半自動人聲配音系統
Sem -automatic Voice Dubbing Syste

古鴻炎　　洪茂松
Hung-Yan Gu　　Mou-Son Hong
國立台灣科技大學電機系，台北市基隆路四段 43 號
Department of Electrical Engineering
National Taiwan University of Science and Technolog
E-mail: root@guhy.ee.ntust.edu.t

摘要

人聲配音或音色轉換就是要將一種原始音色轉變成許多不同的音色。在戲劇作品(如電影)裡的人聲配音工作，通常需要許多的配音員，因此我們希望借由電腦來作音色轉換之處理，而達成只需要一位配音員便能夠配出許多不同音色的功能，以節省人力，實現個人配音工作室的理想。我們研究的人聲配音系統稱之為半自動的，因為喜怒哀樂的表達仍然要由聲音信號的原輸入者來控制。我們的成果是，提出了一種半自動人聲配音的方法，以獨立控制基頻、聲道長度、聲源訊號、及聲道內部比例之方式，來達到改變人聲音色的目標，其中聲道內部比例之控制因素是新提出來的。此外，我們也已將此方法實作成一個可線上即時操作之系統，並且經由實際聽測驗証，的確可得到相當豐富的音色轉換。

關鍵詞：人聲配音，音色轉換，基頻，聲道，語音合成

ABSTRACT

Voice dubbing or timbre translation is meant a processing that can translate a single voice timbre into many distinct timbres. In drama works, differen t actors are usually dubbed with different timbres qnd many dubbing persons are therefore required. To reduce the cost spended to dub the actors, it will be useful if the computer can help to convert a single voice timbre into many distinct timbres. Therefore, we intend to develop a semiautomatic voice-dubbing system. It is called semiautomatic becaus the emotions (like, anger, sad, happy) perceived from th translated voice need still be controlled by the person who provides the original voice. In this pa per, a method for timbre translation is proposed. The goal is accomplished by providing independent control of fundamental frequency vocal-track length, voice source, and internal ratio of vocal track. Among the four factors, the internal ratio of vocal track is newly studied here. In addition, an on -lin operable system is built with this method. It can be used for real-time voice dubbing. Also, according to our perception tests, it can indeed translate a single voice timbre into many distinct timbres.

Key Words: voice dubbing, timbre translation, pitch frequency, vocal track, speech synthesis

1．前言

相信大家都有這樣的經驗，當一個朋友打電話來時，我們接上電話就知道是誰打來的，那麼我們是如何做到聽音辨人的呢？一般而言是依據“音色”和“說話習慣”來作判斷的。說話習慣主要是由後天的學習所形成，而如何形成則是相當的複雜，並且，人是具有相當高的可塑性的，例如電視上的“模仿秀”，某甲模仿某乙說話、唱歌等，其中說話習慣的模仿可說是比音色的模仿更爲重要，所以我們認爲，說話習慣目前在研究上是比較困難的，而音色大致上可說是天生的，音色的不同主要是由於發聲器官構造的差異所造成，在研究上比較有跡可循。雖然一個人由幼年到老年的音色改變是相當明顯的，不過我們知道，造成這種現象的原因是，發聲器官會隨著年齡的增長而有所改變，此外男性與女性的音色不同，也同樣是因爲發聲器官的差異(主要是聲帶的震動頻率)所造成的。因此，發聲器官的各個組件對於音色有何種影響，是我們所感興趣的，我們希望藉由電腦來模擬及控制一些發聲器官的特性而達到改變音色的目的，然後把研究的成果實作成一個即時的半自動人聲配音系統。

以往有關音色轉換的研究，如語音轉換(voice conversion)，幾乎是集中在特定對象與另一特定對象之間的音色轉換關係，也就是將某甲的音色經過處理之後轉換成某乙的音色，這些方法都必須要先錄製甲、乙兩人所說的共同句子，再根據這些句子去求取甲乙兩人的音色對應關係，然後才能依對應關係將甲的語音轉換成乙的語音。這些轉換的方法大致可分成兩大類，第一類是參數式(parametric)的方法，這類方法的效果全視參數是否能夠精確的描述語者的特徵 [1,2,3]；另一類是非參數式(non-parametric)的方法，以各種訓練的程序來獲得兩人之間的音色對應關係，這種方法需要夠多的訓練語句，才能得到較佳的效果 [4,5]。另外，Baudoin 與 Styliano u曾對四種頻域上的方法去比較語音轉換的效果 [6]。

