使用分段式 GMM 及自動 GMM 挑選之語音轉換方法

A Voice Conversion Method Using Segmental GMMs and Automatic GMM Selection

古鴻炎　　蔡松峰
Hung-Yan Gu　　Sung-Fung Tsai

國立台灣科技大學 資訊工程系
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
{guhy, M9615069}@mail.ntust.edu.tw

摘要

本論文提出分段式(segmental)高斯混合模型(Gaussian mixture model, GMM)的觀念，用以改進語音轉換的效能，而爲了應用該觀念於線上(on-line)進行的語音轉換處理，我們也發展了一個基於動態規劃(dynamic programming, DP)之自動 GMM 挑選的演算法。此外，爲了使用單一高斯混合來對映(mapping)離散倒頻譜係數(discrete cepstrum coefficients, DCC)係數，我們也設計了一種高斯混合選取之演算法。關於分段式 GMM 觀念的評估，在此我們建造了三個採取不同功能組合之語音轉換系統，然後使用三個系統所轉換出的語音去作聽測實驗，實驗的結果顯示，分段式 GMM 之觀念確實可用以改進音色相似度(timbre similarity)、及語音品質(voice quality)兩方面的效能。

關鍵詞：語音轉換，離散倒頻譜，高斯混合模型，音色相似度，語音品質

一、緒論

以 GMM 爲基礎的語音轉換方法首先由 Stylianou 提出[1]，之後有許多研究者對這種方法的相關議題再作了更進一步的探討[2-5]，然而幾個棘手的問題至今仍然存在，其中一個最令人注意的是，經語音轉換得到的頻譜大多都會發生頻譜過度平滑化(over smoothing)的現象[2-4]，而使得轉換出的語音聽起來會有明顯的語音品質退化的感覺。此外，另一個也需要注意的問題是，當嘗試以最大加權值之混合來作單一高斯混合之頻譜對映時，某些相鄰音框的轉換出的頻譜，可能會發生頻譜不連續的問題，而使得轉換出的語音會時常聽到怪音(artifact sound)。

在本論文裡，我們嘗試以不同的方向來解決頻譜過度平滑的問題。在 GMM 爲基礎的語音轉換方法中，跨越多個(如 128 個)高斯混合作加權和(weighting sum)的運算，是導致過度平滑的一個重要原因。一個典型的基於 GMM 的對映函數，其公式如下[1]:

$$y = F(x;\mu,\Psi) = \sum_{m=1}^{M}\left[\frac{w_m \cdot N(x;\mu_m^x,\Psi_m^{xx})}{\sum_{m=1}^{M} w_m \cdot N(x;\mu_m^x,\Psi_m^{xx})}\left(\mu_m^y + \left(\Psi_m^{yx}\right)\cdot\left(\Psi_m^{xx}\right)^{-1}\cdot(x-\mu_m^x)\right)\right] \quad (1)$$

其中 x 表示來源語者的頻譜特徵向量，y 表示轉換後得到的頻譜特徵向量，M 是高斯混合 $N(\bullet, \bullet, \bullet)$的總數，而 μ 及 Ψ 分別表示平均向量與共變異矩陣的集合。爲了解決頻譜過度平滑之問題，我們認爲減少公式(1)中高斯混合的個數 M 是必需的，然而當直接減小 M 值時，訓練出的 GMM 所建構的機率密度函數必然會變得粗糙。因此，我們思考去對模型訓練的語句先作切割，使成爲一序列的語音片段，然後將這些語音片段作分類而分成數群，接著拿各群的語音音框分別去訓練出混合個數較少(如 16 個混合)的 GMM，而不同群的語音片段所訓練出的多個 GMM，將來就只從其中挑出一個 GMM 來對屬於同一分類的來源語音(source speech)音框作頻譜對映，如此，基於 GMM 的對映函數(如公式(1))就可使用較少的混合個數。也就是說原先一個複雜的 GMM 對映函數，現在被多個較簡單的 GMM 對映函數所取代了。

在本論文裡，我們探討國語(華語)之語音轉換，而國語是一個音節顯著的語言，因此我們決定以訓練語句裡所標示的各個音節作爲語音片段(segment)，如此一句話若有 7 個音節，就看成是由 7 個語音片段串接而成。再者，國語有 37 種不同的韻母，因此我們就依據韻母來把訓練語句的語音片段分成 37 群。由於我們使用的是平行語料，所以對於各群的平行語音片段，就可分別拿去訓練出一個對應的聯合(joint) GMM 模型，當訓練好 GMM 模型之後，就可拿這 37 個 GMM 模型去作線上的語音轉換處理。至於一個輸入的語音音框，應如何從這 37 個 GMM 中去挑選出一個正確的 GMM 來對它作頻譜轉換？對於這個問題，我們發展了一個以 DP 爲基礎的 GMM 自動挑選之演算法，該演算法將會在 3.1 節中詳細說明。

