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

A Voice Conversion System based on Formant and LSF Mapping without Using Parallel Corpus

吳嘉彧　Chia-Yu Wu
國立清華大學電機工程學系
Department of Electrical Engineering
National Tsing Hua University
u921802@gmail.com

王小川　Hsiao-Chuan Wang
國立清華大學電機工程學系
Department of Electrical Engineering
National Tsing Hua University
hcwang@ee.nthu.edu.tw

摘要

語者特質轉換的研究已有廣泛的運用，早期使用的向量量化碼本對照，與目前被廣爲使用的高斯混合模型，都會使用經動態時軸校準的平行對應語句作訓練。近年來已有減少使用訓練語料與使用非平行句的語料進行語者特質轉換的方法。本論文提出一個不採用平行句的訓練方法，而依據語者音節共振峰映對，並結合線頻譜頻率映對，進行語者特質轉換。

Abstract

Voice conversion has been used in many applications. The methods based on vector quantization codebook and Gaussian mixture models need dynamic time warping on parallel sentence corpus for generating mapping functions. Recent study tries to use less training data, and even without parallel sentence corpus. This paper presents a voice conversion method without using parallel sentence corpus. It applies the formant mapping and line spectral frequency mapping to accomplish a voice conversion system.

關鍵詞：語者特質轉換，平行句語料，共振峰映對，線頻譜頻率映對

Keywords: voice conversion, parallel sentence corpus, formant mapping, LSF mapping

一、緒論

語音轉換和語者特質轉換已被探討多年，目前的研究除了提升轉換相似度以及保持語音品質，也要考慮實用層面會遇到的問題。例如爲了使用者的便利，訓練語料要減少，並要考慮跨語言語者特質轉換等沒有平行對應語句供訓練的情況，因此針對不同用途所使用的轉換方法和訓練語料都要有所調整。

語者特質轉換必須轉換來源語料的頻譜參數與韻律參數，使頻譜與韻律變成有目標語者的特性。最早被提出的頻譜參數轉換採用向量量化碼本對照(Vector Quantization Codebook Mapping)[1]方法，要面對不連續轉換造成音質不佳的問題。其後出現了其他使用類神經網路(Artificial Neural Networks, ANN)[2]和隱藏式馬可夫模型(Hidden Markov Model, HMM) [3]的方法，最被廣泛使用的則是高斯混合模型(Gaussian Mixture Model, GMM)[4]方法，但要面對頻譜過度平滑化的問題。之後有結合高斯混合模型(GMM)和動態頻軸校準(Dynamic Frequency Warping, DFW)[5]的方法被提出，能減低頻譜過度平滑化的情形。以上方法用於訓練的語料，需要是經動態時軸校準(Dynamic Time Warping, DTW)的平行對應語句。

近年來有依據語者共振峰特性所做的頻譜頻軸映對轉換(Formant Mapping)[6]方法，和從線頻譜頻率特性直接對高階數線頻譜頻率做轉換(LSF Mapping)[7]的方法，使用的語料數目減少，但能維持轉換品質。也有結合高斯混合模型於共振峰特性頻譜頻軸映對轉換[8]和多組合高斯混合模型線頻譜頻率轉換[9]等研究，使用的語料數目可以減到更少。

雖然使用經動態時軸校準的平行對應句能增進轉換正確性，但是在實際運用上難以取得平行對應句的語音資料，若是語料沒能準確對應，反而會造成誤差。通常語料內容要平衡音節出現機率，所以設計出的內容不一定符合語者說話習慣，或有些不自然的地方，若是要錄製的語料較多，在錄製者和語者較爲疲憊的情況下，容易有些語料錄製問題發生。

由於針對語言翻譯[10]的跨語言語者特質轉換研究[6][11]興起，如2002年歐洲提出的TC-STAR計劃[10]，辨識完英語後將其翻譯成西班牙文或華文，再用文字轉語音系統結合語者特質轉換系統，將其輸出爲近似原語者發出的西班牙語或華語。因此語者特質轉換要考慮在沒有平行對應語句的情況下，也能調整語者間轉換函式特徵參數，以減少非平行句轉換錯誤[12]。

韻律參數轉換常常使用基週同步疊加法(Pitch Synchronous Overlap and Add, PSOLA)[13]，可以彈性調整語句音調的高低起伏和說話速度。一般是分析語者正常說一句話時的韻律參數，如平均基頻、基頻標準差、音長、音量等等。再針對不同語者之間的韻律參數作轉換，以達到轉換語者特質的目的。

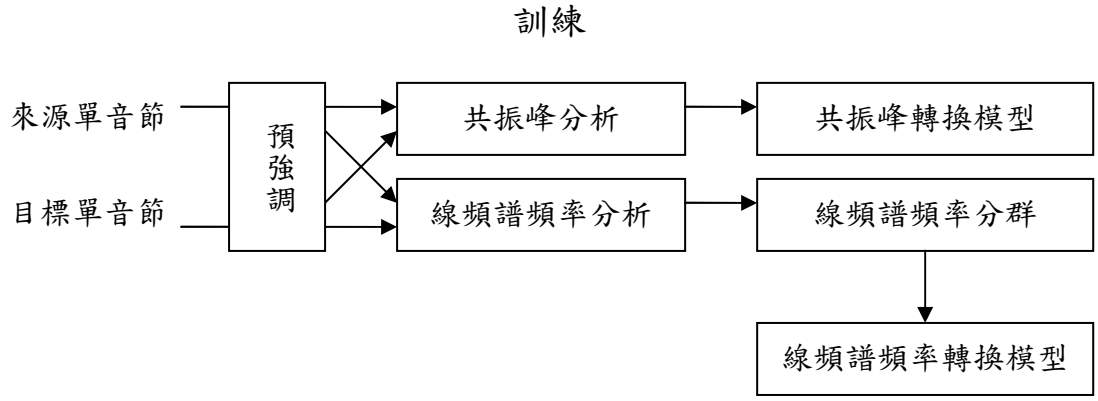
本研究的主要目標，是在不使用平行對應語句訓練的情況下做語者特質轉換，使轉換後的聲音相似於目標語者的聲音，並保持語音品質。在轉換頻譜參數上，本文僅根據國語單音節對共振峰特性做頻譜頻軸映對轉換，並結合線頻譜頻率特性做低階數線頻譜頻率轉換。由較不連續的分段共振峰轉換搭配分群加權平均後頻譜資訊較平滑的線頻譜頻率轉換，達到互補效果。由於單音節的基頻比較不穩定，因此在轉換韻律參數時還是使用短句(不平行)的基頻作爲基週同步疊加法(PSOLA)的基準，考慮單音節基頻變異性，僅將其作爲變異參數進行韻律轉換。

