基於音段式 LMR 對映之語音轉換方法的改進

Improving of Segmental LMR-Mapping Based Voice Conversion Methods

古鴻炎 張家維
Hung-Yan Gu Jia-Wei Chang

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}@mail.ntust.edu.tw

摘要

基於線性多變量迴歸(linear multivariate regression, LMR)頻譜對映之語音轉換方法，轉換出的頻譜包絡仍然存在過度平滑(over smoothing)的現象，因此本論文研究在音段式 LMR 頻譜對映之前加入直方圖等化(HEQ)的處理，並且在 LMR 頻譜對映之後加入目標音框挑選的處理，希望藉以提升轉換出語音的品質。在此，直方圖等化處理包含兩個步驟，首先是把離散倒頻譜係數(DCC)轉換成主成分分析(PCA)係數，接者把 PCA 係數轉換成累積密度函數(CDF)係數；目標音框挑選則是依據一個音框的音段類別編號、及 LMR 對映出的 DCC 向量，到目標語者相同音段類別所收集的音框群中，去搜尋出距離較小的目標語者 DCC 向量、並且取代原先對映出的 DCC 向量，如此以避免發生頻譜包絡之過度平滑現象。對於直方圖等化與目標音框挑選，我們以外部(未參加模型參數訓練)平行語料來量測語音轉換之平均 DCC 誤差，當加入直方圖等化後會使誤差值變大一些，而當加入目標音框挑選後則會使誤差值變大得更多。不過，VR (variance ratio)值量測及主觀聽測的結果卻是相反的方向，亦即直方圖等化可使語音品質提升一些，而目標音框挑選則可使語音品質獲得更為明顯的提升。這種誤差距離值和語音品質聽測之間的不一致性，我們設法去尋找了它的原因，所找到的一個理由在內文裡說明。

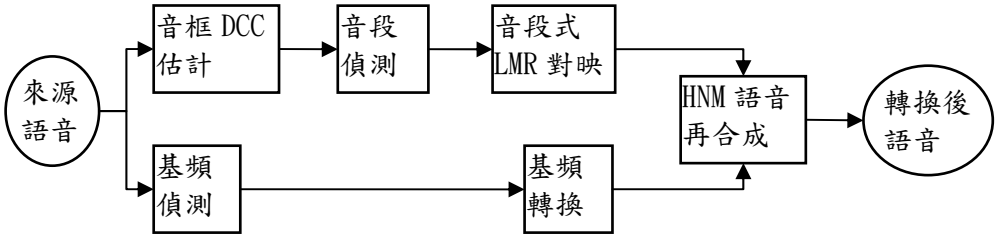
關鍵詞：語音轉換，線性多變量迴歸，直方圖等化，目標音框挑選，離散倒頻譜係數

一、緒論

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，這種處理稱為語音轉換(voice conversion)[1, 2, 3]，語音轉換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。去年我們曾嘗試以線性多變量迴歸(linear multivariate regression, LMR)來建構一種頻譜對映(mapping)的機制[4]，然後用於作語音轉換，希望藉以改進傳統上基於高斯混合模型(Gaussian mixture model, GMM)之頻譜對映機制[3]常遇到的一個問題，就是轉換出的頻譜包絡(spectral envelope)會發生過度平滑(over smoothing)的現象。我們經由實驗發現，音段式(segmental) LMR 頻譜對映機制不僅在平均轉換誤差上可以比傳統 GMM 頻譜對映機制獲得一些改進，並且轉換出語音的音質也

比傳統 GMM 對映的稍好一些。不過，整體而言音段式 LMR 對映機制所轉換出的頻譜包絡，仍然存在有過於平滑的現象，而使得轉換出的語音仍然令人覺得有一些悶悶的，而不像真人發音那樣清晰。前面提到的"音段式" LMR，是指我們對於訓練語料中不同的韻母、有聲聲母(如/m, n, l, r/)的語音要分別去建立各自的 LMR 矩陣，這是為了避免發生一對多(one to many)對映的問題[5]，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)，而不連續的頻譜很可能導致怪音(artifact sound)被合成出來。

去年我們研究的基於 LMR 頻譜對映之語音轉換系統，其主要的處理流程如圖一所

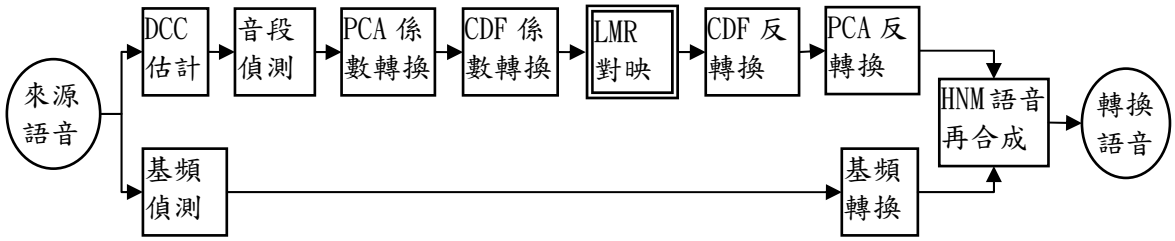


圖一、基於 LMR 頻譜對映之語音轉換的主要處理流程

示，來源語者發出的語音先分割成一序列的音框，然後對各個音框去估計它的 40 階 DCC (discrete cepstral coefficients) 倒頻譜係數[6, 7]及偵測出基頻值；接著，依據各音框的 DCC 係數，可作有聲聲母與韻母的音段(segment)偵測，先前我們曾提出一種基於音段式 GMM 與最大似然率(maximum likelihood)的音段自動偵測方法[8]，實驗顯示即使挑選到錯誤但近似的音段，也仍可轉換出正確的語音，由於在此我們把焦點放在 LMR 對映方塊，所以音段偵測方塊暫時以讀取標記(label)檔案的方式來進行；LMR 對映就是把 LMR 矩陣乘以輸入的 DCC 向量而求得輸出的 DCC 向量，至於 LMR 矩陣的訓練方法，則可參考我們去年發表的論文[4]；之後，LMR 對映出的 DCC 向量，以平均值與標準差轉換出的基頻值，兩者就可送給 HNM (harmonic plus noise model)語音再合成方塊，以合成出轉換後的語音信號，關於使用諧波加雜音模型(HNM)作語音信號合成的細節，可參考前人的論文[9, 7]。