除了採取分段式的多個 GMM 來減少高斯混合的個數之外，我們更進一步採取單一高斯混合作對映的方法，來對來源語者的輸入音框作頻譜轉換，希望如此的組合式處理，能夠用以解決轉換出的頻譜包絡會變得過度平滑的問題。不過，當採取前述的組合式處理時，相鄰的兩個音框的轉換後頻譜，仍然可能發生頻譜不連續的情況，而導致怪音被產生出來。爲了避免發生頻譜不連續的情形，我們就嘗試設計一個基於 DP 的演算法，以便對一序列的音框作整體考量，即同時考慮各高斯混合被使用的似然率(likelihood)及其對頻譜連續性可能造成的危害，這個演算法的細節，我們將在 3.2 節中說明。依據前述提到的幾個作法，我們實際去製作出線上處理之語音轉換系統，然後使用這些系統轉換出的語音來作聽測實驗。

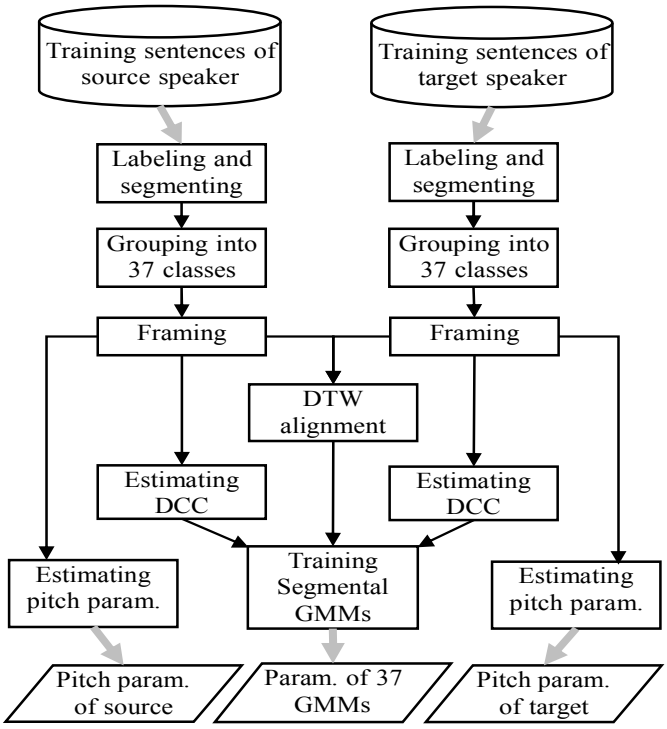
二、系統訓練階段

我們的語音轉換系統，在訓練階段的主要處理步驟如圖一所示。首先我們邀請了三位錄音者，分別到隔音錄音室來錄製 375 句之平行語料，取樣率設爲 22,050Hz，其中二位是男性，在此以 M1 和 M2 作代號，而另一位是女性，以 F1 作代號。在本研究裡，我們把 M1 當作來源語者，而把 M2 和 F1 分別作爲目標語者，也就是說我們要把 M1 的語音轉換成 M2 及 F1 的語音。

2.1 標音與分群

對於各個語者所錄的訓練語句，我們先操作 HTK (HMM tool kit)軟體，經由強制對齊(forced alignment)來作自動標音，把一個語句的各個音節的邊界標示出來。由於自動標記的音節邊界有很多是錯誤的，因此我們再操作 WaveSurfer 軟體，以人工檢查自動標

記的音節邊界是否有錯，若發現錯誤則加以更正。然後，依據各音節的拼音符號標記及音節邊界標記，就可將一個訓練語句的各個音節的語音信號分別擷取、及存成獨立的音檔，再依語句編號、音節序號和音節拼音來命名該音檔。整體來說，375 個訓練語句可擷取出 2,926 個音節音檔。之後，作爲模型訓練之用的音節音檔，我們再依其檔名中的韻母拼音符號，將這些音節音檔分成 37 群。