二、語者特質轉換的系統架構

一個語者特質轉換的系統，可以分成訓練和轉換兩個部份。

(一) 訓練部份

圖一展示頻譜參數的映對轉換模型訓練程序，在分別對來源和目標語者的七個主要韻母(ㄚㄛㄜㄝㄧㄨㄩ)單音節進行共振峰分析之後，建立共振峰轉換模型。再對來源和目標語者的全單音節線頻譜頻率做分群，建立線頻譜頻率的映對轉換模型。



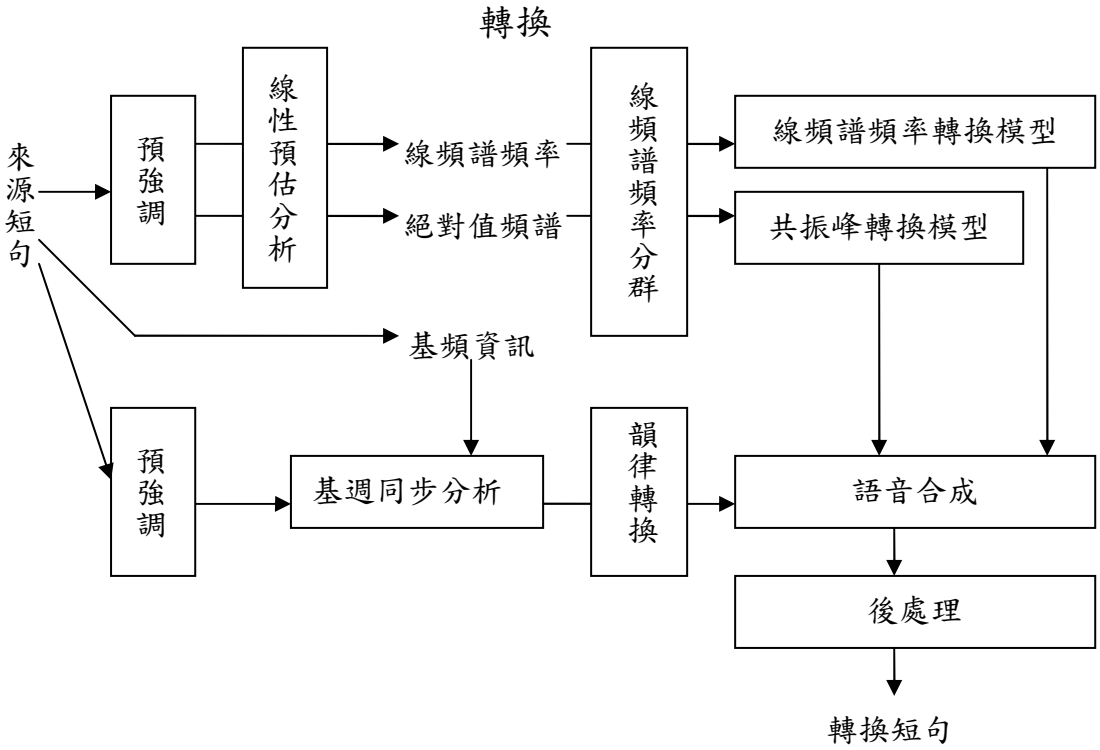
圖一、 訓練部份架構圖

(二) 轉換部份

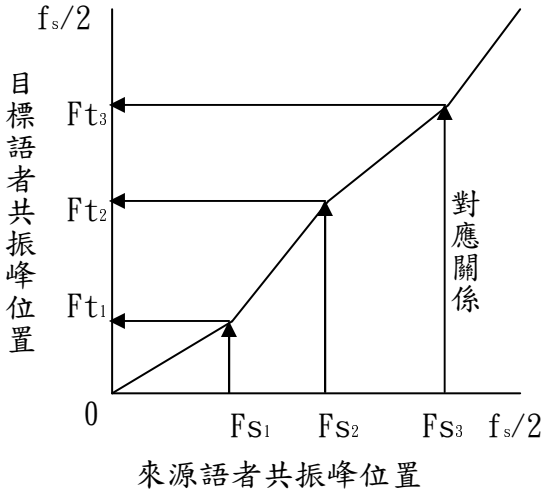
圖二說明轉換方式，對來源語句進行特徵參數抽取後，得到基頻參數，和線性預估絕對值頻譜與線頻譜頻率。基頻參數套進基週同步分析，抓取其基週標記，並取出欲轉換單元(兩倍標記週期)。線性預估絕對值頻譜套進共振峰轉換模型，線頻譜頻率則套進線頻譜頻率轉換模型的參數，搭配韻律轉換的資料進行語音合成，經過後處理就得到轉換後的結果。