為了提升轉換出的語音的音質，我們開始思考在 GMM 對映與 LMR 對映之外，是否還有其它種類的對映方法？後來我們想到一種似乎可行的頻譜對映方法，就是以直方圖等化(histogram equalization, HEQ)來取代 LMR 對映。直方圖等化雖然起源於影像處理領域，但是近年來被應用於語音辨識領域[10, 11]，用以降低環境噪音造成的訓練語音和測試語音之間的頻譜不匹配(mismatch)問題，而使得辨識率獲得了明顯的改進。有鑑於此，我們覺得在語音轉換的問題上，來源與目標語者之間有著差異的頻譜形狀而呈現出差異的音色，這可想像是因為來源語音通過了某一種特殊的通訊通道而使得其頻譜形狀被轉換成目標語音的形狀，以致於造成來源與目標語音之間的頻譜不匹配。因此在觀念上應可應用直方圖等化的處理，來模仿前述的通訊通道之特性，以把來源語音的頻譜轉變成目標語音的頻譜，所以我們構想了如圖二所示的基於直方圖等化之語音轉換的處理流程。

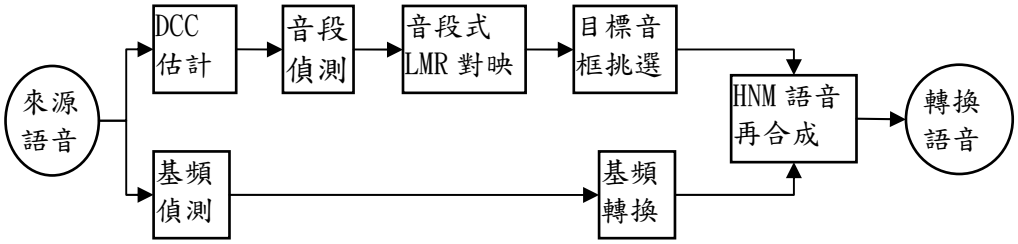
在圖二的處理流程中，我們不直接拿 DCC 係數去作直方圖等化，即計算 CDF (cumulative density function)係數，我們的觀點是，一個音框各維度的 DCC 係數之間有



圖二、基於直方圖等化之語音轉換的處理流程

明顯的相關性存在，而直方圖等化卻是對特徵的各維度獨立去進行，這恐將降低直方圖等化的功用，因此我們決定對各個音段類別所屬的音框 DCC 向量先進行主成分分析(principle component analysis, PCA) [12]，再依據主成分向量把 DCC 係數轉換成 PCA 係數，如此將可讓一個音框各維度的 PCA 係數之間變成是獨立的。此外，圖二中的 LMR 對映方塊，一開始時是未被加入的，不過經由初步的測試實驗發現，當沒有作 LMR 對映的處理時，轉換出語音的音色雖可達到部分近似目標語者的音色，但是仍存在明顯的音色落差，因此我們遂決定把 LMR 對映方塊加上去，以提升音色相似度。

對於圖一處理流程會遇到的頻譜包絡過於平滑的情況，雖然前人曾經提出至少兩種的改進方法，即全域變異數(global variance, GV)之變異數調整方法[13]、和頻率軸校正(frequency warping)的方法[14, 15]，但是 Toda 等人的方法[13]和 Erro 等人的方法[14]都是針對 GMM 對映所設計的，而 Godoy 等人的方法[15]則不是針對 GMM 對映或 LMR 對映所設計的。因此我們就從另外一個方向去思考圖一流程的改進作法，在參考 Dutoit 等人的論文[16]之後，我們想到的一個作法是，在圖一"LMR 對映"方塊之後插入"目標音框挑選"的方塊。既然經過 GMM 或 LMR 對映得到的頻譜包絡會發生過度平滑的現象，那麼就不要直接拿 LMR 對映得到的頻譜係數去作語音再合成處理，而要改變成依據**來源音框**(來源語者音框)的音段類別、及對映出的頻譜特徵係數(如 DCC)，去對同一音段類別的**目標音框**(目標語者音框)群作搜尋，以找出頻譜特徵最相似(或距離最小)的目標音框，然後把找出的目標音框的頻譜係數拿去取代對映出的頻譜係數，如此就可免除發生頻譜包絡過度平滑的問題。由於被找出的目標音框不是經由頻譜對映而得到，所以在此也稱它為**真實音框**(真實語音的音框)，此外，目標音框的音段分類與收集是在訓練階段進行，所以轉換階段就可直接去作搜尋與挑選。當圖一插入"目標音框挑選"的方塊之後，一種基於 LMR 對映及目標音框挑選之改進的語音轉換處理流程就如圖三所示。



圖三、基於 LMR 對映及目標音框挑選之語音轉換的處理流程

除了分別去加入直方圖等化和目標音框挑選的處理動作，我們也考慮了另外一種處理流程，就是同時把這兩種處理動作加入圖一的處理流程中，如此轉換出的語音是否可以獲得最好的音色相似度及語音品質？這將會第四節中作實驗探討。此外，在圖一、二、

三裡都出現的 DCC 估計之方塊，表示我們仍然採用離散倒頻譜係數(DCC)[6, 7]作為頻譜特徵參數，並且階數設為 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。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[6, 7]，然後再依據頻譜包絡、轉換出的基頻值，去設定該音框的 HNM 模型之諧波參數和雜音參數[7, 9]，之後就可拿這些參數去合成出語音信號 [7, 9]。

二、PCA 係數轉換與直方圖等化

若要依據圖二的處理流程來進行語音轉換的處理，則各音框在求取 DCC 係數之後，接著就要作 PCA 係數轉換和 CDF 係數轉換的動作，然後在 LMR 對映之後，還要作 PCA 反轉換和 CDF 反轉換的動作，以將頻譜特徵還原成 DCC 係數。因此，在這一節就說明 PCA 係數轉換和 CDF 係數轉換的細節。