圖一、訓練階段之主要處理流程

2.2 DCC 係數計算

關於一個語音音框的振幅頻譜包絡(magnitude-spectrum envelope)的估計，過去已有一些方法被提出。雖然 STRAIGHT 法[12]可估計出相當準確的頻譜包絡，但是它需求的計算量也很大，而難以用於製作即時處理的系統。因此在本論文裡，我們採用離散倒頻譜之頻譜包絡估計方法[7, 8]，並且以離散倒頻譜係數(DCC)作爲頻譜參數。對於一個語音音框，我們使用先前發展的 DCC 估計程式[8]來計算出 40 維的 DCC 係數，在此一個音框的長度設爲 512 個樣本點(23.2ms)，而音框位移則設爲 110 個樣本點(5ms)。

2.3 分段式 GMM 之訓練

在圖一中經由方塊 ”Grouping into 37 classes” 的處理之後，對於各群的音節片段，我們就分別拿去訓練出一個由 16 個高斯混合所形成的 GMM 模型，所以這樣得到的 37 個 GMM，就稱爲 37 個分段式 GMM。

由於我們使用的是平行語料，每一個來源語者音節和它對應的目標語音音節，可先以動態時間校正(dynamic time warping, DTW)作時間軸對齊的處理，這由圖一裡的 ”DTW alignment” 方塊負責。然後，一個來源語音音框和它所對齊的目標語音音框，兩音框算出的 DCC 係數就可被合併成一個 80 維的頻譜特徵向量，接著我們使用基於 MLE

(maximum likelihood estimate)的GMM訓練方法[9]，來對各群合併後DCC向量進行MLE訓練，如此就可得到各群的聯合機率密度之 GMM 模型。

2.4 音高參數

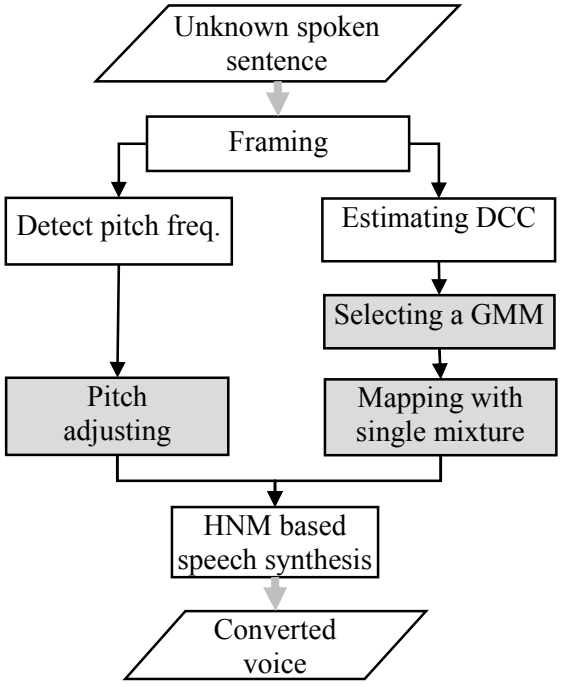
我們使用一種基於自相關函數及AMDF (absolute magnitude difference function)的基週偵測方法[10]，來偵測各音框的音高頻率，然後將一個語者發音中有聲(voiced)音框偵測出的音高頻率值收集起來，據以求出他們的平均值及標準差，這就是我們所需要的音高參數。

三、語音轉換階段

我們研究的語音轉換方法，其主要的處理流程如圖二所示。當一句未知內容的語句輸入後，它首先會被切割成一序列的音框，而音框長度(512 點)和位移(110 點)則和 2.2 節裡使用的一樣。然後，在圖二的左邊流程，會對各音框的音高頻率作偵測，當一個音框被偵測爲無聲時，圖二中的三個灰色方塊就被直接跳過，也就是不需作音高頻率的調整，且 DCC 頻譜參數也未被轉換。相對地當一個音框被偵測爲有聲時，我們在此採用一種簡便的音高調整公式來調整音高頻率，