如何將一種音色轉換成爲許多不同的音色的研究，我們還沒有找到他人的直接相關之研究成果，不過，前述的語音轉換之研究是有一定的參考價值的。關於音色轉換的應用，在許多戲劇作品(如電影)裡的人聲配音工作，通常需要許多的配音員，因此，我們希望能運用電腦來作音色轉換之處理，以達成只需要一位配音員便能夠配出許多不同音色的功能，而使得戲劇作品及有聲書裡關於人聲配音的人力花費可以節省下來，並使得個人配音工作室的理想能夠實現。我們所研究的人聲配音系統稱之爲半自動的，因爲喜、怒、哀、樂的表達，仍然要由聲音信號的原輸入者來控制，我們的系統只負責音色的轉換。我們定義一個全自動的人聲配音系統是，不需要輸入語音信號，並且喜怒哀樂的情緒可由參數來控制，這樣的一個文句翻語音(text-to-speech)系統，很明顯的以目前的技術來說，要製作一個全自動的人聲配音系統是相當困難的。

在本文裡，我們提出了一種音色轉換的方法，以獨立控制基頻、聲道(vocal track)長度、聲源訊號、及聲道內部比例的方式來達到改變人聲音色的目標，並且已經將

這個方法實作成一個可線上即時操作的人聲配音系統，實際聽測的驗證是，的確可得到相當豐富的音色轉換。我們的系統的主要處理流程就如圖 1 所示，這裡先對各個處理方塊做簡略的介紹，詳細的作法則在以後各節中說明。

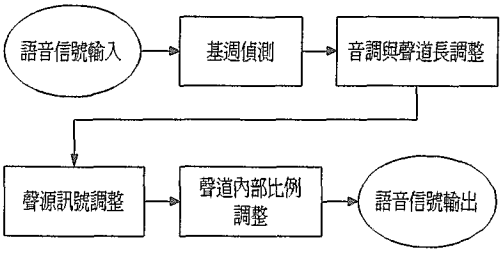


圖 1　半自動人聲配音系統之主要處理流程

基週偵測：

將輸入的語音信號以時框(frame)爲單位進行切割，然後即時求取時框中各基週頂點(pitch peak)的位置。

音調與聲道長調整：

此部份是修改我們先前提出的 TIPW (Time Proportioned Interpolation of Pitch Waveform) 音節信號合成方法 [7]，以便能夠在即時的要求下，將語音信號依據所設定的音高、聲道長參數來作調整。

聲源訊號調整：

此部份是透過 LPC(linear prediction coding) 分析來求取聲源訊號，然後依據前人的研究成果來調整聲源訊號 [8,9]。

聲道內部比例調整：

以 LPC 分析所建構的聲道模型爲基礎，改變聲道前後部分（咽腔、口腔）的長短比例，以模擬不同人的聲道內部比例之差異。

2. 即時基週位置偵測

基週(長度或位置)的求取在許多語音信號處理的應用裡都是一個重要的部分，而對於音色轉換處理來說，基週更是一項重要的參數，並且求出的基週位置的準確性對系統的效能有重大的影響。關於基週求取的研究 [10]，過去被提出的方法可粗略分爲兩大類，其中一類是依據頻譜或相關性(co-relation)分析來求取基週長度，例如使用 LPC 分析的剩餘訊號的自相關係數 [11]，或使用交互相關係數(cross co-relation) [12]，來求取基週長度；另外一類是直接在時域波形上求取基週，例如根據波峰來尋找基週位置的方法 [13,14]。這兩大類的方法，一般來說頻譜與相關性分析方法求出的基週長度之準確度較高，但是所花費的時間較長，而時域波形上的方法相對的準確度較低，但其所需的處理時間較少。

2.1 基週位置偵測的方法

雖然有些語音信號處理的應用裡，只要求取基週長度，不需要求取基週位置，也不需要即時處理，然而我們的系統除了要求，求取基週位置及求取的準確性之外，同時還必須兼顧即時處理的需求，因此採用了時域波形上直接偵測的方法，圖 2 是我們發展的基週偵測方法的流程圖，圖中各方塊的功能將在下面各段中加以說明。