三、以共振峰特性做頻譜頻軸映對轉換

由低頻到高頻的前三個共振峰 F_1、F_2 與 F_3，常被用於描述語音中母音的差異。線性預估階數越高，能找到的共振峰位置越多，但是若過多也會干擾一對一對應的結果。本文抓取訓練用共振峰時所用的線性預估階數爲 16 和 20 (16 爲主，20 爲輔)。使用七個國語主要韻母(ㄚㄛㄜㄝㄧㄨㄩ)的前三個共振峰，做爲對各語者做頻譜頻軸映對轉換的基準。在 0Hz 到第一個共振峰、第一個共振峰到第二個共振峰、第二個共振峰到第三個共振峰、第三個共振峰到 8000Hz 的對應範圍內，來源語者以雙方共振峰頻率爲基準，頻譜形狀被進行線性壓縮或擴張。各區段內的轉換如圖三與(1)式所示。



圖二、 轉換部份架構圖



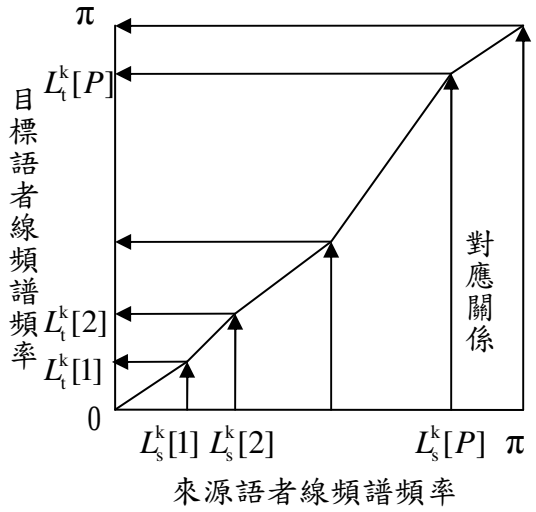
圖三、 共振峰轉換示意圖

$$F(f)=\begin{cases}\dfrac{f\times \mathrm{F_{t1}}}{\mathrm{F_{s1}}} & 0<f\le \mathrm{F_{s1}}\\ \mathrm{F_{t1}}+\dfrac{(f-\mathrm{F_{s1}})\times(\mathrm{F_{t2}}-\mathrm{F_{t1}})}{(\mathrm{F_{s2}}-\mathrm{F_{s1}})} & \mathrm{F_{s1}}<f\le \mathrm{F_{s2}}\\ \mathrm{F_{t2}}+\dfrac{(f-\mathrm{F_{s2}})\times(\mathrm{F_{t3}}-\mathrm{F_{t2}})}{(\mathrm{F_{s3}}-\mathrm{F_{s2}})} & \mathrm{F_{s2}}<f\le \mathrm{F_{s3}}\\ \mathrm{F_{t3}}+\dfrac{(f-\mathrm{F_{s3}})\times(\frac{\mathrm{f_s}}{2}-\mathrm{F_{t3}})}{(\frac{\mathrm{f_s}}{2}-\mathrm{F_{s3}})} & \mathrm{F_{s3}}<f\le \dfrac{\mathrm{f_s}}{2}\end{cases} \quad (1)$$

本研究在進行共振峰轉換後，搭配後續的線頻譜頻率轉換，藉著相對低階(和[7]相比)的線頻譜頻率線性對應，在線頻譜頻率轉換時所用的多個單音節分群加權平均，降低因爲不穩定單音節造成的影響，使頻譜表現較平滑，讓轉換能維持轉換相似度和語音品質的平衡。

四、以線頻譜頻率特性做線頻譜頻率轉換

線頻譜頻率爲線性預估分析中常用到的參數，可以穩定的表現出聲音的頻域特性。透過分析來源和目標語者的線頻譜頻率，將 404 個國語單音節語料分成 K 群，對這 K 群(K = 32)資料求出各階線頻譜頻率的對應轉換函式。



圖四、 線頻譜頻率轉換示意圖

圖四中的 $L_s^k[1]$ 、 $L_s^k[2]$ 到 $L_s^k[P]$ 是來源語者的 P 階(P= 16)線頻譜頻率，直接對應到目標語者的 P 階(P= 16)線頻譜頻率， $L_t^k[1]$ 、 $L_t^k[2]$ 、・・・、 $L_t^k[P]$ 。0 和 π 是邊界條件，

$$L_s^k[0]=L_t^k[0]=0 \quad (2)$$

$$L_s^k[P+1]=L_t^k[P+1]=\pi \quad (3)$$

(4)式和(5)式描述前後階的線頻譜頻率應符合同一線性對應函式 $f_j^k(\bullet)$。

$$f_j^k(L_s^k[j-1]) = L_t^k[j-1] \qquad (4)$$

$$f_j^k(L_s^k[j]) = L_t^k[j] \qquad (5)$$

L_s^k 和 L_t^k 分別代表來源語者和目標語者的第 k 群線頻譜頻率值。來源語料線頻譜頻率值 $x[i]$ 若是在第 j-1 到 j 階範圍，就要套進轉換函式 $f_j^k(\bullet)$，$\tilde{y}[i]$ 即爲轉換後的線頻譜頻率。

$$x[i] \in \left\langle L_s^k[j-1], L_s^k[j] \right\rangle \qquad (6)$$

$$\tilde{y}[i] = f_j^k(x[i]) = a_j^k x[i] + b_j^k \qquad (7)$$

若使用的線性預估階數爲 P，要進行轉換必須先求出這 K 群轉換參數 a_j^k 與 b_j^k，j= 1, 2, …, P+1，k = 1, 2, …, K。每群資料可以使用 P+1 個最小平方差解進行，先將(2)式到(7)式的內容表示成 P+1 個式子，

$$\begin{cases} a_j^k \times L_s^k[j-1] + b_j^k = L_t^k[j-1] \\ a_j^k \times L_s^k[j] + b_j^k = L_t^k[j] \end{cases} \qquad j = 1, 2, \cdots, P+1 \qquad (8)$$

