以線性多變量迴歸來對映分段後音框之語音轉換方法

A Voice Conversion Method Mapping Segmented Frames with Linear Multivariate Regression

古鴻炎 Hung-Yan Gu　張家維 Jia-Wei Chang　王讚緯 Zan-Wei Wang

國立臺灣科技大學 資訊工程系
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
e-mail: {guhy, m9815064, m10015078}@mail.ntust.edu.tw

摘要

基於 GMM 對映之語音轉換方法常遇到的一個問題是，轉換出的頻譜包絡會發生過於平滑(over smoothing)的現象，因此本論文嘗試以線性多變量迴歸(linear multivariate regression, LMR)來建構另一種頻譜對映的方法，希望能夠改進頻譜過平滑的問題。首先，我們推導了 LMR 對映矩陣的解析求解公式，然後我們錄製平行語料，採用離散倒頻譜係數作為頻譜特徵，分割語音信號成聲、韻母之音段，再使用 LMR 對映方法來建造出一個語音轉換系統。應用此系統，我們就可進行內部、外部之平均轉換誤差的量測，並且和傳統 GMM 對映法所量測出的誤差距離作比較，量測的結果顯示，本論文研究的 LMR_F 對映法，不論是在內部或外部之測試情況，都可以獲得比傳統 GMM 對映法較小的平均轉換誤差。此外，我們也進行了主觀的語音品質聽測之實驗，聽測實驗的結果顯示，我們研究的 LMR_F 對映法，其轉換出的語音品值，能夠比傳統 GMM 對映法的稍好一些。

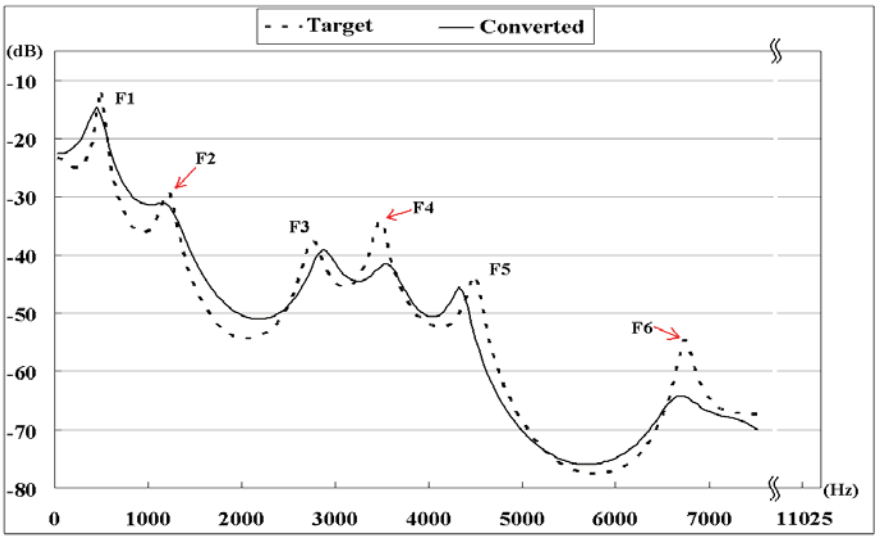
關鍵詞：語音轉換，線性多變量迴歸，高斯混合模型，離散倒頻譜係數

一、緒論

語音轉換(voice conversion)研究的目標是，要把一個來源語者(source speaker)的語音轉換成另一個目標語者(target speaker)的語音。這種語音轉換的處理，可應用於銜接語音合成處理，以獲得多樣性的合成語音音色，此外亦可應用於作戲劇配音的處理，以讓一個配音員可以為多個角色配音。過去在語音轉換領域，先前研究者提出的轉換方法包括了:頻譜特徵之向量量化(VQ)對映(mapping)[1]，共振峰(formant)頻率對映[2, 3]，基於高斯混合模型(Gaussian mixture model, GMM)之對映[4, 5]，基於類神經網路(artificial neural network, ANN)之對映[6]，基於隱藏式馬可夫模型(hidden Markov model, HMM)之對映[7, 8]等。

最近幾年有不少研究者採取基於 GMM 對映之方向來作語音轉換，並且嘗試去解決

原始 GMM 對映方式[4]所碰到的問題，例如轉換出的頻譜包絡(spectral envelope)會出現過於平滑(over smoothing)的現象，一個例子如圖一所示，虛線曲線代表目標語者一個音框的頻譜包絡，實線曲線則代表由來源語者音框轉換出的頻譜包絡，明顯可看出虛線曲線的 F2、F4、F6 等共振峰(formant)的頻寬變寬了很多，也就是山鋒至山谷的深度減少了，這種過於平滑的頻譜包絡，將使得據以合成出的語音信號，發生語音品質衰退的情況，也就是語音聽起來，會讓人覺得悶悶的、不夠清晰。