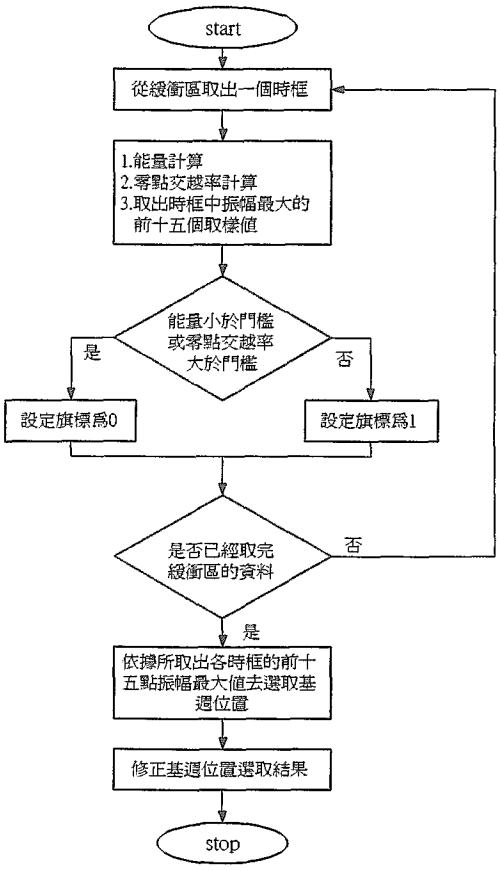


圖2　基週位置求取流程圖

2.1.1 緩衝區與時框長度

緩衝區(buffer)的長度目前設爲 500 * *sampling_rate* / 11,025 個取樣點，因此可以包含 22Hz 以上的聲音信號，而緩衝區的取法如圖 3 所示，圖 3(A)中的緩衝區裡的信號波形經過基週位置偵測之後可得圖 3(B)中所示的基週位置，而下一次緩衝區的設定也如圖 3(B)裡所示的。另外，一個緩衝區內的信號樣本要分成時框來處理，時框長度設爲緩衝區的一半，且有 50%的部分與下一個時框重疊，所以一個緩衝區可劃分爲三個時框，這樣劃分是爲了增加基週偵測的準確度。

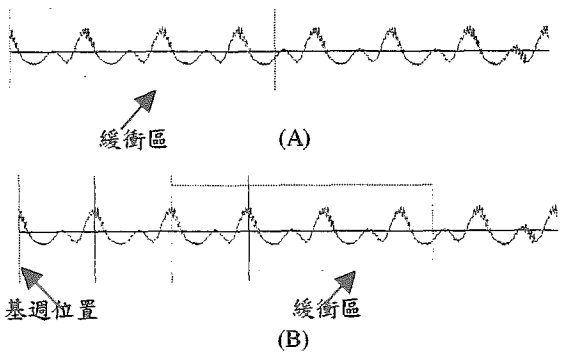


圖 3　緩衝區設定之例子

2.1.2 能量與零點交越率

這裡我們定義能量是一個時框內信號樣本的振幅平方

的加總，通常一個具有週期性的信號其能量都會比靜音或雜訊信號能量大許多，然而一個時框能量的門檻應該要設定爲多少才合理呢？依據我們的經驗，以一個時框具 250 個 16 bits 的樣本而言，其能量門檻可設爲 256,000，如此當一個時框的能量小於 256,000 時，就將該時框的週期性旗標設爲零，否則設爲 1。零點交越率是指單位時間內信號通過時間軸的次數，週期性的信號通常零點交越率會比雜訊信號爲低，在本論文中，我們以一個時框爲單位來計算零點交越率，以一個 22.67ms (取樣率 11025Hz 時，時框長爲 250 個樣本點) 的時框來說，我們設門檻爲 60，如此當一個時框的零點交越率大於 60，就認爲該時框中沒有週期性訊號，而將該時框的週期性旗標設爲零，否則設爲 1。

綜合能量與零點交越率的結果，只有在三個時框的週期性旗標都等於零時，我們才認爲這個緩衝區爲不具有週期性信號。

2.1.3 基週位置選取

以下說明基週位置選取的處理步驟。

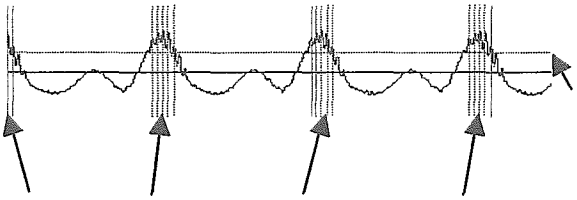
步驟(1): 首先判斷週期性旗標是否皆爲零，若皆爲零(即無週期性信號)，直接傳回零個週期。

步驟(2): 合併三個時框中各取出的 15 個最大振幅值，依時間次序加以排序，然後存入陣列 Y[1] ~ Y[45]。

步驟(3): 計算振幅門檻值並存於變數 Cli ，我們設定 Clip 的程序是，本次緩衝區的三個時框各取出一個振幅極大值，設爲 max1,max2,max3，接著判斷前一個緩衝區是否具有週期性訊號，如果沒有，就令 Clip = (max1+max2+max3)*0.2，如果有週期性訊號，就令 Clip = min(max1,max2,max3)*0.6。

步驟(4): 由於 Y[1]～Y[45]並不是每一個點都在峰值的位置，而是形成一群群集中在峰值及峰值附近的點上，因此我們依序取出 Y[1]～Y[45]中振幅大於 Clip 且爲鋒值者，將它們存入陣列 X[1]～X[K] 。圖 4 爲陣列元素 X[1]～X[K] 分佈的例子。