簡化成矩陣函式，

$$S_j^k A_j^k = T_j^k \qquad j = 1, 2, \cdots, P+1 \qquad (9)$$

其中 $S_j^k = \begin{bmatrix} L_s^k[j-1] & 1 \\ L_s^k[j] & 1 \end{bmatrix}$ $A_j^k = \begin{bmatrix} a_j^k \\ b_j^k \end{bmatrix}$ $T_j^k = \begin{bmatrix} L_t^k[j-1] \\ L_t^k[j] \end{bmatrix}$

利用公式 $\hat{A} = (S^T S)^{-1} S^T T$，即可求出第 k 群所需的參數 a_j^k 與 b_j^k，j= 1, 2, …, P+1。

來源語料每一個音框的線頻譜頻率進行轉換前，必須先將其經過一個分群加權，得到對這 K 群的個別加權值，$w_k(x)$，再乘上該群的轉換，這個新的 $\tilde{y}[i]$ 才是此部分轉換的輸出。

$$\tilde{y}[i] = \sum_{k=1}^{K} w_k(x) f_j^k(x[i]) \qquad (10)$$

$w_k(x)$ 是使用馬氏距離(Mahalanobis distance)計算得到。

$$w_k(x) = \frac{1/d_k(x)}{\sum_{n=1}^{K} 1/d_n(x)} \qquad (11)$$

(11)式中的 $d_k(x)$ 即爲馬氏距離，

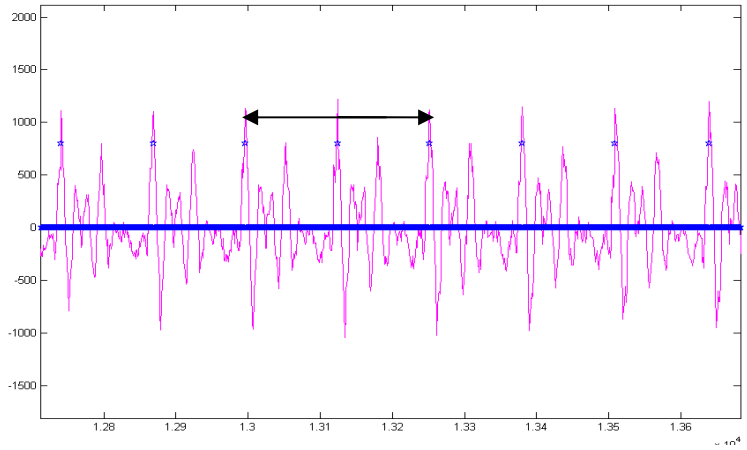
$$d_k(x) = [(x - L_s^k)^T (\sum\nolimits_s^k)^{-1} (x - L_s^k)]^{\gamma} \qquad (12)$$

(12)式中的 $\sum_s^k$ 爲 $L_s^k[1]$ 到 $L_s^k[P]$ 的對角共變異數矩陣(diagonal covariance matrix)，γ 爲一可調整參數，其值越大，對距離越近的分群比重越大。本研究使用的 γ 爲 4，和[7]相同。所使用的線性預估階數爲 16，相較於使用高階數的做法[7]，在頻譜上的表現較爲平滑，用以補償經過分段共振峰轉換造成的不連續。

五、基週同步分析及韻律參數轉換

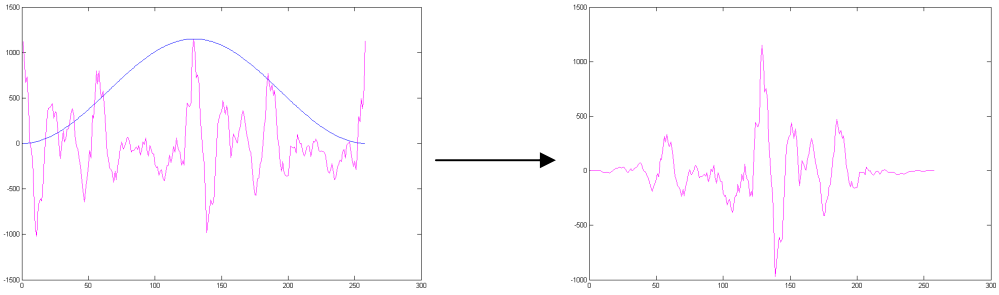
(一) 基週同步分析

爲了後續語音合成使用的基週同步疊加法(Pitch Synchronous Overlap and Add, PSOLA)，必須先對語料進行基週同步分析。以 ACF(Autocorrelation function)除以 AMDF (Average Magnitude Difference Function)的值估算基頻，完成基頻估算後，在每個有聲音段找到能量最大音框中振幅最大的取樣點，作爲第一個標記。從此標記依序往前與往後搜尋對應音框基本週期範圍，出現該區域最大值的位置，即爲下一個基週標記，重覆此動作可得到有聲音段的所有基週標記。



圖五、 基週標記

圖五爲基週標記示意圖，星號位置就是基週標記點，以基週標記前後兩個基週範圍(箭號所示)作爲運算單位，在此單位中的語音加上漢寧窗後，即爲韻律轉換所需的基本單元，如圖六。



圖六、 基本單元

(二) 發音腔道模型

利用線性預估分析，發音腔道模型可以表示成一個全極點系統(all-pole system)，

$$X(z) = \frac{\Theta_0}{1 - \sum_{j=1}^{P} a_j z^{-j}} E(z) \qquad (13)$$

$X(z)$是語音訊號，Θ_0是增益常數，$E(z)$是激發訊號。分母的$1 - \sum_{j=1}^{P} a_j z^{-j}$是一個逆向濾波器(inverse filter)，a_j (j= 1, 2, ⋯, P)是線性預估係數(linear prediction coefficients)，P爲線性預估階數。