$$q_t = \mu^y + \frac{\sigma^y}{\sigma^x}(p_t - \mu^x) \tag{2}$$

其中 p_t 表示偵測出的音高頻率值，μ^x 和 σ^x 分別表示來源語者的音高頻率平均值和標準查，而 μ^y 和 σ^y 是目標語者的。



圖二、轉換階段之主要處理流程

在圖二裡的右邊流程，基本上是一個音框接著一個音框來作處理，但是在 “Selecting a GMM” 之方塊裡，我們提出一種 GMM 自動挑選之演算法，該演算法是以每 20 個有聲音框爲一個批次(batch)來作 GMM 的挑選，以便爲各個有聲音框從 37 個 GMM 中選出正確的(或鄰近的)一個 GMM。之後，在 “Mapping with single mixture” 之方塊裡，我們再從一個音框所選取到的 GMM 裡，選取出一個高斯混合來作單一高斯混合之 DCC 係數對映，以便避免頻譜曲線過度平滑的情形發生。不過，我們不能只依據加權值的大小來分別爲各個音框挑選出單一個高斯混合，因爲相鄰音框的轉換後頻譜的連續性也必需被考慮，以避免怪音被產生出來。對於單一高斯混合選取的問題，我們也發展了一個基於 DP 的演算法，該演算法和前人提出的[4]不一樣，基本上是把一序列的有聲音框(左、右兩邊被無聲音框包夾)，當作一個批次來作單一高斯混合選取的處理。接著在圖二裡左右流程合併之方塊 ”HNM based speech synthesis”，我們使用一個基於 HNM (harmonic plus noise model)的信號合成方法[8, 11]，去依據轉換出的 DCC 係數及音高頻譜，把語音信號再合成出來。

3.1 分段 GMM 之選取方法

對於一個線上處理的語音轉換系統來說，輸入語音的說話內容是事先不知道的，因此當要對一個音框的 DCC 係數作對映時，我們如何知道 37 個 GMM 當中的那一個應被選取? 這樣的問題必須先被解決，而該問題是一種語音辨識的問題，不過它不需要像語音辨識那樣嚴厲地被對待，因爲選取到錯誤但近似的 GMM 是可以容忍的。

在語音辨識領域，隱藏式馬可夫模型(hidden Markov model, HMM)是最常被採用的統計模型，不過在此我們希望以所訓練出的 37 個 GMM 來取代 HMM 的角色，如此就不需另外去訓練 HMM。此外，我們觀察到一個非常接近真實的現象是，一個人不可能在一個很短暫的時間如 100ms 之內，發出多於 2 個的語音片段，在此語音片段指的是音節。所以，我們決定把每 20 個連續的有聲音框(含蓋 100ms 之時間範圍)作爲一個批次，去作 20 個音框整批的 GMM 選取之處理，如此一個批次裡就只需選出一個或二個的 GMM。本論文研發了一個 DP 爲基礎的 GMM 挑選之演算法，該演算法會依據最大似然率(maximum likelihood)去選出一個或二個 GMM。

令第 t 個輸入音框的 DCC 係數是由第 s 個 GMM 所產生的機率是 $G_t(s)$，其詳細計算公式爲

$$G_t(s) = \sum_{m=1}^{M} w_m(s) \cdot N\left(x_t;\ \mu_m^x(s),\ \Psi_m^{xx}(s)\right)\ , \tag{3}$$

其中 $W_m(s)$表示第 m 個高斯混合的加權，x_t 表示第 t 個音框的 DCC 向量。此外，令 $R(t, s)$表示從時刻 1 到時刻 t 的音框都是由第 s 個 GMM 所產生的似然率對數值，而令 $D(t, s)$表示從時刻 1 到時刻 t 的音框是由 2 個 GMM 所產生，並且第 t 個音框是由第 s 個 GMM 所產生的似然率對數值。依據前述的定義，我們可以推導出如下的兩個遞迴公式:

$$R(t,s) = \log\left(G_t(s)\right) + R(t-1, s)\ , \tag{4}$$

$$D(t,s) = \log\left(G_t(s)\right) + \max\left\{\max_{0\le v<37,\, v\ne s}\left[R(t-1,v)\right],\ D(t-1,s)\right\}, \tag{5}$$

其所需設定的邊界值是，$D(1, s)=0$ 和 $R(1, s)=G_1(s)$，$s=0, 1, ..., 36$。接著，依據公式(4)

和(5)，我們可得到 T 個音框整體的最大似然率爲

$$A(T) = \max\left\{ \max_{0\le v<37}\left[R(T,v)\right],\ \max_{0\le v<37}\left[D(T,v)\right] \right\} \ , \tag{6}$$

其中 T 在本論文裡設爲 20。在依據公式(4)，(5)和(6)得到 $A(20)$之最大似然率數值之後，我們可作回溯(backtrack)處理，去找出 $A(20)$數值的最佳行走路徑，而得到 20 個音框各自所被指派的 GMM 編號。