圖一、過於平滑之轉換出的頻譜包絡

為了避免發生頻譜過於平滑的情況，而造成音質的衰退，在此論文裡我們遂決定採取以最小均方(least mean square, LMS)誤差為準則，去研究線性多變量迴歸(linear multivariate regression, LMR)方式的頻譜對映方法，希望用以提升轉換出語音的音質。線性多變量迴歸對映(簡稱為 LMR 對映)的觀念是，在訓練階段使用平行語料，以訓練出一個 $d \times d$ 的線性對映矩陣 M， d 表示一個音框頻譜特徵係數的維度，然後在轉換階段，就可將來源語者第 k 個音框的頻譜特徵向量 S_k (維度為 $d \times 1$)，作 LMR 對映而得到轉換出的頻譜特徵向量 V_k，即令 $V_k = M \cdot S_k$。雖然 Valbret 等人已於 1992 年提出使用 LMR 對映來作頻譜轉換的想法[9]，但是他們對於前述矩陣 M 的數值的求解，只提出了一個逼近的作法，因此在本論文裡，我們遂去研究、推導矩陣 M 的解析(analytic)求解公式，詳細情形在第二節裡說明。

另外，我們由前人的研究得知[5, 10]，所採取的頻譜對映機制，如果不先依據語音內容(如音素或音節)來建立分段式(segmental)的對映模型，則容易發生一對多(one to many)對映的問題[10]，而造成某些相鄰的音框之間，相鄰音框所轉換出的頻譜卻出現劇烈的頻譜形狀差異(即頻譜不連續)，以致於怪音(artifact sound)被合成出來。為了減少發生怪音的機會，因此我們決定以聲、韻母為單位，對訓練用的語音作音段切割，並且各音段(segment)裡的語音音框就交由所屬之聲、韻母去收集，然後使用各個聲、韻母所收集到的音框，去分別訓練出專屬的 LMR 對映矩陣。至於在轉換階段，一個輸入的語音音框如何知道它是屬於那一個聲、韻母的？這樣的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，因為選取到錯誤但近似的聲、韻母是可以容忍的。過去，我們研究分段式 GMM 對映之語音轉換方法[5]，曾提出一種自動挑選音段 GMM 的演算法，該演算法也可以搬過來使用。

關於頻譜係數的選擇，我們仍然採取先前研究過的離散倒頻譜係數(discrete

cepstrum coefficients, DCC)[11, 12]，階數設為 40 階，即一個音框要計算出 $c_0, c_1, c_2, \ldots, c_{40}$ 等 41 個係數，但是只拿 $c_1, c_2, \ldots, c_{40}$ 去作頻譜轉換的處理，所以維度 d 的值是 40。當轉換出各個音框的 DCC 係數之後，我們就可依據各音框的 DCC 係數去計算出頻譜包絡[11, 12]，然後再依據頻譜包絡、轉換出的基頻值，去設定該音框的諧波加雜音模型(harmonic plus noise model, HNM)之諧波參數和雜音參數[12, 13]，之後就可拿這些參數去合成出語音信號 [12, 13]。

二、LMR 對映矩陣

在訓練階段，把平行訓練語料各音框算出 DCC 係數之後，再經由動態時間校正(DTW)，就可得知一個來源語者音框(來源音框)所對應的目標語者音框(目標音框)。在此令某一個聲、韻母類別所收集到的 N 個來源音框是: $S_1, S_2, \cdots, S_N$，而其對應的 N 個目標音框是: $T_1, T_2, \cdots, T_N$，也就是 d×1 大小之 DCC 向量 T_k 經由 DTW 被匹配到 DCC 向量 S_k。為了方便推導，我們在此令矩陣 $S = [S_1, S_2, \cdots, S_N]$，而矩陣 $T = [T_1, T_2, \cdots, T_N]$，很明顯地矩陣 S 和 T 的大小都是 d×N。理想上，我們希望找出一個大小為 d×d 的 LMR 對映矩陣 M，來讓如下的關係式獲得成立，

$$M \bullet S = T \quad . \tag{1}$$

實際上，由於 N 的值通常都比 d 大很多，所以不會存在理想的 M 矩陣，也就是會出現對映的誤差，在此令 E 表示大小為 d×N 之誤差矩陣，其定義是

$$E = M \bullet S \; - \; T \, . \tag{2}$$

若要找出最佳的對映矩陣 M，就相當於要把矩陣 E 的所有元素的絕對值都加以最小化。由於矩陣 E 有 d×N 個元素，而矩陣 M 只有 d×d 個元素，所以我們採取 LMS 準則，先去計算誤差平方和矩陣 $\boldsymbol{\mathcal{E}}$ ，其定義是:

$$\boldsymbol{\mathcal{E}} = E \cdot E^{\mathrm{t}} = (M \cdot S - T)(M \cdot S - T)^{\mathrm{t}}, \quad \mathrm{t}: \text{transpose.} \tag{3}$$

然後拿 $\boldsymbol{\mathcal{E}}$ 的跡數(trace)，即 $\mathrm{tr}(\boldsymbol{\mathcal{E}}) = \boldsymbol{\mathcal{E}}_{1,1} + \boldsymbol{\mathcal{E}}_{2,2} + \ldots + \boldsymbol{\mathcal{E}}_{d,d}$ ，去對 M 作偏微分，並且令偏微分的結果為 0 矩陣 [11, 12]，公式如下，

$$\frac{\partial(\mathrm{tr}(\boldsymbol{\mathcal{E}}))}{\partial M} = 2(M \cdot S - T) \cdot S^{\mathrm{t}} = 0 \quad , \tag{4}$$