(一)、PCA 係數轉換

要能夠把一個來源音框的 DCC 係數轉換成 PCA 係數，則在訓練階段就要先對來源語者各個音段類別所收集到的 DCC 向量作 PCA 分析，以求取來源語者各個音段類別的主成分向量。相對地，要能夠把一個 LMR 對映後音框的 PCA 係數反轉換成 DCC 係數，則在訓練階段也要先對目標語者各個音段類別所收集到的 DCC 向量作 PCA 分析，以求取目標語者各個音段類別的主成分向量。然而關於 PCA 分析的作法，我們曾經思索的一個疑問是，雖然直覺上我們會認為來源音框和目標音框應該要分開去收集，並且分開去作 PCA 分析以求取各自的主成分向量，但是，為什麼不能夠把同一音段類別的來源音框和目標音框放在一起去作 PCA 分析？又為什麼不讓來源音框和目標音框共用一組主成分向量呢？因此，我們將以實驗評估的方式來探討此一疑問。

PCA 分析是由 K. Pearson 於 1901 年提出，在 1933 年時再由 H. Hotelling 加以發展[17]。PCA 轉換是一種正交變換，它可以將原本維度間相關的原始數據轉換成各維度獨立的新數據，再者作 PCA 轉換後的新數據，它們的總變異數(variance)與原始數據集的總變異數相等，也就是說 PCA 轉換能保留原始數據的訊息。

1、主成分分析

對於某一音段類別的所有訓練語音作音框切割及求取 DCC 係數，以建立一個 40 維 DCC 係數的數據集，接著再對這個數據集作 PCA 分析以得到該種音段的主成分向量，詳細的分析流程如下:

(a) 假設某一音段類別的訓練語音總共可切成 M 個音框，而每個音框經由計算可得到一個 DCC 係數的向量，然後把全部音框的 DCC 向量並列成各欄(column)的方式，表示成大小為 $L \times M$ 的矩陣 $\Gamma=[\Gamma_1, \Gamma_2, ..., \Gamma_M]$，其中 L 表示 DCC 係數的階數，M 的值大於 L。

(b) 接著求出這 M 個音框之 DCC 向量的平均向量 Ψ，Ψ 代表著這 M 個音框共有的 DCC 向量成分。

(c) 將第 i 個音框的 DCC 向量作標準化，即減去平均向量 Ψ，而得到一個差值向量 Φ_i。

(d) 使用所有的差值向量Φ_i，來計算出一個共變異矩陣Λ。

$$\Lambda = \sum_{i=1}^{M} \Phi_i \Phi_i^{\mathrm{T}} \tag{1}$$

(e) 對矩陣Λ求其特徵值(eigen value)λ_i與特徵向量(eigen vector)γ_i。

$$\Lambda \cdot \gamma_i = \lambda_i \cdot \gamma_i, \quad i = 1,2,...,L \tag{2}$$

(f) 求得特徵向量γ_i後，進一步對γ_i作正規化，以取得L個主成分基底向量μ_i。

$$\begin{aligned} \upsilon_i &= \sqrt{(\gamma_{i1})^2 + (\gamma_{i2})^2 + \cdots + (\gamma_{iL})^2}, \quad i = 1,2,...,L \\ \mu_i &= \left[\frac{\gamma_{i1}}{\upsilon_i}, \frac{\gamma_{i2}}{\upsilon_i}, \cdots \frac{\gamma_{iL}}{\upsilon_i} \right]^{\mathrm{T}}, \quad i = 1,2,...,L \end{aligned} \tag{3}$$

2、主成分係數轉換

當我們對某一個音段類別做完主成分分析後，就可得到該類別的 DCC 平均向量Ψ、L個主成分基底向量μ_i。接著，要把各個音框的 DCC 係數轉換成 PCA 係數，首先把一個音框的 DCC 向量Γ_i減去 DCC 平均向量Ψ而得到差值向量Φ_i，再將Φ_i分別投影到各個主成分基底向量μ_i，投影公式為:

$$\omega_{ij} = \mu_j^{\mathrm{T}} \cdot \Phi_i, \quad j = 1,2,...,L \tag{4}$$

如此就可得到 DCC 向量Γ_i的L個主成分係數(亦稱為 PCA 係數)，再用以形成L維度的主成分係數(PCA 係數)之向量：

$$\Omega_i = [\omega_{i1}, \omega_{i2}, ..., \omega_{iL}]^{\mathrm{T}} \tag{5}$$

3、主成分係數反轉換

在圖二的處理流程中，“PCA 反轉換”方塊就是要將轉換後的 PCA 係數還原到 DCC 係數的向量空間，以得到轉換後的 DCC 係數。假設我們取得一序列音框的 PCA 係數之向量，則首先要知道各個音框分別所屬的音段類別，如此才能對各個音框分別去作還原，令第i個音框所屬的音段類別之編號為k，則我們就要取出訓練階段目標語者在第k類音段所計算出的 DCC 平均向量Ψ、及L個主成分基底向量μ_j，來把轉換後的 PCA 向量Ω_i還原成轉換後的 DCC 向量Γ_i，如公式(6)所示:

$$\Gamma_i = \Psi + \sum_{j=1}^{L} \mu_j \cdot \omega_{ij} \tag{6}$$

(二)、直方圖等化

直方圖等化所指的是圖二流程裡“CDF 係數轉換” 與 “CDF 反轉換”兩方塊的處理。要

能夠把一個來源音框的 PCA 係數轉換成 CDF 係數，則在訓練階段就要先對來源語者各個音段類別所收集到的 PCA 向量作 HEQ 分析，以建造來源語者各個音段類別的 HEQ 表格。相對地，要能夠把一個 LMR 對映後音框的 CDF 係數反轉換成 PCA 係數，則在訓練階段也要先對目標語者各個音段類別所收集到的 PCA 向量作 HEQ 分析，以建造目標語者各個音段類別的 HEQ 表格。這裡提到 HEQ 表格，意謂我們採取基本的表格法來建立 PCA 係數和 CDF 係數之間的直方圖等化關係。

1、HEQ 表格建造

選定一個來源(或目標)語者的音段類別，令該類別裡收集到的音框總數為 M，則將 M 個維度為 L 的 PCA 係數向量作為輸入資料，依照下列步驟來建造 HEQ 表格:

(a) 令區間數為 N，並且對各個維度 i，i=1, 2, …, L，分別作下列步驟的處理。

(b) 將 M 個音框中所有位於第 i 維度的 PCA 係數挑出，然後依係數值作由小到大之排序，排序後則把 M 個 PCA 係數依順序且平均地分配到 N 個區間。