(三) 韻律參數轉換

將語音訊號經過其對應的逆向濾波器，就得到剩餘訊號。如果對剩餘訊號的基頻軌跡作轉換，得到新的基頻軌跡，這個轉換後的剩餘訊號，用以產生轉換後的語音訊號。對平均基頻做調整，可以轉換語者說話的音高。對基頻標準差做調整，可以轉換語者說話的起伏。本文用於實驗的平均基頻和基頻標準差，是在語料庫中任意選取三個短句算出的平均值。由於考慮不同音節的基頻變異性，在基頻標準差比值上多乘上一個變異參數C_{kst}。

$$f_{0c} = C_{kst} \times \frac{\sigma_{ft}(f_{0s} - \mu_{fs})}{\sigma_{fs}} + \mu_{ft} \qquad (14)$$

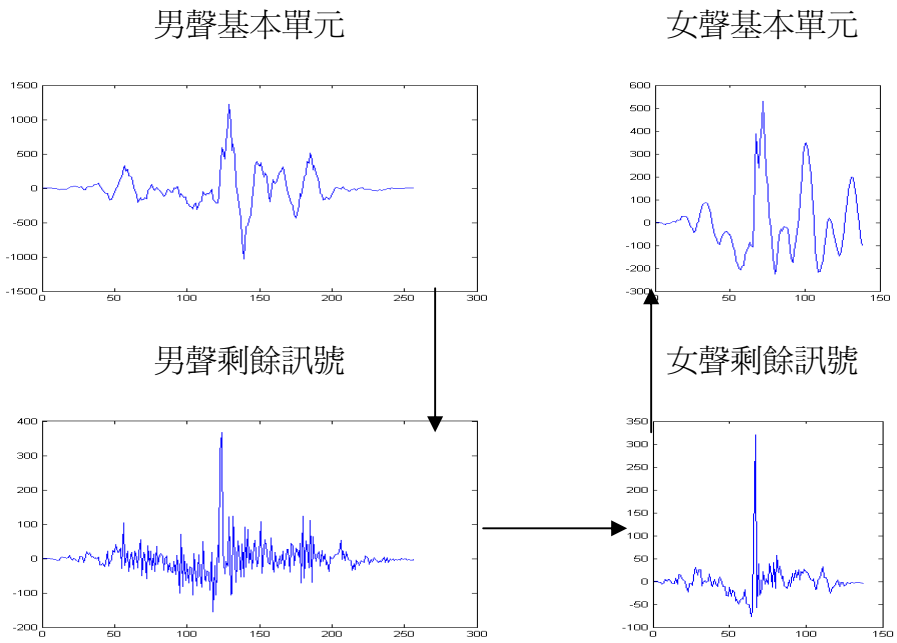
C_{kst}是以線頻譜頻率爲分群基準，計算「來源—目標」基頻比值F_{kst}得出：

$$C_{kst} = F_{kst} \Big/ \frac{1}{K}\left(\sum_{n=1}^{K} F_{nst}\right) \qquad (15)$$

六、語音合成

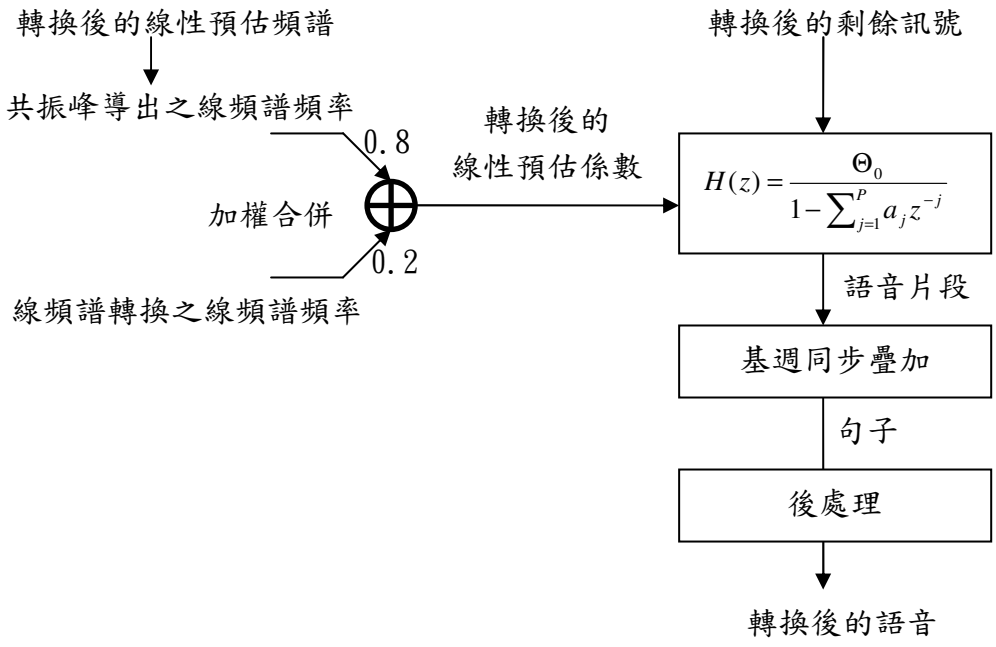
以 16ms 爲固定音框長度，8ms 爲音框移動距離，使用線性預估分析計算出該音框的線性預估頻譜和線頻譜頻率後，所得之絕對值頻譜形狀經由第三節共振峰頻譜分析轉換後，再將其表示成線頻譜頻率，稱之爲「共振峰導出之線頻譜頻率」。利用第四節線頻率頻譜轉換得到的線頻譜頻率則叫做「線頻譜轉換之線頻譜頻率」。將「共振峰導出之線頻譜頻率」和「線頻譜轉換之線頻譜頻率」兩者進行加權合併，產生新的線頻譜頻率，再轉爲線性預估係數，用以構成語音合成時的發音腔道全極點模型。每個不同長度的基本單元依其時間點，可以對應到進行頻譜資訊轉換時使用的固定音框序列。在時間上相對應的固定音框線性預估係數結合經過基週同步分析基本單元的剩餘訊號，可合成出一段語音訊號。