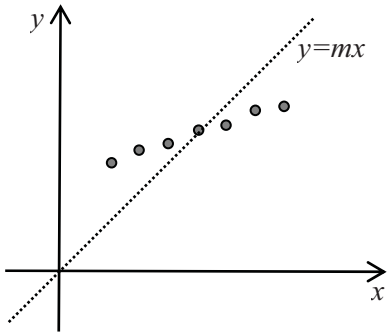
上式中矩陣形式之 $\partial(\mathrm{tr}(\boldsymbol{\mathcal{E}})) / \partial M$ 其實是表示 $\partial(\mathrm{tr}(\boldsymbol{\mathcal{E}})) / \partial M_{i,j}$ ，j=1, 2, …, d，i=1, 2, …, d，也就是分別拿 M 矩陣第 i 列第 j 行的元素 $M_{i,j}$ 去對 $\mathrm{tr}(\boldsymbol{\mathcal{E}})$ 作偏微分。公式(4)經過移項整理後，就可解出 M 的數值，公式如下，

$$M \cdot S \cdot S^{\mathrm{t}} = T \cdot S^{\mathrm{t}} \, , \tag{5}$$

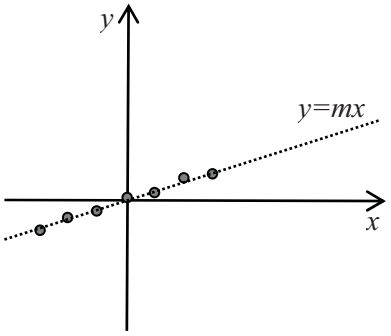
$$M = T \cdot S^{\mathrm{t}} \cdot (S \cdot S^{\mathrm{t}})^{-1} \quad . \tag{6}$$

現在，我們已可使用公式(6)來找出 LMS 準則下局部最佳的 M 矩陣，說它僅是局部最佳，其原因可以圖二(a)所示的單變量線性迴歸的例子來說明，也就是如公式(1)的 M

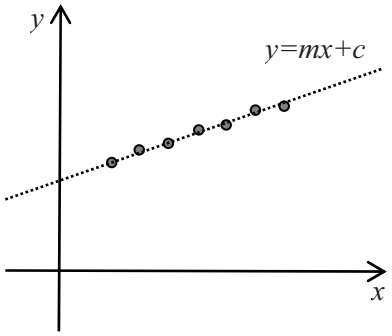
矩陣的定義，相當於是在單變量線性迴歸情況下，限定迴歸之直線必須通過原點，因此迴歸所導入的誤差，會比圖二(b)情況或圖二(c)情況的都大。若要作改善，第一種作法是，設法把圖二(a)的情況轉變成圖二(b)的情況，如此公式(6)則仍然可繼續使用，作轉變的實際方法是，先計算出來源音框 $S_1, S_2, \cdots, S_N$ 的平均向量 S^m，再以 $S_k - S^m$ 取代原先的 S_k，同樣地對於目標音框 $T_1, T_2, \cdots, T_N$，也要去計算平均向量 T^m，然後作類似的取代。採用此種作法時，平均向量 S^m 與 T^m 必須儲存下來，如此在轉換階段才可拿出來使用。



(a) $y = m \cdot x$



(b) $y = m \cdot x$



(c) $y = m \cdot x + c$

圖二、單變量線性迴歸例子

在本論文裡，我們不希望另外作儲存平均向量的動作，因此研究了把圖二(a)情況轉變成圖二(c)情況的作法，也就是要導入常數項。我們想到的一個作法是，先依照下列公式把原先公式(1)裡的矩陣 M、S、T 的定義作擴充，

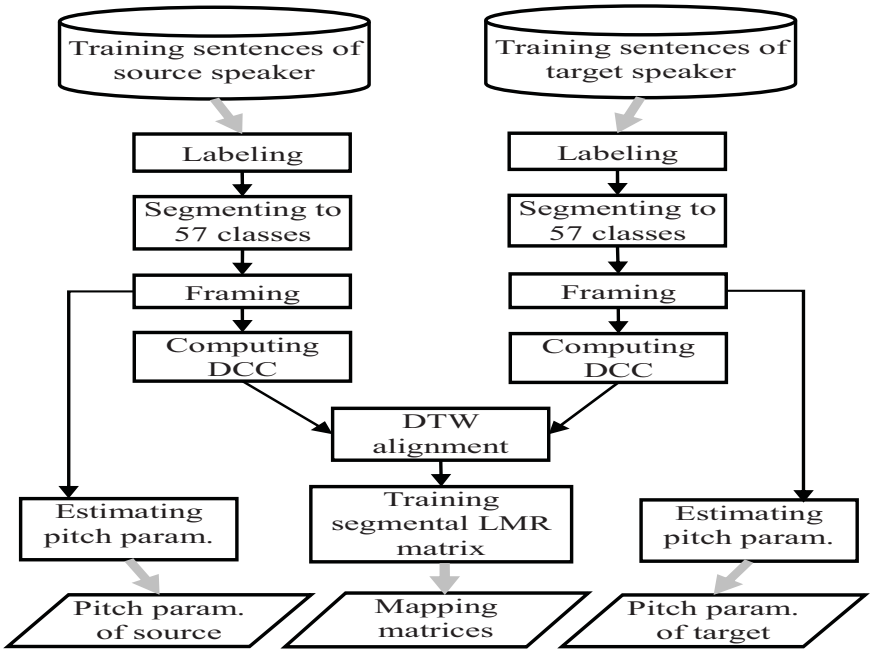
$$\tilde{M} = \begin{bmatrix} & & M_{1,d+1} \\ & M & M_{2,d+1} \\ & & : \\ & & M_{d,d+1} \\ 0, 0, ..., 0, & & 1 \end{bmatrix}, \qquad \tilde{S} = \begin{bmatrix} S_1 & S_2 & \ldots & S_N \\ 1, & 1, & \ldots & 1 \end{bmatrix}, \qquad \tilde{T} = \begin{bmatrix} T_1 & T_2 & \ldots & T_N \\ 1, & 1, & \ldots & 1 \end{bmatrix}, \tag{7}$$