(c) 區間編號 j 從 1 變到 N，對於第 j 個區間內的 PCA 係數，挑選排序位於中間(median)的 PCA 係數數值，然後記錄該 PCA 係數值為 Fp_i^j，並且記錄其對應的 CDF 值為 Fc_i^j，CDF 值就是該 PCA 係數在全體(M 個)係數排序中的順序值除以 M。

(d) 記錄第 i 維度 PCA 係數的最大值為 Fp_i^{N+1}，且記錄其對應的 CDF 值為 $Fc_i^{N+1}=1$；此外，記錄第 i 維度 PCA 係數的最小值為 Fp_i^0，且記錄其對應的 CDF 值為 $Fc_i^0=\frac{1}{M}$。

當所有維度都完成上述步驟，則該音段類別的 HEQ 表格就建立完成了。對於區間數 N 的選擇，我們在評估實驗裡嘗試了 32, 64, 128 等三種。HEQ 表格建造後的外觀為何？在此舉一個簡化的例子，設有 20 個音框，PCA 係數向量維度為 1 維，且 PCA 係數序列排序後為 1, 2, …, 20，若設定的區間數為 N=4，則建造出的 HEQ 表格如下所列。

表一、一個簡化的 HEQ 表格例子

區間 j	0	1	2	3	4	5
Fp_1^j	1(min)	3	8	13	18	20(max)
Fc_1^j	0.05	0.15	0.4	0.65	0.9	1

2、CDF 係數轉換

假設有一個音框的 PCA 係數向量 $P=[P_1, P_2, \cdots, P_L]$ 要被轉換，而該音框所屬的音段類別資訊，已經在圖二的“音段偵測”方塊決定出來，所以我們可以取出該音段類別的來源音框所訓練出的 HEQ 表格，然後以線性內插的方式來計算出該音框的 CDF 係數向量 $Q=[Q_1, Q_2, \cdots, Q_L]$，線性內插之公式如下：

$$Q_i = Fc_i^j + (Fc_i^{j+1} - Fc_i^j)\cdot\left[\frac{(P_i - Fp_i^j)}{(Fp_i^{j+1} - Fp_i^j)}\right], \quad i=1,2,...,L. \tag{7}$$

公式(7)中 i 表示維度編號，Fp_i^j 、Fc_i^j 分別為 HEQ 表格裡所記錄的第 j 區間的 PCA 係數值、CDF 值，並且假設我們已作過搜尋而得知 P_i 的值落於 Fp_i^j 與 Fp_i^{j+1} 之間。

3、CDF 反轉換

假設有一個音框的 CDF 向量 $Q=[Q_1, Q_2, \cdots, Q_L]$ 要被反轉換成 PCA 係數向量，而該音框所屬的音段類別資訊，已經在圖二的"音段偵測"方塊決定出來，所以我們可以取出該音段類別的目標音框所訓練出的 HEQ 表格，然後以線性內插的方式來計算出該音框的 PCA 係數向量 $P=[P_1, P_2, \cdots, P_L]$，線性內插之公式如下：

$$P_i = Fp_i^j + (Fp_i^{j+1} - Fp_i^j) \cdot \left[\frac{(Q_i - Fc_i^j)}{(Fc_i^{j+1} - Fc_i^j)} \right], \quad i = 1, 2, ..., L. \tag{8}$$

公式(8)中 i 表示維度編號，Fp_i^j 、Fc_i^j 分別為 HEQ 表格裡所記錄的第 j 區間的 PCA 係數值、CDF 值，並且假設我們已作過搜尋而得知 Q_i 的值落於 Fc_i^j 與 Fc_i^{j+1} 之間。

三、目標音框挑選

在訓練階段，我們可預先把目標語者的訓練語音依據標示檔的資訊拿去作音段分類、及對各種音段分別作音框的收集，之後在轉換階段，就可依據所偵測出的音段代號去取出對應的音框集，再依據所轉換出的 DCC 向量去作真實音框的搜尋與挑選。

令 $Y_1, Y_2, \cdots, Y_T$ 是一序列 T 個被轉換出的 DCC 向量，轉換可以是直接經由圖三"LMR 對映"方塊得到，或是 LMR 對映後再作 CDF 反轉換與 PCA 反轉換而得到(圖二的流程)。為了改進轉換出的語音的品質，所以在此要依據 Y_t 及其對應的音段類別代號 $I(t)$，從目標語者的 $I(t)$音段的音框集去挑選出一個非常靠近 Y_t 的真實音框的 DCC 向量 Z_t。然而挑選 Z_t 的準則，不僅只是考慮 Y_t 與 Z_t 的匹配距離 dist(Y_t, Z_t)，也要考慮相鄰音框之間的連接距離 dist(Z_{t-1}, Z_t)，以避免發生頻譜之不連續，而導致怪音被合成出來。在本論文裡，距離函數 dist(·, ·)是量測幾何距離。除了依循 Dutoit 等人的論文[16]去考慮音框連接的距離，我們還更加考慮了另外一種距離量測，即動態頻譜(dynamic spectral)距離，以把轉換出的相鄰兩 DCC 向量之間的頻譜改變 $\Delta Y_t = Y_t - Y_{t-1}$ 納入考慮。在此，動態頻譜距離是量測 dist(ΔY_t, ΔZ_t)，而 $\Delta Z_t = Z_t - Z_{t-1}$ 表示相鄰兩個挑選出的 DCC 向量之間的頻譜改變。

依據前述的三種距離，即匹配距離、連接距離與動態頻譜距離，我們發展了一種基於動態規劃的演算法來作目標音框的挑選。首先，對於各個轉換出的 DCC 向量 Y_t，我們依其音段編號 $I(t)$，從第 $I(t)$個音框集去尋找出 K 個最靠近 Y_t (即離 Y_t 的距離最小)的真實音框的 DCC 向量，在此 K 的值設為 16。接著，令 $U(t, i)$表示從時刻 1 到時刻 t 的最小的累積距離，而條件是在時刻 t 時所挑選到的目標音框必須是 K 個中的第 i 個。如此，我們就可得到如下的遞迴公式:

$$U(t,i)=\min_{0\le j<K}\left[U(t-1,j)+\alpha\cdot dist\left(Z_{t-1}^{j},Z_{t}^{i}\right)+\alpha\cdot dist\left(Y_{t}-Y_{t-1},Z_{t}^{i}-Z_{t-1}^{j}\right)\right]+dist\left(Y_{t},Z_{t}^{i}\right),\quad(9)$$