3.2 單一高斯混合之對映

所謂使用單一高斯混合來對映一個輸入音框的 DCC 係數，其實際作法是把公式(1)裡的累加符號及加權項移除，如此轉換出的 DCC 向量 y 就變成以下列公式來計算，

$$y = F^k(x) = \mu_k^y + \left(\Psi_k^{yx}\right)\cdot\left(\Psi_k^{xx}\right)^{-1}\cdot(x-\mu_k^x)\,, \tag{7}$$

其中 x 表示輸入音框的 DCC 係數，$F^k(x)$表示使用第 k 個高斯混合所建立的對映函數。

關於公式(7)裡 k 值(即高斯混合之編號)的選取的問題，我們設計了一個基於 DP 的高斯混合選取之演算法。首先令 3.1 節中爲第 t 個音框自動挑出之 GMM 編號爲 $I(t)$，接著以 $F_{I(t)}^k(x_t)$ 表示使用第 k 個高斯混合來對第 t 個音框之 DCC 向量 x_t 作對映，此外以 $C(t, k)$ 表示從時刻 1 到時刻 t 的累積距離，但是限定在時刻 t 時使用編號爲 k 的高斯混合，如此我們設計的遞迴公式就可寫成

$$C(t,k) = \min_{\substack{0\le m<M,\\ w_m(I(t-1))>H}} \left[dist\left(F_{I(t)}^k(x_t), F_{I(t-1)}^m(x_{t-1})\right) + C(t-1,m)\right] \ , \tag{8}$$

其中 $dist(\bullet,\ \bullet)$表示對兩 DCC 向量之間作幾何距離的量測，H 是一個門檻參數，我們依經驗設定它的值爲 0.3，而 $W_m(I(t\text{-}1))$表示第 $I(t\text{-}1)$個 GMM 的第 m 個混合的加權。

公式(8)的意義是，在各個時刻 t 先依 $W_m(I(t)) > H$ 之條件篩選出加權夠大的幾個高斯混合，然後從各時刻篩選出的高斯混合中，以 DP 的觀念去串接出行走的路徑，最後在結束的時刻 T 時，以下列公式找出最小的累積距離 $B(T)$，

$$B(T) = \min_{0\le k<M,\ w_k(I(T))>H} \left[C(T,k)\right] \ , \tag{9}$$

所以依據公式(8)和(9)，我們可求得最小的累積距離，然後經由回溯的程序找出行走的路徑，如此就可決定時刻 1 到時刻 T 各個音框所應選取的高斯混合。至於公式(8)裡 $C(t, k)$在 t=0 時的邊界數值，我們可直接設定成 $C(0, k)$=0，$0 \le k < M$。

3.3 基於 HNM 之語音信號合成

在諧波加雜音模型(harmonic plus noise model，HNM)中，一個有聲音框的頻譜被分割成低頻的諧波部分和高頻的雜音部分，而分割這兩部分的邊界頻率稱爲最大有聲頻率(maximum voiced frequency，MVF)。關於 MVF 值的偵測，在 Stylianou 的博士論文裡[11]，提出了一個對各個音框逐一作偵測的方法，不過爲了簡化語音信號合成處理的程序，在此我們把各個有聲音框的 MVF 值都直接設爲 6,000Hz。

假設第 i 和第 i+1 個音框都是有聲的，並且分別有 L^i 和 L^{i+1} 個諧波成分(harmonic partials)，L^i 的值以 MVF / q_i 作計算，q_i 表示第 i 個音框的轉換過的基頻值。當要對這兩個音框之間的第 t 個樣本點產生出信號樣本值，首先我們以線性內插來計算第 t 個樣本點上的各個諧波成分的頻率值 $f_k(t)$ 和振幅值 $a_k(t)$ ，計算方式如公式(10)所示，

$$\begin{aligned} f_k(t) &= f_k^i + \frac{f_k^{i+1} - f_k^i}{N} \cdot t, \quad k = 1, 2, ..., L, \\ a_k(t) &= a_k^i + \frac{a_k^{i+1} - a_k^i}{N} \cdot t, \quad k = 1, 2, ..., L \end{aligned} \tag{10}$$