第一步把 M 矩陣擴充成大小為$(d+1)\times(d+1)$之 $\tilde{M}$ 矩陣，亦即在原先的 M 矩陣裡加入第$(d+1)$列和第$(d+1)$行，而新增的元素如公式(7)所示；然後在 S 矩陣內加入第$(d+1)$列，並且把該列的元素值全設為常數 1，因此擴充後的 $\tilde{S}$ 矩陣大小為$(d+1)\times N$；接著以類似的擴充方式也把 T 矩陣擴充成 $\tilde{T}$ 矩陣。之後，就可以把擴充後的 $\tilde{M}$ 、 $\tilde{S}$ 、 $\tilde{T}$ 矩陣代入公

式(6)，去求取 $\tilde{M}$ 矩陣的數值。如此，當應用求得的 $\tilde{M}$ 矩陣於公式(1)時，就可以讓線性迴歸所導入的誤差減小。

三、系統製作 -- 訓練階段

我們製作的語音轉換系統，在訓練階段主要的處理步驟如圖三所示。首先我們邀請了二位男性和二位女性錄音者，其中二位男性，在此以 M_1 和 M_2 作代號，而另二位女性，則以 F_1 和 F_2 作代號。我們請四位錄音者分別到隔音錄音室去錄製 375 句(共 2,926 個音節)之國語平行語料，取樣率設成 22,050Hz。在本論文裡，我們實驗了四種語者配對方式，分別是(a)M_1 至 M_2、(b)M_1 至 F_1、(c)F_1 至 M_1、(d)F_1 至 F_2，這四種配對方式裡，前者就當來源語者，而後者則當目標語者。



圖三、訓練階段之主要處理步驟

3.1 標音與切割音段

對於各個語者所錄的訓練語句(即前 350 句之平行語句)，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，把一個語句的各個聲母、韻母的邊界標示出來。由於自動標記的聲、韻母邊界有許多是錯誤的，因此我們再操作 WaveSurfer 軟體，以人工檢查自動標記的邊界是否有錯，有錯則加以更正。

接著，依據各個聲、韻母的拼音符號標記和邊界位置，就可作音段切割和分類的動作。對於各個訓練語句，依據其所屬的標記檔案，一一讀出各個音段(即聲、韻母)的資訊，就可依拼音符號將該音段作分類，我們一共分成 57 類(21 類聲母和 36 類韻母)，分類後再將該音段所在的語句編號、時間邊界資料寫出至分類記錄檔案。

3.2 DCC 係數計算

在本論文裡，我們採用離散倒頻譜之頻譜包絡估計方法[11, 12]，並且以 DCC 係數作為

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[12]來計算出 41 維的 DCC 係數。在此一個音框的長度設為 512 個樣本點(23.2ms)，而音框位移則設為 128 個樣本點(5.8ms)。

3.3 DTW 匹配和 LMR 矩陣計算

由於平行語料已經過音段切割和分類，所以在此就逐一對各個聲、韻母類別所收集的平行發音音段作 DTW 匹配，再依匹配出的音框對應序列去計算各類別的 LMR 對映矩陣。由於來源語者和目標語者的發音速度會有差異，因此對於兩人發音同一個句子所取出的平行音段(如/a/)，必需先作 DTW 匹配，以便為來源語者音段所切出的各個音框 S_k，去目標語者之平行音段內找出正確的音框來對應。如此，經由平行音段之間作 DTW 匹配，就可建立兩語者的平行音段內的音框對應關係(S_k, $T_{w(k)}$)，k=1, 2, …, K_n，K_n 表示第 n 個平行音段之來源語者發音的音框數量。接著，把各個平行音段的音框對應關係作串接，就可求得一個聲、韻母類別的一序列的來源音框和目標音框的對應組合。

關於 LMR 矩陣的求取，在此也是逐一對各個聲、韻母類別去計算，先把各類別求得的一序列的來源音框和目標音框的對應組合，拿去建造如公式(1)裡的 S 和 T 矩陣，然後代入公式(6)以計算出基本型 LMR 對映所需的 M 矩陣。此外，我們也依據公式(7)，把矩陣 S 和 T 擴充成 $\tilde{S}$ 和 $\tilde{T}$ ，再代入公式(6)，以算出完整型 LMR 對映所需的 $\tilde{M}$ 矩陣。

3.4 音高參數

我們先計算零交越率(ZCR)，以把 ZCR 很高的無聲(unvoiced)音框偵測出來；再使用一種基於自相關函數及 AMDF (absolute magnitude difference function)的基週偵測方法[14]，來偵測剩餘音框的音高頻率。之後，把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，據以算出該語者音高的平均值及標準差，而平均值及標準差就是本論文所使用的音高參數。