其中 α 是加權常數，我們經過試驗後將它的值設為 0.5，Z_t^i 表示時刻 t 時所尋找出的 K 個音框中的第 i 個音框 DCC 向量。另外，前人論文[16]中曾提到一個技巧，當 Z_t^i 和 Z_{t-1}^j 被檢查出是來自同一次發音的相鄰音框時，就機動地把公式(9)中 α 的值改設為 0，以便優先選取相鄰的目標音框來提升頻譜連接的自然性。在此我們也應用了這個技巧，並且把條件放寬，就是當 Z_t^i 和 Z_{t-1}^j 不是直接相鄰而是存在另一個音框在它們之間，我們也接受此一情況而會把 α 的值機動地改設為 0。

當到達最後時刻 T 時，全部路徑中的最小累積距離 A(T)可以下列公式來計算，

$$A(T)=\min_{0\le j<K}\left[U(T,j)\right]\quad,\qquad(10)$$

此外，我們可再作回溯(backtrack)處理，以找出在最佳路徑上各個時刻 t 所選到的目標音框編號 $k(t)$，然後把 t 時刻所選到的第 $k(t)$個目標音框的 DCC 向量，拿去取代被轉換出的 DCC 向量 Y_t。

四、測試實驗

我們邀請了二位男性和二位女性錄音者，其中二位男性以 MA 和 MB 為代號，而另二位女性則以 FA 和 FB 為代號。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，這 375 句的語料中，前 350 句被拿來作模型參數的訓練之用，而剩下的 25 句則保留作為外部測試之用。在此我們實驗了四種語者配對方式，分別是(a)MA 至 MB、(b)MA 至 FA、(c)FA 至 MA、(d)FA 至 FB，這四種配對方式中，前者就當來源語者，而後者則當目標語者。

(一)、語音轉換系統之訓練

首先，我們操作 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，然後操作 WaveSurfer 軟體，以檢查自動標記的邊界是否有錯，有錯則作人工更正。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，我們一共分成 57 類，即 21 類聲母和 36 類韻母。

對於各個語音音框，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 的基週偵測方法[18]，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。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)。此外，對於一個音框的頻譜係數，我們使用先前發展的 DCC 估計程式[7]來計算出 41 維的 DCC 係數。

在訓練 LMR 對映矩陣之前，我們逐一對各個聲、韻母類別所收集的平行發音音段

作 DTW 匹配，以便為來源語者音段所切出的各個音框，去目標語者之平行音段內找出正確的音框來對應。然後，把各個平行音段的音框序列串接起來，就可為一個聲、韻母類別準備好一序列的來源音框和目標音框的 DCC 向量對應組合，(S_i, R_i)，i=1, 2, …, Nr，其中 S_i 表示第 i 個來源音框的 DCC 向量，R_i 表示第 i 個經 DTW 配對到的目標音框的 DCC 向量，Nr 表示此一序列的音框總數。再來，依照所建構系統的結構，若是如圖三的流程，則各個聲、韻母類別的一序列的來源與目標音框對應的 DCC 向量組合，就可直接拿去訓練計算 LMR 對映所需的對映矩陣[4]；然而當系統的結構是如圖二所示的流程時，則各個聲、韻母類別的 DCC 向量組合序列，(S_i, R_i)，i=1, 2, …, Nr，其中各個組合的 S_i 與 R_i 就必須先作 PCA 係數轉換和 CDF 係數轉換，以形成 CDF 係數的向量組合，然後才拿去訓練 LMR 對映之映矩陣。

設 $\tilde{S}$ 、 $\tilde{R}$ 矩陣的定義如下所列，

$$\tilde{S} = \begin{bmatrix} S_1 & S_2 & \dots & S_{Nr} \\ 1, & 1, & \dots & 1 \end{bmatrix}, \quad \tilde{R} = \begin{bmatrix} R_1 & R_2 & \dots & R_{Nr} \\ 1, & 1, & \dots & 1 \end{bmatrix}, \tag{11}$$

其中各行的 S_i 與 R_i 都被附加一列的常數 1，以增加一個常數項至多變量線性迴歸的各個維度裡，如此，LMR 對映所需的最佳(least squared error)對映矩陣 $\tilde{M}$ ，就可以下列公式[4]來求得，

$$\tilde{M} = \tilde{R} \cdot \tilde{S}^{\mathrm{t}} \cdot (\tilde{S} \cdot \tilde{S}^{\mathrm{t}})^{-1} \quad . \tag{12}$$

然後，我們就可用矩陣 $\tilde{M}$ 來作 LMR 對映，即令$[Y^{\mathrm{t}}, 1]^{\mathrm{t}} = \tilde{M} \cdot [X^{\mathrm{t}}, 1]^{\mathrm{t}}$，其中 X 表示一個來源語者音框的 DCC 或 CDF 係數向量，而 Y 表示經由 LMR 對映出的係數向量。

(二)、共用主成分向量之測試

圖二的處理流程裡，PCA 係數轉換與 PCA 反轉換兩個處理方塊，若讓兩者共用一組主成分向量是否會比較好？原先不共用主成分向量的情況，表示“PCA 係數轉換”方塊使用的主成分是由**來源**音框作完音段分類後再作 PCA 分析得到，而“PCA 反轉換”方塊使用的主成分則是由**目標**音框作完音段分類後再作 PCA 分析得到；若是共用主成分向量，就表示同一音段類別的來源音框和目標音框要放在一起去作 PCA 分析，以求得共用的一組主成分向量。

我們以量測語音轉換的平均轉換誤差的方式，來比較共用與不共用主成分向量之優劣。在此，我們只拿平行語料最後的 25 句來作語音轉換之外部測試，當一個來源音框經過轉換而得到 DCC 向量之後，我們就可量測此 DCC 向量與對應的目標音框 DCC 向量之間的幾何距離，這樣的距離也稱為轉換誤差，當把全部音框的轉換誤差加總及取平均，就可算出平均的轉換誤差。此外，我們也把圖二流程裡的直方圖等化(即 CDF 係數轉換與反轉換)分成三種情況來作實驗，就是分別設定區間的數量 N 為 32、64、與 128，經過實驗量測後，我們得到如表二所示的平均轉換誤差值。