其中 N 表示兩相鄰音框之間的樣本點總數(在此設爲 110，即音框位移的點數)，L 表示 L^i 和 L^{i+1} 兩者的較大值，此外 f_k^i 和 a_k^i 分別表示第 i 個音框的第 k 的諧波成分的頻率值和振幅值，f_k^i 可以 $f_k^i = k \times q_i$ 作計算，而 a_k^i 則必需依據第 i 個音框對映得到的 DCC 係數，轉換成頻譜包絡後再去求取它的數值，關於 a_k^i 數值求取的細節請參考我們先前發表的論文[8]。另外，如果 L^i 小於 L^{i+1}，我們就直接設定 $a_k^i = 0$， $k = L^i + 1, \ldots, L^{i+1}$ 。然後，第 t 個樣本點上的諧波信號 $h(t)$就可以公式(11)來作計算，

$$\begin{aligned} h(t) &= \sum_{k=1}^{L} a_k(t) \cdot \cos(\phi_k(t)), \quad 0 \le t < N, \\ \phi_k(t) &= \phi_k(t-1) + 2\pi \cdot f_k(t) / 22,050 \end{aligned} \tag{11}$$

其中 $\phi_k(t)$ 表示第 k 個諧波成分在樣本點 t 時的累積相位，22,050 是取樣率。至於 $\phi_k(t)$ 的初值 $\phi_k(-1)$ ，我們可令它等於前一個音框最後一個樣本點時的累積相位(即 $\phi_k(N-1)$)，以保持相位的連續性。如果本音框是第一個音框(即沒有前一個音框)，則可令 $\phi_k(-1)$ 的值爲一個隨機值，使用隨機值是符合語音信號特性的。

四、實驗評估

爲了評估所提出的轉換方法，我們建造了三個語音轉換系統，分別以 SOG，SSG 和 SLG 作爲代號，在代號 SOG (system using original GMM for mapping)的系統裡，我們使用 350 個訓練語句來訓練出一個由 256 個高斯混合形成的 GMM，然後使用公式(1)來對各個輸入音框的 DCC 係數作對映。另外，在代號 SSG (system using single Gaussian mixture for mapping)的系統裡，我們仍然使用 350 個語句所訓練出的一個具有 256 個高斯混合的 GMM，不過在轉換階段，3.2 節裡說明的高斯混合選取方法被用來爲一序列的輸入音框選取出各音框的單一高斯混合，然後各輸入音框的 DCC 係數就使用所選出的單一高斯混合及公式(7)來作對映。至於在代號 SLG (system using selected GMM for mapping)的系統裡，我們首先以 350 個語句來訓練出 37 個分段式 GMM，而各分段式 GMM 都只有 16 個高斯混合，然後在轉換階段，我們採用 3.1 節裡說明的 GMM 選取方法，來爲每 20 個有聲音框選取出最大似然率的一個或兩個分段 GMM，接著採用 3.2 節裡的高斯混和選取方法，來爲各輸入音框選取出單一個高斯混合，再依據公式(7)作對映。

當把一個來源語者的發音檔，分別輸入到前述的三個語音轉換系統，我們就可得到三個轉換出語音檔。然後使用轉換出的音檔，我們進行了兩種類型的聽測實驗，分別是音色相似度之聽測、和語音品質之聽測。在這二類型的聽測實驗裡，我們都邀請了 25 位人士來聆聽音檔並給予相對分數，而在這 25 位人士中，有 20 位是不熟悉語音轉換之研究的。

4.1 音色相似度測試

首先我們準備了 5 個音檔，它們的代號分別是 VS(由來源語者發音)，VT(由目標語者發音)，VX1(經由 SOG 系統轉換得到)，VX2(經由 SSG 系統轉換得到)，VX3(經由 SLG 系統轉換得到)，其中 VS 與 VT 具有相同的說話內容，而 VX1、VX2 和 VX3 三者也有相同的內容，但不同於 VS 和 VT 的，這 5 個音檔可從網頁 http://guhy.csie.ntust.edu.tw/VoiceConv/去下載。在進行聽測實驗時，我們以 ABX 的次序來撥放前述的音檔，在此 A 固定爲 VS，B 固定爲 VT，而 X 則隨機由 VX1、VX2 和 VX3 三者中選出，每次以 ABX 次序播放完音檔後，受測者就被要求給一個分數。在此分數的定義是，9 分(或 1 分)表示 X 的音色確定就是 B(或 A)的音色，7 分(或 3 分)表示 X 的音色比較接近 B(或 A)的音色，而 5 分表示 X 的音色無法判斷是接近 A 或接近 B。