四、系統製作 -- 轉換階段

我們製作的語音轉換系統，在轉換階段的主要處理流程如圖四所示。當一句測試語句輸入後，它首先會被切割成一序列的音框，至於音框長度和位移則和 2.2 節裡使用的一樣，分別是 512 點和 128 點。然後，在圖四的左邊流程，系統會去偵測各音框的音高頻率，如果一個音框被偵測為無聲時，圖四中的三個灰色方塊就被直接跳過，也就是不作音高頻率的調整，且 DCC 頻譜參數也不會被轉換。相對地如果一個音框被偵測為有聲時，系統就會使用如下的音高調整公式，

$$q_t = \mu^{(y)} + \frac{\sigma^{(y)}}{\sigma^{(x)}}(p_t - \mu^{(x)}) \tag{8}$$

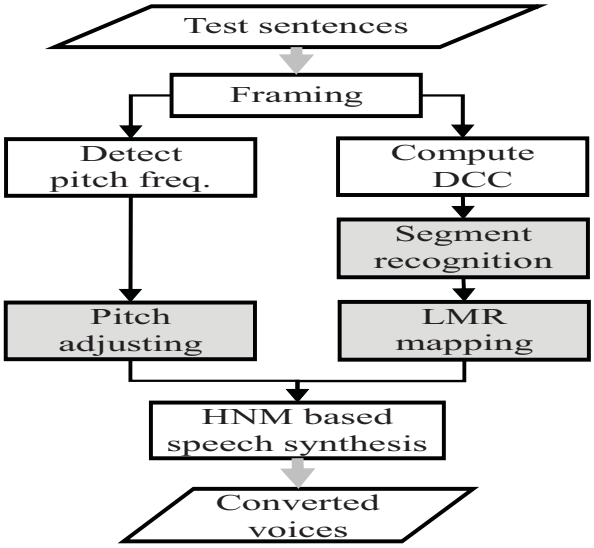
來調整音高頻率，其中 p_t 表示偵測出的音高頻率值，$\mu^{(x)}$和 $\sigma^{(x)}$分別表示來源語者音高頻率的平均值和標準查，而 $\mu^{(y)}$和 $\sigma^{(y)}$則是目標語者的。

4.1 聲、韻母音段辨識

當進行實驗以比較不同型式的 LMR 對映矩陣時，重點是放在 LMR 矩陣本身，所以我

們跳過此步驟(聲韻母音段辨識)的處理，而直接依據各語句所屬的標記檔案，來讀出各音段的拼音標記和時間邊界資料。

如果要處理一個線上即時輸入的語句，那麼"聲韻母音段辨識"步驟就必須實際地執行，關於這個步驟的製作，目前我們是透過呼叫 HTK 所提供的辨識命令來達成。不過，在能夠呼叫 HTK 的辨識命令之前，要先操作 HTK 的 HMM 訓練命令，以便拿 350 句來源語者的訓練語句去訓練出各個聲、韻母的 HMM 模型。



圖四、轉換階段之主要處理步驟

4.2 基於 HNM 之語音信號合成

在諧波加雜音模型(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為最大有聲頻率(maximum voiced frequency，MVF)[13]。為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 值都直接設為 6,000Hz。

使用 HNM 來對轉換出的頻譜包絡作語音信號合成，觀念上是分別去合成出諧波部分的信號，及合成出雜音部分的信號，然後把兩部分的信號加總，即是所合成的語音信號。由於我們在先前發表的論文裡[5, 12]都已說明 HNM 語音信號合成方法的細節，所以在本論文裡就不再重複敘述。

五、測試實驗

在第二節中我們說明了兩種 LMR 對映的作法，第一種作法是，採取如公式(1)定義的 M 矩陣來作為對映的矩陣，這種作法稱為基本型 LMR 對映，在此以 **LMR_B** 表示；至於第二種作法是，採取如公式(7)定義的 $\tilde{M}$ 矩陣來作為對映的矩陣，這種作法稱為完整型 LMR 對映，在此以 **LMR_F** 表示。

此外，我們也研究了一種把向量量化和 LMR 對映作結合的作法，稱為 **LMR_FC**，

該作法的細節是，訓練階段時，在圖三中的”DTW alignment”和”Train LMR matrix”兩方塊之間，增加一個”VQ clustering”方塊，先對一個聲韻母類別所收集到的 DCC 接合向量(joint vector, 維度 80)作 K-means 分群的處理，以分成 L 群的 DCC 接合向量，並且記錄 L 群向量的中心向量，之後對各群的 DCC 向量分別去訓練出一個對應的 LMR 對映矩陣 $\tilde{M}$ ，在此我們只將群數 L 設為 4，因為設太多群時，有一些聲韻母會發生音框數過少的情況。

接著在轉換階段時，圖四中的”LMR mapping”方塊之前就必須增加一個”Select mapping matrix”方塊，以從 L 個對映矩陣中挑選出一個，我們採取的挑選方法是，將輸入音框的 DCC 向量(維度 40)和訓練階段記錄下來的 L 個中心向量的前 40 維，逐一量測幾何距離，然後把距離最小的那個中心向量所對應的對映矩陣，選取出來再用以作 LMR 對映。