從表二的轉換誤差平均值可以看出，圖二中的 PCA 係數轉換與反轉換方塊若是使用共用的 PCA 主成分向量，則平均轉換誤差可從 0.5447 降到 0.5414，這說明了使用共用的 PCA 主成分向量，可以略微提升來源與目標音框之間 PCA 係數的相關性，而稍微減小 LMR 對映的誤差。此外，關於直方圖等化的區間數的設定，依據表二的轉換誤差平均值可知，設為 64 區間或 128 區間是沒有差異的。

表二、共用與不共用主成分向量之平均轉換誤差

誤差	不共用 PCA 向量			共用 PCA 向量		
配對	32 區間	64 區間	128 區間	32 區間	64 區間	128 區間
MA=> MB	0.5442	0.5438	0.5442	0.5389	0.5389	0.5389
MA=> FA	0.5159	0.5158	0.5156	0.5155	0.5154	0.5154
FA => MA	0.5387	0.5386	0.5384	0.5369	0.5344	0.5344
FA => FB	0.5807	0.5806	0.5805	0.5773	0.5768	0.5768
平均	**0.5449**	**0.5447**	**0.5447**	**0.5422**	**0.5414**	**0.5414**

(三)、PCA 轉換之必要性測試

對於圖二的流程裡，加入"PCA 係數轉換"與"PCA 反轉換"方塊是否為必要的？在此我們以量測語音轉換的平均轉換誤差的方式，來比較 PCA 係數轉換加入與不加入的優劣，所用的測試語料和誤差的量測方式，和 4.2 節裡敘述的一樣，亦即使用平行語料最後 25 句來作外部測試，並且量測轉換得到的 DCC 向量與對應的目標音框 DCC 向量之間的幾何距離，再計算全部音框的平均誤差。此外，直方圖等化也分成三種區間數來作實驗，即 32、64、與 128 個區間。經過實驗量測後，我們得到如表三所示的平均轉換誤差值，其中右邊三欄的數值是取自表二的右邊三欄。

表三、作與不作 PCA 係數轉換之平均轉換誤差

誤差	不作 PCA 係數轉換			作 PCA 係數轉換		
配對	32 區間	64 區間	128 區間	32 區間	64 區間	128 區間
MA=> MB	0.5454	0.5450	0.5446	0.5389	0.5389	0.5389
MA=> FA	0.5177	0.5172	0.5171	0.5155	0.5154	0.5154
FA => MA	0.5410	0.5402	0.5399	0.5369	0.5344	0.5344
FA => FB	0.5826	0.5825	0.5823	0.5773	0.5768	0.5768
平均	**0.5467**	**0.5462**	**0.5460**	**0.5422**	**0.5414**	**0.5414**

從表三的數值可以看出，作 PCA 係數轉換的確可使得語音轉換的誤差平均值下降，在 64 區間直方圖等化的情況下，平均轉換誤差可從 0.5462 降到 0.5414，這說明了直方圖等化之前先作 PCA 係數轉換是有用的、需要的。

(四)、目標音框挑選之轉換誤差

目標音框挑選可用以避免發生頻譜過度平滑的問題，其詳細的作法已在第三節說明。在此我們依據圖三之處理流程，測試目標音框挑選是否可以讓語音轉換的平均誤差減少？是否可以比圖二處理流程的好？圖三流程的語音轉換方法，我們稱為基本型目標音框挑選法，此外，我們也測試了另外一種語音轉換方法，稱為複合型目標音框挑選法，就是在圖二流程中"PCA 反轉換"與"HNM 語音再合成"兩方塊之間插入"目標音框挑選"之方塊，至於直方圖等化(CDF 轉換與反轉換)所用的區間數，這裡就設為 64。

對於前述的基本型與複合型目標音框挑選法，我們使用的測試語料和誤差的量測方式，和 4.2 節裡敘述的一樣，亦即使用平行語料最後 25 句來作外部測試，並且量測轉換得到的 DCC 向量與對應的目標音框 DCC 向量之間的幾何距離，再計算全部音框的平均誤差。經過實驗量測後，我們得到如表四所示的平均轉換誤差值，由表四可知基本

型目標音框挑選的轉換誤差平均值會變大成為 0.6029，這明顯比表三的 0.5414 增加了許多；再者，複合型目標音框挑選的轉換誤差平均值也變得更大，0.6121。根據這二個變大很多的誤差平均值，直覺上會讓人認為基本型與複合型目標音框挑選法，所轉換出的語音應會在音色相似度和語音品質上衰減很多，然而實際上當我們去聽轉換出的語音時，發現經由基本型或複合型目標音框挑選所轉換出的語音，語音品質卻是會變得更為清晰(應是使用真實音框 DCC 的緣故)，並且音色相似度也沒有衰減。所以，基於量測兩 DCC 向量之間幾何距離的轉換誤差平均值，其數值大小和語音品質之間似乎不是正比例的關係。

表四、目標音框挑選之平均轉換誤差

誤差 配對	基本型	複合型
MA=> MB	0.5990	0.6087
MA=> FA	0.5706	0.5791
FA => MA	0.5925	0.6032
FA => FB	0.6493	0.6574
平均	**0.6029**	**0.6121**