做完聽測實驗之後，25 位受測者所給的分數被用來計算出三個系統各自的平均分數(AVG)和標準差(STD)，所得到的分數數值就如表 1 所列出的。由表一的平均分數可知，不同性別之間的語音轉換(即從 M1 到 F1)，會比同性別之間的(即從 M1 到 M2)獲得明顯較高的分數。此外，拿三個系統的平均分數作比較，可從表一的數值得知，SLG 系統的表現明顯比 SSG 系統的好許多(7.05 vs 6.24，7.60 vs 7.24)，而 SSG 系統的表現則是比 SOG 系統的稍微好一些(6.24 vs 6.08，7.24 vs 6.92)。所以本論文提出的分段式 GMM 之觀念及自動 GMM 挑選之演算法，的確可幫忙改進所轉換出語音的音色相似度。

表一、音色相似度聽測之平均分數與標準差

		SOG	SSG	SLG
M1=>M2	AVG	6.08	6.24	7.05
	STD	1.11	1.09	0.93
M1=>F1	AVG	6.92	7.24	7.60
	STD	1.13	1.07	1.10

4.2 語音品質測試

在此我們使用三個系統轉換出的語音檔 VX1、VX2 和 VX3，來進行語音品質的聽測實驗。音檔撥放的次序爲 AX，A 固定設爲 VX1，而 X 則隨機由 VX2 和 VX3 兩者中取出，每次以 AX 次序播放完音檔後，受測者就被要求給一個分數。在此分數的定應是，9 分(或 1 分)表示 X 的語音品質明顯比 A 的好(或差)，7 分(或 3 分)表示 X 的品質比 A 的稍微好(或差)一些，5 分則表示 X 和 A 的語音品質無法分辨優劣。

作完聽測實驗之後，我們收集 25 位受測者所給的分數，來計算出 SSG 和 SLG 兩系統各自的平均分數和標準差，結果得到的數值如表二裡列出的。依據表二的平均分數可看出，同性別之間(即從 M1 到 M2)的轉換語音的品質，會比不同性別之間(即從 M1 到 F1)

的較好約 0.5 分，這顯示不同性別之間的轉換語音的品質，是比較難作改進的。此外，依據 SLG 和 SSG 兩系統的平均分數作比較，我們可看出 SLG 的分數都比 SSG 的高約 0.7 分，並且 SLG 的平均分數都高於 5 分，所以分段式 GMM 之觀念及自動挑選 GMM 之演算法，確實可用以改進所轉換出語音的語音品質。

表二、語音品質聽測之平均分數與標準差

		SSG vs SOG	SLG vs SOG
M1=>M2	AVG	5.23	6.04
	STD	1.43	1.45
M1=>F1	AVG	4.89	5.55
	STD	1.50	1.47

4.3 倒頻譜距離量測

在所錄音的 375 句平行語料中，最後 25 句並未被用於訓練 GMM 模型，因此這 25 句來源語者發音的語音檔，在此就分別被輸入到三個語音轉換系統 SOG、SSG 和 SLG，去作語音轉換的處理，以便量測轉換出語音和目標語音(目標語者發音)之間的倒頻譜距離，用以作爲轉換後頻譜和目標頻譜之間的接近程度的客觀量測。

對於轉換出的語音音檔的每一個有聲音框，我們先依先前作 DTW 時間對齊的資料，來找出目標語者發音檔中對應的音框，然後將兩對應音框的 DCC 係數，拿去計算幾何距離，接著再依所有有聲音框量測到的距離去計算出平均距離，結果對於三個語音轉換系統，我們計算出的平均距離就如表三裡所列出的。

表三、轉換後語音的平均倒頻譜距離

	SOG	SSG	SLG
M1=>M2	0.543	0.609	0.601
M1=>F1	0.598	0.634	0.612

依據表三列出的數值，可發現 SOG 系統會得到最小的平均距離，然而由聽測實驗的結果可知，SOG 系統在音色相似度方面是最差的，並且在語音品質方面也是比 SLG 系統差，如此的不一致性，其原因應是公式(1)裡的加權和運算，會導致於對映出的頻譜變得過度平滑化，而造成音質變差。另一方面，SLG 系統比起 SSG 系統所表現出的效能改進，則是有反應在所量測出的平均距離上，SLG 比 SSG 多增加了選取分段式 GMM 之處理步驟。

4.4 分段 GMM 選取之例子

當我們使用 3.1 節的方法，來爲有聲的音框序列挑選似然率最高的分段 GMM 時，有時會發生挑錯 GMM 的情況，一個例子如圖三所示，此音框序列爲/song-4/ (“送”)的發音，第一欄數字表示音框的編號，第二欄數字表示各音框偵測出的音高，第三欄數字表示各音框的對數能量值，第四欄資料則是我們的演算法爲各音框所挑選出的分段 GMM(以對應的韻母表示)。觀察第四欄的 GMM 挑選結果可以發現，在音框編號 543 時會發生韻