5.1 誤差距離量測

由於 375 句平行語句中，只有前 350 句拿去訓練 LMR 對映矩陣，因此對於轉換出的 DCC 向量和目標 DCC 向量之間的誤差距離，我們分成內部測試(使用前 350 句)和外部測試(使用後 25 句、共 209 個音節)兩種情況分別去量測。設 $R = R_1, R_2, \cdots, R_N$ 為一序列被轉換出的 DCC 向量，而 $T = T_1, T_2, \cdots, T_N$ 為 R 所對應的目標 DCC 向量序列，在此我們以如下公式,

$$D_{avg} = \frac{1}{N} \sum_{1 \le k \le N} dist(R_k, T_k), \tag{9}$$

去量測轉換誤差之平均距離，公式(9)中 $dist(\ )$表示幾何距離之量測函數。

對於前述三種對映方法，我們分別在內部測試與外部測試兩種情況下，去量測四組語者配對各自的平均轉換誤差距離，然後再取四組語者配對之平均轉換誤差的平均值，結果得到如表一所列的數值。從表一前二欄的數值可知，完整型的 LMR 對映方法(LMR_F)比起基本型的對映方法(LMR_B)，不論在內部測試或外部測試皆可讓轉換誤差減小(分別是 1.6%和 1.7%)，這樣的改進和我們預期的一致；不過，比較表一後二欄的數值，我們發現內部測試和外部測試出現不一致的情況，結合 VQ 和 LMR 對映的方法(LMR_FC)，在內部測試時獲得了非常顯著的改進，平均轉換誤差由 0.4956 降低至 0.4672，即改進 5.7%，然而在外部測試時，平均轉換誤差卻由 0.5382 變大成 0.5493，即變差了 2.1%。另一個觀點是，我們覺得 LMR_FC 法之內部測試的平均誤差值 0.4672 有一個含意，它表示將來我們有機會把外部測試的平均誤差再加以改進至 0.5 以下；相對來說，LMR_B 法之內部測試的平均誤差值 0.5038 已經很大，應不可能直接用 LMR_B 法去把外部測試的平均誤差值改進至 0.5 以下。

另外，為了和 GMM 為基礎的對映方法作比較，在此我們也使用相同的語者配對語料和相同維度的 DCC 頻譜係數，去訓練出傳統 GMM 對映模型[4]的參數，以及音段式 GMM 對映模型(Segmental GMM) [5]的參數，其中傳統 GMM 對映模型使用 128 個高斯分佈，而每一種音段的音段式 GMM 模型則使用 8 個高斯分佈。然後，在內部測試與外部測試兩種情況下，我們分別去量測四組語者配對各自的 GMM 對映模型的平均轉換誤差距離，然後再取四組語者配對之平均轉換誤差的平均值，結果得到如表二所列的數值，雖然從表二的轉換誤差平均值可發現，音段式 GMM 對映模型的轉換誤差，不論在內部或外部測試情況，都會比傳統 GMM 對映模型的小，但是，本論文研究的完整型

LMR 對映法(LMR_F)，則更進一步地讓轉換誤差平均值減小了，比較表一 LMR_F 法的誤差值和表二列出的誤差值，可知 LMR_F 法在內部測試情況，能夠將轉換誤差改進 7.1%(比傳統 GMM 法)、和 4.5%(比音段式 GMM 法)，而在外部測試情況，則能夠將轉換誤差改進 1.5%、和 0.7%。因此，對於分段後的語音音框，LMR 為基礎的對映方法，確實可用於改進語音轉換的誤差。

表一、三種 LMR 對映方法之平均轉換誤差

平均轉換誤差		LMR_B	LMR_F	LMR_FC
內部測試	M_1=> M_2	0.4890	0.4794	0.4475
	M_1=> F_1	0.4782	0.4705	0.4451
	F_1 => M_1	0.4967	0.4881	0.4612
	F_1 => F_2	0.5514	0.5443	0.5149
	平均	**0.5038**	**0.4956**	**0.4672**
外部測試	M_1=> M_2	0.5467	0.5331	0.5398
	M_1=> F_1	0.5174	0.5106	0.5188
	F_1 => M_1	0.5388	0.5307	0.5413
	F_1 => F_2	0.5867	0.5782	0.5973
	平均	**0.5474**	**0.5382**	**0.5493**

表二、兩種 GMM 對映模型之平均轉換誤差

平均轉換誤差		GMM (128 mix.)	Segmental GMM (8 mix.)
內部測試	M_1=> M_2	0.5058	0.5096
	M_1=> F_1	0.5012	0.4910
	F_1 => M_1	0.5412	0.5095
	F_1 => F_2	0.5853	0.5673
	平均	**0.5334**	**0.5194**
外部測試	M_1=> M_2	0.5346	0.5403
	M_1=> F_1	0.5147	0.5146
	F_1 => M_1	0.5551	0.5361
	F_1 => F_2	0.5806	0.5766
	平均	**0.5463**	**0.5419**

5.2 語音品質聽測

我們使用未參加模型訓練的來源語者語句，來準備 6 個作語音品質聽測的音檔，它們的代號分別是 X1、X2、Y1、Y2、Z1、Z2，在此 X1 與 X2 表示使用傳統 GMM 對映模型[4]所轉換出的音檔，Y1 與 Y2 表示使用 LMR_F 對映方法所轉換出的音檔，而 Z1 與 Z2 表示使用 LMR_FC 對映方法所轉換出的音檔；此外，代號 X1、Y1、Z1 中的 1 表示使用 M_1 至 M_2 之語者配對的語料去訓練模型參數，而代號 X2、Y2、Z2 中的 2 表示使用 M_1 至 F_1 之語者配對的語料去訓練模型參數。這 6 個音檔可從如下網頁去下載試聽: http://guhy.csie.ntust.edu.tw/VCLMR/LMR.html。