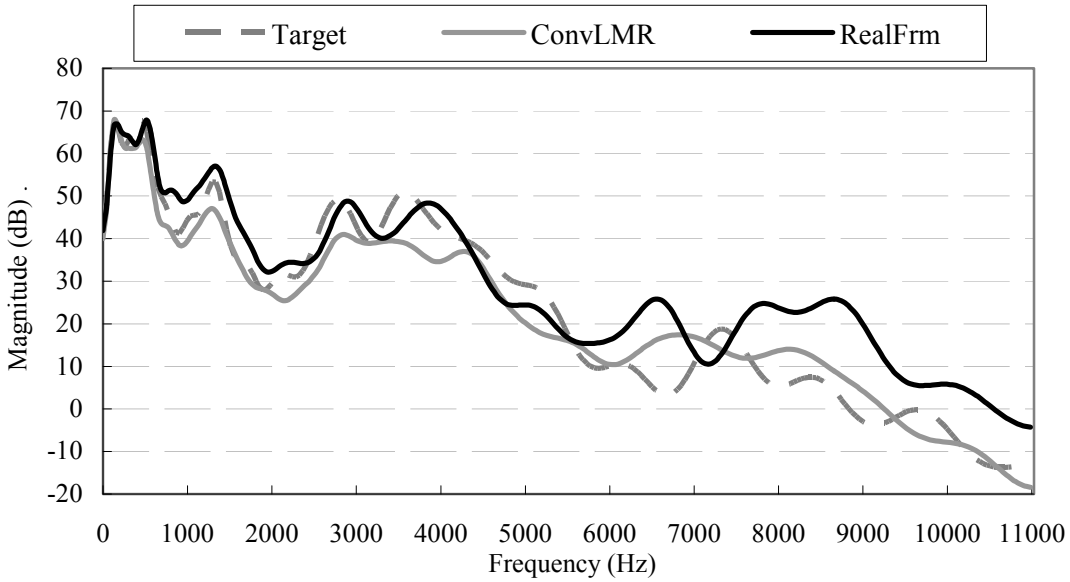
前述的**不一致性**情況，即誤差距離變大反而得到更好的語音品質，是什麼原因造成的？為了瞭解其原因，我們就找一些目標音框來觀察它們的頻譜包絡曲線。對於各個目標音框，我們把 LMR 對映出的 DCC 向量、經目標音框挑選得到的 DCC 向量、及該目標音框的 DCC 向量，計算出三者的頻譜包絡曲線並且畫出來作比較，結果我們發現了一個現象可用以解釋前述的不一致性。一個例子如圖四所示，圖四中的虛線代表/song/音節的一個目標音框的頻譜包絡線，淺灰色實線代表 LMR 對映得到的 DCC 向量所算出的頻譜包絡線，深黑色實線則代表**目標音框挑選**得到的 DCC 向量所算出的頻譜包絡線，比較這三條包絡線，我們可發現在橫軸頻率範圍 2,500 Hz 至 4,500 Hz 之間，深黑色實線的形狀比起淺灰色實線的形狀較為接近虛線曲線的共振峰起伏，所以這可以解釋為什麼**目標音框挑選**能夠改進轉換出語音的品質；此外，在橫軸頻率範圍 5,500 Hz 至 11,000 Hz 之間，淺灰色實線會比深黑色實線更為靠近虛線曲線，所以這可以解釋為什麼 LMR 對映所導入的轉換誤差，會比**目標音框挑選**所導入的轉換誤差來得小。

轉換後音框與目標音框的頻譜向量之間，誤差距離平均值的大小並不能夠代表語音品質的好壞，這樣的情形在前人的研究中已經注意到了，所以 Godoy 等人[15]採用以變異數比值(variance ratio, VR)來量測轉換後語音的品質，變異數比值的量測公式為:

$$VR = \frac{1}{C}\sum_{i=1}^{C} \frac{1}{L}\cdot\sum_{k=1}^{L}\frac{\hat{\sigma}_i^k}{\sigma_i^k} \quad , \tag{13}$$

其中 C 表示音段的類別數，L 表示頻譜向量的維度，$\hat{\sigma}_i^k$ 表示轉換後音框中第 i 類音段第 k 維頻譜係數的變異數，σ_i^k 則表示目標音框第 i 類第 k 維頻譜係數的變異數。

對於前面提到的四種處理流程，即作與不作直方圖等化(含 PCA)、作與不作目標音框挑選之四種組合，我們依據公式(13)去量測轉換後音框與目標音框之間的變異數比值，結果得到如表五所示 VR 值。由表五的 VR 值可發現，若不作目標音框挑選，則平均 VR 值只有 0.2 左右，但是當加入目標音框挑選之後，就可讓平均 VR 值提升到 0.5



圖四、音節/song/一個音框的三條頻譜包絡曲線

以上，所以客觀上來看，目標音框挑選之處理應可以讓語音品質獲得明顯的提升。至於直方圖等化，做了此種處理反而讓 VR 值下降一些，而 VR 值下降一些是否在主觀聽測上就會感覺到語音品質的衰退? 這尚需進行聽測實驗來驗證。

表五、變異數比值之比較

配對	無 目標音框挑選		有 目標音框挑選	
	DCC+LMR	HEQ+LMR	DCC+LMR	HEQ+LMR
MA=> MB	0.2463	0.1671	0.5893	0.5245
MA=> FA	0.1994	0.1290	0.5182	0.4485
FA => MA	0.2367	0.1775	0.5814	0.5383
FA => FB	0.2063	0.1375	0.5648	0.5303
平均	**0.2222**	**0.1528**	**0.5634**	**0.5104**

(五)、語音品質主觀聽測

我們使用未參加模型訓練的來源語句，來準備 4 組作語音品質聽測的音檔，這 4 組音檔的代號是 VD、VH、WD、WH，並且每一組中含有兩個音檔，分別是使用 MA=>MB 與 MA=>FA 之語者配對來作語音轉換而產生出的音檔，在此以_1 與_2 之代號來作區分。代號 VD 與 VH 中的 V 表示未作目標音框挑選，而 WD 與 WH 中的 W 則表示有作目標音框挑選；此外，VD 與 WD 中的 D 表示直接拿 DCC 向量去作 LMR 對映，就如圖一之處理流程，而 VH 與 WH 中的 H 表示 DCC 向量要先作 PCA 係數轉換及 CDF 係數轉換，然後才作 LMR 對映，就如圖二之處理流程。這 4 組音檔可從如下網頁去下載試聽: http://guhy.csie.ntust.edu.tw/vcHeqLmr/。

使用這 4 組音檔，我們先編排成二項的聽測實驗，第一項聽測實驗裡，受測者先、

後點播(VD_1, VH_1)與(VD_2, VH_2)兩對音檔來試聽，然後受測者分別給每一對音檔一個評分，以顯示左邊音檔的語音品質比起右邊音檔的品質是好或壞；第二項聽測實驗裡，受測者先後點播(WD_1, WH_1)與(WD_2, WH_2)兩對音檔來試聽，然後受測者分別給每一對音檔一個評分，以顯示左邊音檔的語音品質比起右邊音檔的品質是好或壞。在二項聽測實驗裡，受測者都是同樣的 12 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)分表示右(左)邊音檔的語音品質比左(右)邊音檔的明顯地好，1 (-1)分表示右(左)邊音檔的語音品質比左(右)邊音檔的稍為好一點，0 分表示分辨不出左、右兩音檔的語音品質。在二項聽測實驗之後，我們將受測者所給的評分作整理，結果得到如表六所示的平均評分。從表六的二項平均評分(即 0.583 與 0.375)可得知，評分分數都是正值，表示先作直方圖等化再作 LMR 對映，比起 DCC 向量直接作 LMR 對映會得到更好一些的語音品質；此外，第二項聽測的平均評分(0.375)，比起第一項聽測的平均評分(0.583)要稍微低一點，表示在作過目標音框挑選的處理之後，直方圖等化所帶來的語音品質改進，就會變得較不明顯。