圖七爲基本單元的轉換示意圖(男聲→女聲)，將這些轉換完的語音片段經過基週同步疊加後，再經過後處理，即爲轉換後的語音。



圖七、 基本單元轉換

整個語音合成的流程見圖八。



圖八、 語音合成流程圖

七、實驗

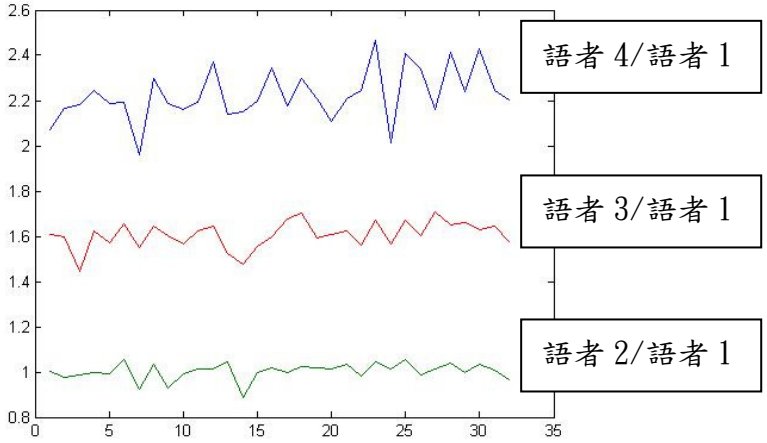
(一)實驗語料

所用語料經由麥克風錄製，取樣頻率爲 16kHz, 16 bits PCM。共有四位語者(兩男兩女)，每位皆有 404 個國語單音節及 110 句國語短句。本論文的轉換實驗訓練部份皆使用單音節進行，短句僅作爲韻律轉換的參考基準。使用語者資料如下：

表一、 語者資料-基頻

	性別	單音節平均基頻 (Hz)	短句平均基頻(Hz)
語者 1	男	126.07	125.22
語者 2	男	128.89	124.44
語者 3	女	206.79	188.10
語者 4	女	288.57	229.79

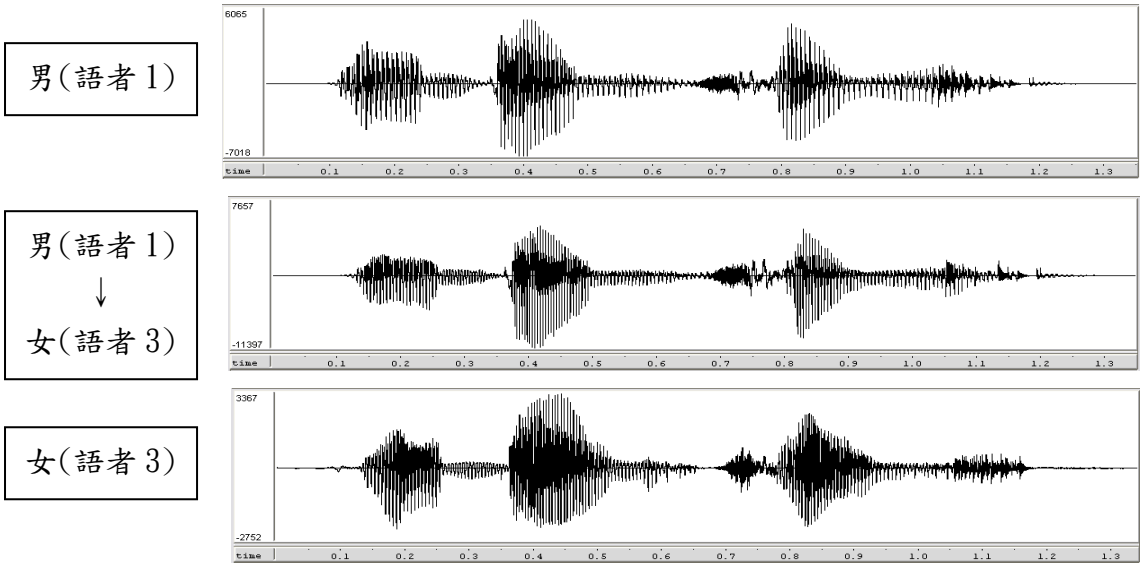
由於語者 4 發單音節的音高皆比平常說話時還高出許多，所以無法以平均基頻變異性做分段韻律轉換基準。但若以語者 1 的線頻譜頻率爲基準概分 32 群後，語者 2、3、4 的對應分群基頻和語者 1 的分群基頻比值，可以看出不同音節的基頻比值會有不同趨勢，考慮音節基頻變異性，可用於調整基頻標準差於改變音高起伏程度。



