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 Method

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

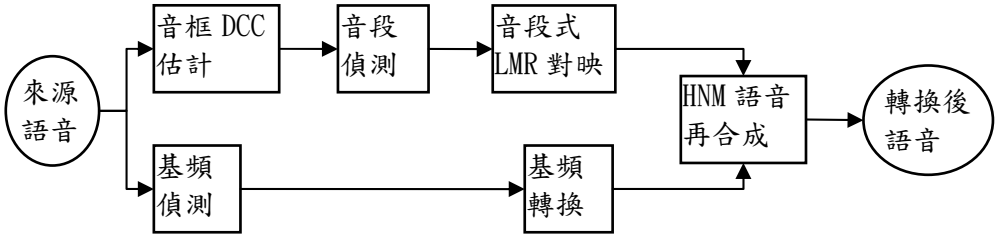
基於線性多變量迴歸(linear multivariate regression, LMR)頻譜對映之語音轉換方法，轉換出的頻譜包絡(spectral envelope)仍然存在過度平滑(over smoothing)的現象，因此本論文研究在音段式 LMR 頻譜對映之前加入直方圖等化(histogram equalization, HEQ)的處理，並且在 LMR 頻譜對映之後加入目標音框挑選的處理，希望藉以提升轉換出語音的品質。原先我們提出的基於 LMR 頻譜對映之語音轉換系統，其主要的處理流程如圖一所示，而在本論文裡所嘗試的改進則如圖二與圖三所示。

直方圖等化近年來被應用於語音辨識領域，用以減緩環境噪音造成的訓練語音和測試語音之間的頻譜不匹配問題，因此在觀念上應可用直方圖等化的處理，來把來源語音的頻譜轉變成目標語音的頻譜。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(discrete cepstral coefficient, 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。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。

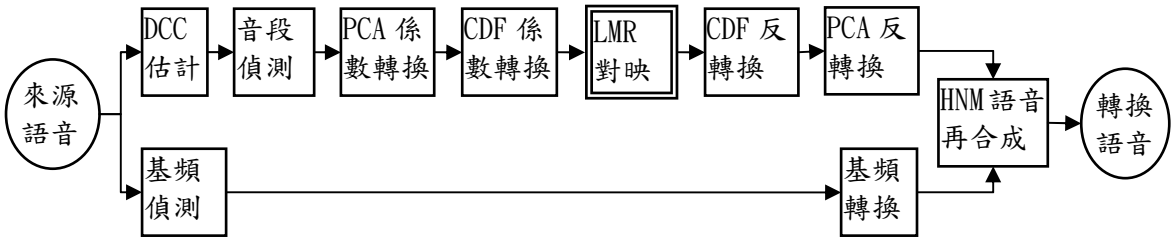
關於圖一流程遇到的頻譜包絡過於平滑的情況，雖然前人曾經提出至少兩種的改進方法，如全域變異數(global variance, GV)法、和頻率軸校正(frequency warping)法，但是他們的方法一者是針對 GMM (Gaussian mixture model)對映所設計的，或者不是針對 LMR 對映所設計的，因此我們在參考 Dutoit 等人的論文之後想到的一個作法是，在圖一“LMR 對映”方塊之後插入一個改進過的“目標音框挑選”方塊。目標音框挑選(target frame selection)的作法是，依據一個輸入音框的音段類別編號、及 LMR 對映出的 DCC 向量，到目標語者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。

在圖一、二、三裡都出現的 DCC 估計之方塊，表示我們採用離散倒頻譜係數作為頻譜特徵參數，並且階數設為 40 階，即一個音框要拿 $c_1, c_2, \ldots, c_{40}$ 等係數去作頻譜轉換的處理(c_0 則未轉換)。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 去計算出頻譜包絡，然後再依據頻譜包絡、轉換出的基頻值，去設定該音框的 HNM (harmonic-plus-noise model)模型之諧波參數和雜音參數，之後就可拿這些參數去合成出

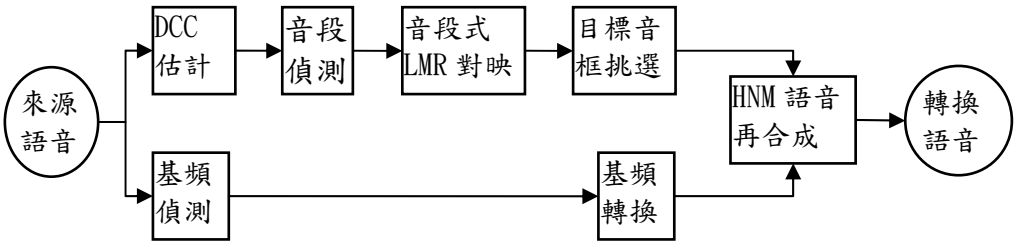
語音信號。關於 DCC 之估計、HNM 語音信號之合成，可參考前人的論文。



圖一、基於 LMR 頻譜對映之語音轉換的主要流程



圖二、基於直方圖等化及 LMR 頻譜對映之語音轉換流程



圖三、基於 LMR 對映及目標音框挑選之語音轉換流程

對於直方圖等化與目標音框挑選，我們使用 350 句平行語料來訓練模型參數，然後以外部(未參加模型訓練) 25 句平行語料來量測語音轉換之平均 DCC 誤差(以幾何距離作量測)。當加入直方圖等化後，平均 DCC 誤差會從 0.5382 變大成為 0.5414，而當加入目標音框挑選後，平均 DCC 誤差值則變大得更多，從 0.5382 變大成為 0.6029。但是，主觀聽測的實驗結果卻是相反的方向，也就是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。主觀聽測實驗裡，我們對“直方圖等化”與“目標音框挑選”之四種組合準備了轉換出語音的音檔，這些音檔可從如下網頁去下載試聽: http://guhy.csie.ntust.edu.tw/vcHeqLmr/。

前述之客觀誤差距離量測、和主觀聽測結果的不一致情形，多篇前人的語音轉換論文已經提到此種情形。因此，我們再以前人提出的 VR (variance ratio)量測法去進行客觀量測，結果量測到的 VR 值分別是，使用圖一流程時得到 0.2222，使用圖二流程時得到 0.1528，而使用圖三流程時則得到 0.5634。所以，量測出的 VR 值大小，大體上和聽測實驗的結果呈現一致的走勢，並且“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，這也可從 VR 值獲得呼應。此外，對於平均 DCC 誤差和聽測實驗的不一致情形，我們也設法從頻譜包絡上去尋找了它的原因，結果我們找到了一個可作說明的理由。

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

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