音處理研討會*，南投，第 281-295 頁，(2010)。

[7] 校園民歌回顧，一品文化出版，台北，1985。

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

[9] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction," Speech Communication 27, pp. 187-207 , 1999.

[10] HMM-based Speech Synthesis System (HTS), http://hts.sp.nitech.ac.jp/.

[11] J. Bonada, X. Serra, "Synthesis of the singing voice by performance sampling and spectral models," IEEE Signal Processing Magazine March 2007.

[12] K. Saino, H. Zen, Y. Nankaku, A. Lee, K. Tokuda, "An HMM-based Singing Voice Synthesis System," INTERSPEECH - ICSLP, Pittsburgh, Pennsylvania, USA, pp. 2274-2277, 2006.

[13] K. Tokuda, H. Zen, and A.W. Black. "An HMM-based speech synthesis system applied to English," Proc. IEEE 2002 Workshop on Speech Synthesis, Santa Monica, USA, Sep. 2002.

[14] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi and S. Imai, "An Algorithm for Speech Parameter Generation from Continuous Mixture HMMs with Dynamic Features," Proc. EUROSPEECH-95, Madrid, Spain, pp. 757-760, 1995.

[15] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proc. of ICASSP 2000, Istanbul, Turkey, pp. 1315-1318, June 2000.

[16] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[17] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK Book (for HTK version 3.2.1), Cambridge University Engineering Department, 2002.

[18] Sinsy, "HMM-based Singing Voice Synthesis System," http://www.sinsy.jp/.

[19] T. Toda and K. Tokuda, "A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis," IEICE trans. Inf. & Syst., Vol. E90-D, No. 5, May 2007.

[20] 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.

[21] Yamaha, VOCALOID, New Singing Synthesis Technology, http://www.vocaloid.com/en/.