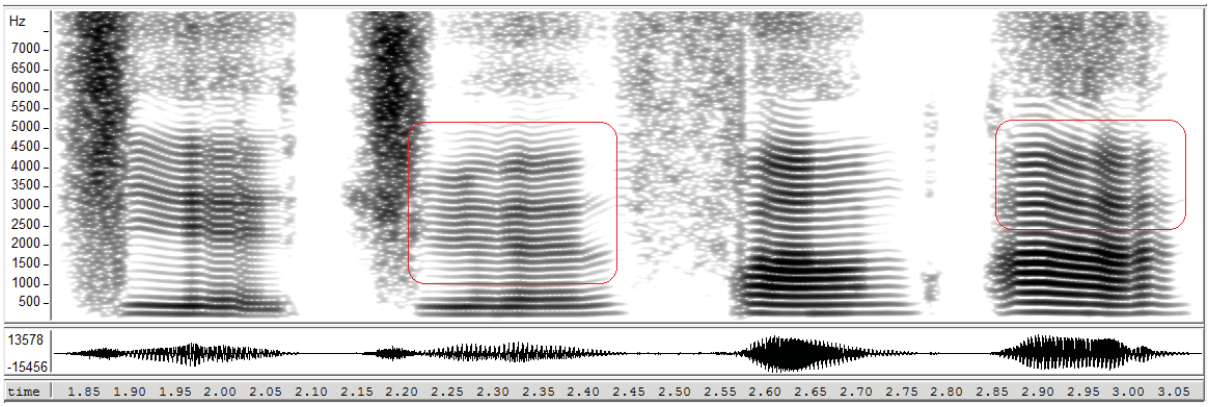
使用這 6 個音檔，我們編排成四次的聽測實驗，第一次實驗裡，隨機指派 X1、Y1 成為 A 與 B 音檔，然後依序播放 A、B 音檔給受測者聽，再要求受測者給一個評分，以顯示 B 音檔的語音品質比起 A 音檔的是好或壞；第二次實驗裡，隨機指派 Y1、Z1 成為 A 與 B 音檔，然後播放給受測者聽；第三次實驗裡，隨機指派 X2、Y2 成為 A 與 B 音檔，然後播放給受測者聽；第四次實驗裡，則隨機指派 Y2、Z2 成為 A 與 B 音檔，然後播放給受測者聽。在四次聽測實驗裡，受測者都是同樣的 15 位學生，他們大部分都不熟悉語音轉換之研究領域，至於評分的標準是，2 (-2)分表示 B (A)音檔的語音品質比 A (B)音檔的明顯地好，1 (-1)分表示 B (A)音檔的語音品質比 A (B)音檔的稍為好一點，0 分表示分辨不出 A、B 兩音檔的語音品質。

在四次聽測實驗之後，我們將受測者所給的評分作整理，結果得到如表三所示的平均評分。從表三第一欄的平均評分(即 0.867 與 0.467)可發現，和傳統 GMM 對映方法比起來，本論文研究的 LMR_F 對映方法能夠轉換出品質稍好一些的語音；另外，從表三第二欄的平均評分(即 0.267 與 0.000)可發現，LMR_F 對映法和 LMR_FC 對映法，兩者所轉換出語音的品質，不能被感覺出有差異，雖然我們自己聽音檔後覺得，LMR_FC 對映法所轉換出語音的品質要比 LMR_F 對映法的稍好一些。

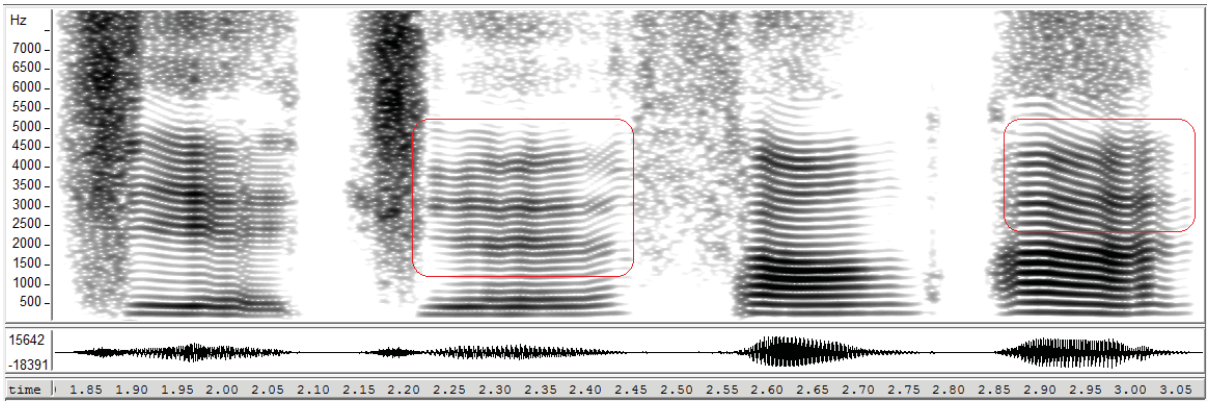
表三、語音品質聽測之平均評分

平均評分		GMM (128mix.) vs LMR_F	LMR_F vs LMR_FC
M_1 => M_2	AVG (STD)	0.867 (0.640)	0.267 (0.915)
M_1 => F_1	AVG (STD)	0.467 (0.704)	0.000 (0.378)