表六、語音品質聽測--比較 DCC 與 HEQ

	DCC vs. HEQ (無 目標音框挑選)	DCC vs. HEQ (有 目標音框挑選)
平均評分 AVG (STD)	0.583 (0.776)	0.375 (0.824)

接著，我們再將前述的 4 組音檔作編排以進行另二項聽測實驗，在第三項聽測實驗裡，受測者先、後點播(VD_1, WD_1)與(VD_2, WD_2)兩對音檔來試聽，然後受測者分別給每一對音檔一個評分，以顯示左邊音檔的語音品質比起右邊音檔的品質是好或壞；在第四項聽測實驗裡，受測者先後點播(VH_1, WH_1)與(VH_2, WH_2)兩對音檔來試聽，然後受測者分別給每一對音檔一個評分，以顯示左邊音檔的語音品質比起右邊音檔的品質是好或壞。在第三、第四項聽測實驗裡，受測者也共有 12 位學生，他們大部分不熟悉語音轉換之研究領域，至於評分的標準與分數範圍則和前一段所說的一樣。在這二項聽測實驗之後，我們將受測者所給的評分作整理，結果得到如表七所示的平均評分。從表七的二項平均評分 0.917 與 1.125 可得知，只要加入目標音框挑選的處理，就可讓轉換出語音的品質獲得明顯的提升，並且這樣的提升要比表六裡的更明顯很多，所以這二項聽測實驗的結果，和表五裡量測出的 VR 值是相互呼應的。

表七、語音品質聽測--比較有、無目標音框挑選之差異

TFS (Target Frame Selection)	TFS_no vs. TFS_yes (DCC+LMR)	TFS_no vs. TFS_yes (HEQ + LMR)
平均評分 AVG (STD)	0.917 (0.584)	1.125 (0.680)

五、結論

我們研究改進了線性多變量迴歸(LMR)頻譜對映為基礎的語音轉換方法，在處理流程中加入**直方圖等化**及**目標音框挑選**之處理步驟，用以提升轉換出語音的品質。當我們在圖一流程的 DCC 估計與 LMR 對映之間插入“直方圖等化”處理(包含 PCA 係數轉換與 CDF 係數轉換)之後，雖然語音轉換的平均誤差距離會由 0.5382 [4]變大成為 0.5414，但是主

觀聽測實驗的結果顯示，轉換出語音的品質卻是比未加直方圖等化時的好，所以直方圖等化處理可用以紓解 LMR 對映所造成的頻譜過度平滑之問題。此外，關於來源語者和目標語者是否應共用主成分向量的疑問，實驗的結果顯示，讓兩語者共用主成分向量是比較好的作法，可讓語音轉換的平均誤差從 0.5447 減小成 0.5414。

另一種改進語音品質的方法是，在圖一流程的 LMR 對映與 HNM 語音再合成之間插入“目標音框挑選”之處理，雖然語音轉換的平均誤差距離會由 0.5382 變大成為 0.6029，但是客觀 VR 值的量測及主觀聽測實驗的結果都顯示，轉換出語音的品質確實是明顯地提升了，不論 LMR 頻譜對映方塊之前有否作過直方圖等化的處理，所以“目標音框挑選”比起“直方圖等化”，對於轉換出語音之品質提升更為有功效，並且 VR 值大體上可反應出語音的品質。另外，對於平均誤差距離愈大反而得到愈好的語音品質，這種不一致性的情況，我們觀察一些音框的頻譜包絡曲線後發現，轉換出之語音聽起來比較模糊者，通常其頻譜包絡在 2,500 Hz 至 4,500 Hz 之頻率範圍，會顯現過度平滑的情形，並且比起清晰者較為遠離目標頻譜包絡曲線；然而在 5,000 Hz 之後的頻率範圍，雖然模糊者的頻譜包絡也是顯現過度平滑的情形，但是比起清晰者卻較為接近目標頻譜包絡曲線，所以會計算出比較小的誤差距離。

致謝

感謝國科會計畫之經費支援，國科會計畫編號 101-2221-E-011-144。

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

[3] 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.

[4] 古鴻炎、張家維、王讚緯，「以線性多變量迴歸來對映分段後音框之語音轉換方法」，*第 24 屆自然語言與語音處理研討會*，中壢，Session 1 (speech processing)，2012。

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] A. Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Bentez and A. J. Rubio, "Histogram Equalization of Speech Representation for Robust Speech Recognition," *IEEE trans. Speech and Audio Processing*, Vol. 13, No. 3, pp. 355-366, 2005.

[11] S. H. Lin, Y. M. Yeh, and B. Chen, "A Comparative Study of Histogram Equalization (HEQ) for Robust Speech Recognition," *Computational Linguistics and Chinese Language Processing*, Vol. 12, No. 2, pp. 217-238, 2007.

[12] I. T. Jolliffe, *Principal Component Analysis*, second edition, New York: Springer-Verlag, 2002.

[13] T. Toda, 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, pp. 2222-2235, 2007.

[14] D. Erro, A. Moreno, and A. Bonafonte, "Voice Conversion Based on Weighted Frequency Warping," *IEEE trans. Audio, Speech, and Language Processing*, Vol. 18, pp. 922-931, 2010.

[15] E. Godoy, O. Rosec, and T. Chonavel, "Voice Conversion Using Dynamic Frequency Warping with Amplitude Scaling, for Parallel or Nonparallel Corpora," *IEEE trans. Audio, Speech, and Language Processing*, Vol. 20, pp. 1313-1323, 2012.

[16] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, and Y. Stylianou, "Towards a Voice Conversion System Based on Frame Selection," *Int. Conf. Acoustics, Speech, and signal Processing*, Honolulu, Hawaii, pp. 513-516, 2007.

[17] H. Hotelling, “Analysis of a Complex of Statistical Variables into Principal Components,” *Journal of Educational Psychology*, Vol. 24, No. 6, pp. 417-441, 1933.

[18] 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.