圖九、 單音節分群後的基頻比值趨勢圖

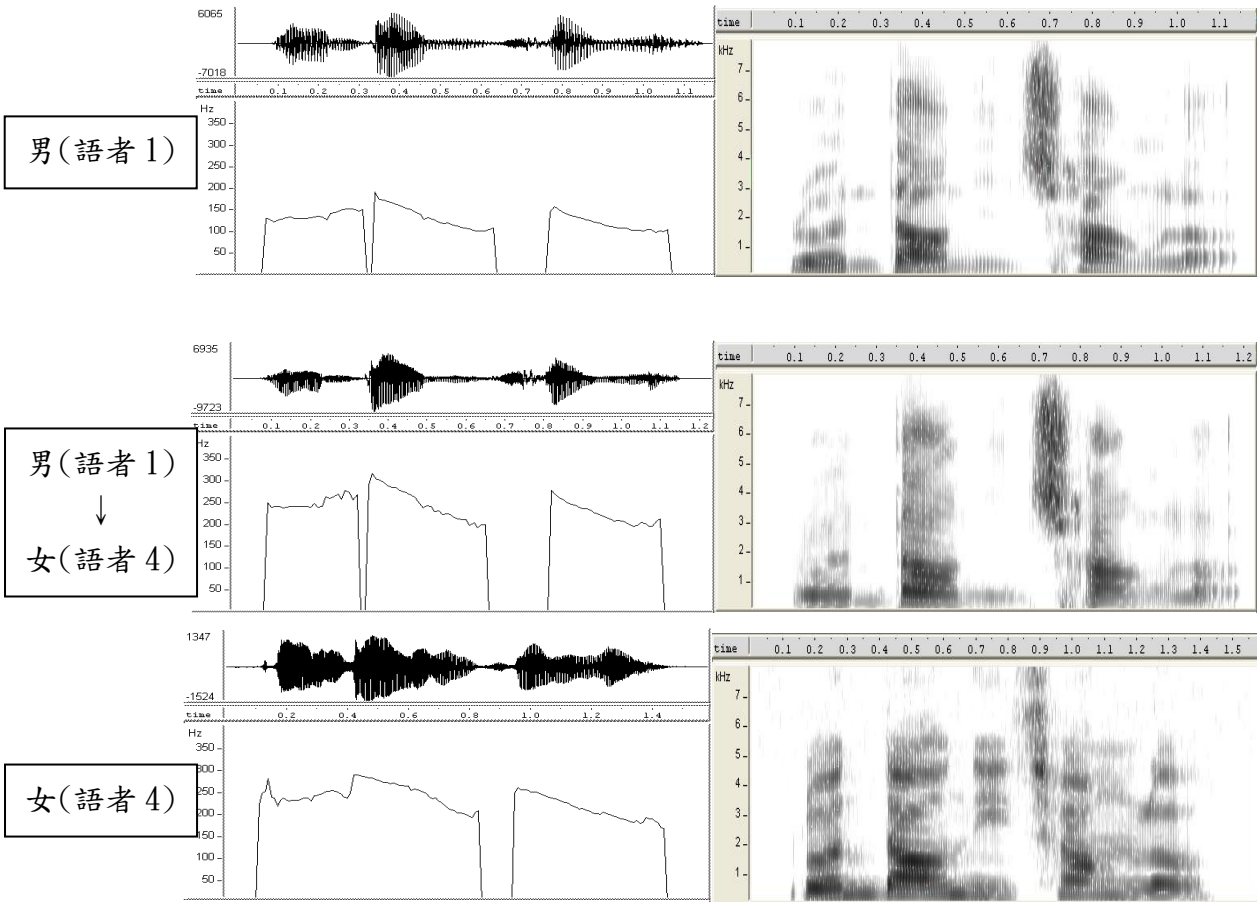
(二)轉換結果

圖十是語者 1(男)轉換到語者 3(女)的波形對照，轉換後的語音波形與來源語音波形比較，已有所改變。其中語者 3 的語料音量原本就較小，所以轉換的語音振幅強度對照語者 1 的強度做了調整。基本上轉換後的語音節奏受到來源語音的影響，但是音質接近目標語者的語音。



圖十、 語者特質轉換波形對照圖(輪到你唱了)

圖十一是語者 1(男)轉語者 4(女)的波形、基頻、聲譜對照圖，從圖中可看到將原本男生(語者 1)較低的基頻軌跡轉爲女生(語者 4)較高的基頻軌跡，而基頻軌跡的形狀由來源語音決定。



圖十一、 語者特質轉換波形、基頻、聲譜對照圖(輪到你唱了)

(三)主觀聽覺實驗

主觀聽覺實驗包含音質和相似度兩部份，兩個實驗都有 9 位未經過特別聽力訓練的受測者參與，其中包含了一位不熟悉國語，平常僅使用越南語和英語的受測者。受測者在實驗中各聽 35 個不重複的國語短句，其中混合了對不同目標進行的轉換結果及未轉換的語者原始語料。受測者不知道轉換內容及其是否爲轉換過的語料，藉此可得到語料經語者特質轉換前後的比較。評定方法如下：

(1) 音質

使用平均主觀分數(Mean Opinion Score, MOS)，依照語音品質由高到低分爲五級。

(2) 相似度

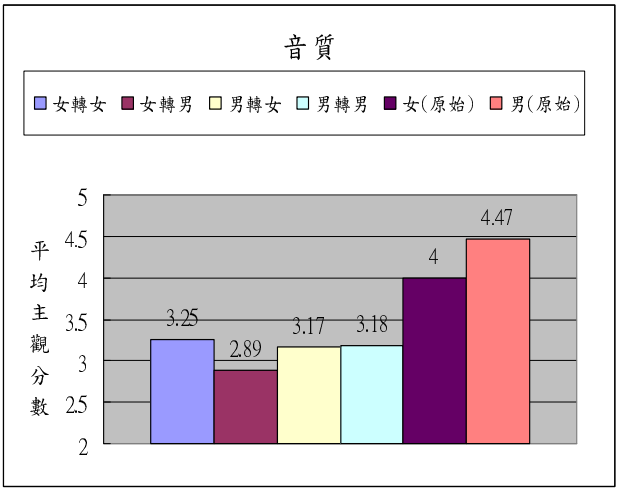
先讓受測者聽兩位語者(A 和 B，句子內容相同)的原始語料，再請受測者聽一個未知語料(X，句子內容不同於 A 和 B)，此未知語料可能是轉換後的語料(由 A 轉 B 或由 B 轉 A)，也可能是兩位語者的其他語料(A 或 B 的其他句子)，相似度部分有五個選項，不依選項大小判斷相似與否，而是要再從選項內容判斷。