母的切換，從韻母/ong/ (/ㄨㄥ/) 切換到韻母 /eng/ (/ㄥ/)，但是正確的答案應是不切換韻母的，即此序列的音框都要挑選到韻母/ong/。不過這兩個韻母(/ong/與/eng/)的發音，其共同點是具有相同的韻尾音素/ng/，所以我們認爲類似圖三的韻母挑選節果是可以接受的。

Frame	Pitch	Energy	Vowel
526	0	58.99	
527	0	56.65	
528	185	63.61	ong
529	185	67.59	ong
530	182	70.43	ong
531	170	71.34	ong
532	165	70.41	ong
533	160	70.83	ong
534	156	71.16	ong
535	149	70.57	ong
536	146	69.66	ong
537	141	68.95	ong
538	136	67.60	ong
539	130	64.12	ong
540	128	64.08	ong
541	125	63.74	ong
542	121	62.00	ong
543	116	61.37	eng
544	113	60.39	eng
545	113	59.01	eng
546	164	58.54	eng
547	164	56.07	eng
548	135	50.28	eng
549	0	51.49	
550	0	47.81	

圖三、語音/song-4/的有聲音框之 GMM 挑選情形

五、結論

依據聽測實驗的結果，我們可說 SLG 系統是三個系統之中效能最好的，不管是在音色相似度、還是在語音品質上都表現得最好，因此 SLG 系統所採用的處理方法，即本論文提出的分段式 GMM 之觀念及自動 GMM 挑選之演算法，經由聽測實驗的驗證，的確可用以改進 GMM 爲基礎的語音轉換機制。另一方面，依據客觀量測出的平均倒頻譜距離，可知使用原始 GMM 轉換方法之 SOG 系統，仍然可得到三個系統中最小的平均距離，不過 SOG 系統轉換出的語音，在音色相似度上卻是表現最差的，並且在語音品質上也是比 SLG 系統的差。目前我們僅根據韻母來作語音的分段與分群，將來可再考慮把有聲聲母(如/m/, /n/, /l/)的部分獨立切成語音段，這樣應可進一步改進語音轉換的效能。

參考文獻

[1] Y. Stylianou, O. Cappe, and E. Moulines, ”Continuous probabilistic transform for voice conversion,” *IEEE trans. Speech and Audio Processing*, vol. 6, no. 2, pp.131–142.1998.

[2] T. Toda, H. Saruwatari, and K. Shikano, “Voice conversion algorithm based on Gaussian

mixture model with dynamic frequency warping of STRAIGHT Spectrum," *Int. Conf. Acoust., Speech, and Signal Processing*, Salt Lake City, pp. 841-844, 2001.

[3] T. Toda and A. W. Black and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE trans. Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222-2235, 2007.

[4] Z. H. Jian and Z. Yang, "Voice Conversion Using Viterbi algorithm based on Gauaaian mixture model", *Int. Symposium on Intelligent Signal Processing and Communication Systems*, pp. 32-35, Xiamen, China, 2007.

[5] Z. Yue, X. Zou, Y. Jia, and H. Wang, "Voice conversion using HMM combined with GMM", *2008 Congress on Image and Signal Processing*, Sanya, China, pp. 366-370, 2008.

[6] E. Godoy, O. Rosec, and T. Chonavel, "Alleviating the one-to-many mapping problem in voice conversion with context-dependent modeling", *Proc. INTERSPEECH*, pp. 1627-1630, Brighton, UK, 2009.

[7] O. Cappé and E. Moulines, "Regularization techniques for discrete cepstrum estimation," *IEEE Signal Processing Letters*, vol. 3, no. 4, pp. 100-102, 1996.

[8] H. Y. Gu and S. F. Tsai, "A discrete-cepstrum based spectrum-envelope estimation scheme and its example application of voice transformation," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 14, no. 4, pp. 363-382, 2009.

[9] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195-239, 1984.

[10] H. Y. Kim, et al., "Pitch detection with average magnitude difference function using adaptive threshold algorithm for estimating shimmer and jitter," 20-th Annual *Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, Hong Kong, China, 1998.

[11] Y. Stylianou, *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.

[12] H. Kawahara, I. Masuda-katsuse, and A. De. Cheveign, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187-207, 1999.