對於前一段得到的語音品質聽測之結果，在此我們嘗試以聲譜圖(spectrogram)來解釋其原因。當使用傳統 GMM 對映法來對一個來源語句作轉換，語句內的四個字(“解決方案”)所轉換出語音的聲譜就如圖五(a)所顯示的；而當使用 LMR_FC 對映法來對相同的來源語句作轉換，則得到如圖五(b)所示的聲譜圖。比較圖五(a)和(b)可發現，圖五(b)裡的共振峰(formant)條紋比圖五(a)裡的清晰，例如第二個字“決”的共振峰條紋，在圖五(b)裡的峰、谷對比(即黑、白顏色的對比)顯得較強烈，而在圖五(a)裡的峰、谷對比，就相對地比較緩和，因此，圖五(b)對應的語音聽起來會比圖五(a)的清晰一些。



(a) 傳統 GMM 法轉換出語音之聲譜圖



(b) LMR_FC 法轉換出語音之聲譜圖

圖五、兩種方法轉換/jie-3 jyei-2 fang-1 an-4/(“解決方案”)之聲譜圖

六、結論

本論文嘗試以線性多變量迴歸(LMR)作為頻譜對映之機制，去建構出一個語音轉換的系統，並且我們推導了 LMR 對映矩陣的解析求解公式。在使用平行語料、DCC 頻譜係數、和語音信號先分割成聲、韻母音段的情況下，我們經實驗測試發現，LMR_F 對映法進行語音轉換所導入的平均誤差距離值，不論在內部或外部測試之情況，都可以獲得比傳統 GMM 對映法更小的誤差距離值，內部測試時，平均的轉換誤差比起傳統 GMM 對映法的改進了 7.1%，而在外部測試時，平均的轉換誤差則比傳統 GMM 對映法的改進了 1.5%。此外，我們也進行了主觀的語音品質聽測之實驗，實驗的結果顯示，我們研究的 LMR_F 對映法，其轉換出的語音品值，可以比傳統 GMM 對映法的稍好一些。

另外，我們自己試聽轉換出的語音，發現 LMR_F 對映法轉換出的語音聽起來仍有一些模糊的感覺，我們認為這是因為轉換出的頻譜仍存在過平滑的現象，就像傳統 GMM 對映法所遇到的。不過，當使用 LMR_FC 對映法時，這樣的模糊感覺可以減少一些，LMR_FC 對映法能夠轉換出比較清晰的語音，我們覺得它的解釋是，LMR_FC 對映法裡要先作向量量化分群，而分群可以讓頻譜相近的音框聚集在一起，如此就可以減少發生頻譜過平滑的現象。LMR_FC 對映法導入的轉換誤差，內部測試時會比 LMR_F 對映法的小許多，但是外部測試時則比 LMR_F 對映法的大一些，因此，將來可再繼續研究對 LMR_FC 對映法作改進。

參考文獻

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice Conversion through Vector Quantization,” *Int. Conf. Acoustics, Speech, and Signal Processing*, New York, Vol. 1, pp. 655-658, 1988.

[2] H. Mizuno and M. Abe, “Voice Conversion Algorithm Based on Piecewise Linear Conversion Rules of Formant Frequency and Spectrum Tilt,” *Speech Communication*, Vol. 16, No. 2, pp. 153-164, 1995.

[3] 吳嘉彧、王小川，”不需平行語料而基於共振峰與線頻譜頻率映對之語者特質轉換系統”，*第二十一屆自然語言與語音處理研討會*(ROCLING 2009)，台中，第 319-332 頁，2009。

[4] 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.

[5] H. Y. Gu and S. F. Tsai, “An Improved Voice Conversion Method Using Segmental GMMs and Automatic GMM Selection”, *Int. Congress on Image and Signal Processing*, pp. 2395-2399, Shanghai, China, 2011.

[6] S. Desaiy, E. V. Raghavendray, B. Yegnanarayanay, A. W Blackz, and K. Prahallad, “Voice Conversion Using Artificial Neural Networks,” *Int. Conf. Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, pp. 3893-3896, 2009.

[7] E. K. Kim, S. Lee, and Y. H. Oh, “Hidden Markov Model Based Voice Conversion Using Dynamic Characteristics of Speaker,” *Proc. EuroSpeech*, Rhodes, Greece, Vol. 5, 1997.

[8] C. H. Wu, C. C. Hsia, T. H. Liu, and J. F. Wang, “Voice Conversion Using Duration-Embedded Bi-HMMs for Expressive Speech Synthesis,” *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 14, No. 4, pp. 1109-1116, 2006.

[9] H. Valbret, E. Moulines, J. P. Tubach, “Voice Transformation Using PSOLA Technique,” *Speech Communication*, Vol. 11, No. 2-3, pp. 175-187, 1992.

[10] 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.

[11] O. Cappé and E. Moulines, “Regularization Techniques for Discrete Cepstrum Estimation,” *IEEE Signal Processing Letters*, Vol. 3, No. 4, pp. 100-102, 1996.

[12] 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.

[13] 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.

[14] 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.