表二、 相似度選項依據 (A 爲正確答案)

5	X 聽起來肯定是 A
4	X 聽起來接近 A，但不確定是否真的是 A
3	無法分辨 X 比較像 A 還是 B
2	X 聽起來接近 B，但不確定是否真的是 B
1	X 聽起來肯定是 B

(3) 實驗結果

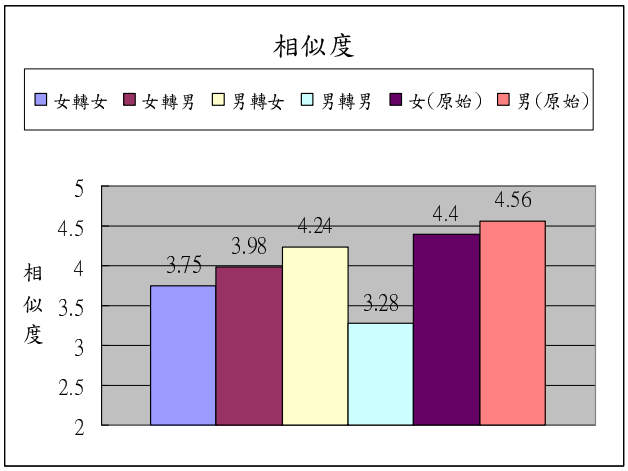
由於受測者都未接受過特別的聽力訓練，所以經過轉換的語料和語者原始的語料都進行音質評定，作爲轉換前後音質落差的參考。



圖十二、 各項轉換的音質比較

圖十二明顯看出經過轉換處理的語音，其音質變差，整體來說，平均主觀分數從 4.2 下變成 3.1，下降約 1.1。

由於本研究的語者特質轉換沒有使用平行對應句，所以相似度測驗也包含了各語者未經轉換的不同句子，作爲轉換前後相似度落差的參考。



圖十三、 各項轉換的相似度比較

圖十三展示經過語音轉換之後，與目標語者語音的接近程度，其中男聲轉男聲時轉換語音與目標語音的相似度最低，因爲這兩位男性語者的聲音比較相似。男聲轉女聲時，其效果最好。受測者並不知道是原始語音或是轉換語音，所以照樣評分，這兩個原始語音的相似度分數只作爲參考。

八、 結論

考量語料取得容易度對語者特質轉換系統實用上的影響，本研究不使用平行對應句，僅使用對應單音節建立轉換模型。對應單音節容易取得，但由於資料量小，對其穩定度的要求就要相對提高，才能使轉換相似度和聲音品質和使用平行對應句一樣好。期望將來的研究能再減低對特定語料的依賴性，並且維持良好的轉換相似度和轉換後的聲音品質。

參考文獻

[1] Abe, M., Nakamura, S., Shikano, K., Kuwabara, H., “Voice Conversion through Vector Quantization,” Proc. IEEE ICASSP pp.665-658, 1988.

[2] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, “Transformation of formants for voice conversion using artificial neural networks,” Speech Communication, vol. 16, pp. 207-216, 1995.

[3] E. K. Kim, S. Lee, and Y. H. Oh, “Hidden Markov model based voice conversion using dynamic characteristics of speaker,” Proc. EUROSPEECH, vol. 5, Rhodes, Greece, 1997.

[4] Stylianou, Y., Cappe, O., and Moulines E., “Continuous Probabilistic Transform for Voice Conversion,” IEEE Trans.on Speech and Audio Processing, vol.6, no.2, pp.131-142, 1998.

[5] Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, “Voice Conversion Algorithm based on Gaussian Mixture Model with Dynamic Frequency Warping of STRAIGHT Spectrum,” Proc. IEEE ICASSP, pp. 841-844, 2001.

[6] Zhiwei Shuang, Raimo Bakis, and Yong Qin, “Voice Conversion Based On Mapping Formants,” TC-STAR Workshop on Speech-to-Speech Translation, pp.219-223, 2006.

[7] Zdenek Hanzlicek, Jindrich Matousek, “On Using Warping Function for LSFs Transformation in a Voice Conversion System,” Proc. IEEE ICSP 2008.

[8] Kun Liu, Jianping Zhang, and Yonghong Yan, “High Quality Voice Conversion through Combining Modified GMM and Formant Mapping for Mandarin,” IEEE ICDT 2007.

[9] Elina Helander, Jani Nurminen, and Moncef Gabbouj, “LSF Mapping for Voice Conversion with very small training sets,” IEEE ICASSP 2008.

[10] H. Höge, “Project Proposal TC-STAR - Make Speech to Speech Translation Real,” Proc. LREC’, Las Palmas, Spain, 2002.

[11] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, and J. Hirschberg, “TC-Star: cross-language voice conversion revisited,” TC-Star Workshop on Speech-to-Speech Translation, 2006.

[12] Athanasios Mouchtaris, Jan Van der Spiegel, and Paul Mueller, “Nonparallel Training for Voice Conversion Based on a Parameter Adaptation Approach, ” IEEE Trans.on Speech and Audio Processing, vol.14, no.3, 2006.

[13] H. Valbret, E. Moulines, and J. P. Tubach, “Voice Transformation using PSOLA Technique, ” Proc. IEEE ICASSP. San Francisco, USA, pp. 145-148, 1992.

[14] Yinqiu Gao, Zhen Yang, “Pitch modification based on syllable units for voice morphing system,” IFIP International Conference on Network and Parallel Computing Workshops 2007.