步驟(5): 上、下週期長度之門檻值的設定：



X[1]～X[2] X[3]～X[7] X[8]~X[12] X[13]~X[17]

圖 4 陣列元素 X[1]～X[K] 分佈的例子

若 前一個緩衝區具有週期性訊號

則 下週期長度門檻＝前一個緩衝區的週期平均值（ave_pitch）＊0.75

上週期長度門檻＝前一個緩衝區的週期平均值（ave_pitch）＊1.75 (1)

否則 下週期長度門檻

＝35 * sampling_rate / 11,025

上週期長度門檻

＝200 * sampling_rate / 11,025 (2)

也就是當第一個週期信號開始出現時，我們設定其基頻必須在 55Hz 至 315Hz 的範圍內，而當目前所分析的緩衝區不是連續週期信號的起始點時，則隨著到目前爲止的平均週期長度作調整。

步驟(6): 由陣列 X [1]～X [K]中找出本次緩衝區中所有的週期頂點位置。作法爲，以緩衝區起點爲參考點，將 X[1]～X[K]中距離在上、下週期長度門檻之間的 X[i] 找出，從中取出一個具有最大振幅的點當作是週期的邊界點；再以此邊界點爲參考點，往前找出 X [1]～X [K]中下一批介於上、下週期長度門檻之間的 X[i]，從中挑出振幅最大之 X[i]當作是下一個週期邊界點，如此繼續找下去。

2.1.4 修正基週位置選取結果

經過以上的基週位置選取之後，大約已可取得 80%的正確基週位置，不過這還未達到我們所要求的準確性標準，而且這樣的取法還很容易忽略一些振幅小於 Cli 值的正確基週位置，或者將一個較長的週期分成了兩個週期，導致後面改變音色處理時的品質降低，爲了改善這些缺點，因此我們發展了修正基週選取結果的方法，就是根據相鄰的兩個週期長度不會有很劇烈的改變的觀點，再配合前一小節所提的平均週期長度及相對的振幅改變量來修正所取得的基週位置。

根據多次的實驗觀察，我們定義兩個週期長度 T_1 和 T_2 是相近的，如果它們滿足如下之條件

$$ABS(T_1 - T_2) < MAX(T_1, T_2) \times \frac{11}{50} \quad (3)$$

依據此條件，就可偵測出圖 5 中的 T_2 與 T_3 不是相近的，圖 5 爲一個緩衝區經過基週位置選取後的例子，其

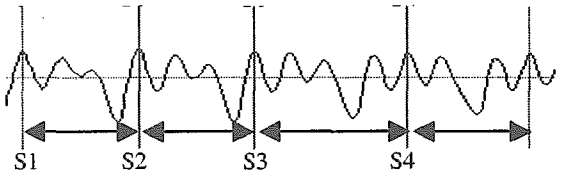


圖 5 有錯誤的基週選取結果

($T_1 = 61, T_2 = 60, T_3 = 79, T_4 = 62$)

中 P_i 表示 S_i 點上的振幅，接著，我們就進行如下所列的處理步驟，來將它更正回來。

步驟(a): 由週期序列中取出 T_1 及 T_2 兩個週期長度，由於 T_1、T_2 符合公式（3），且滿足振幅變化量的限制，即

$$MAX(P_1, P_2) / MIN(P_1, P_2) < 2 \quad (4)$$

因此我們將 T_1 及 T_2 保留不作修正，然後將平均週期長度改爲

$$ave\_pitch = (T_1 + T_2) / 2 \quad (5)$$

步驟(b): 比較 T_2 及 T_3，由於不符合公式(3)，所以便開始修正 T_3，即修正 S_4 的位置。如圖 6 所示，S_4 可能的位置範圍我們定義在 a ~ b 之間，而 a、b 點的位置則由公式(6)

$$a = S_3 + ave\_pitch * 90\ \%,$$

$$b = S_3 + ave\_pitch * 110\ \% \quad (6)$$

來設定，然後從〔a, b〕區間找尋一個具有最大振幅值的波峰，將所找到的最大波峰點設爲新的 S_4，並計算出新的 T_3 值，而平均週期長度則改爲

$$ave\_pitch = (ave\_pitch + T_3) / 2 \quad (7)$$

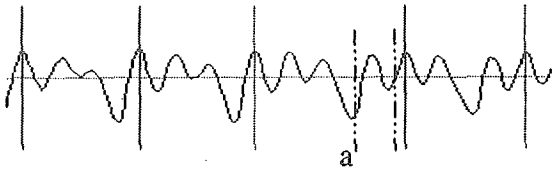


圖 6　基週修正範圍

步驟(c): 繼續檢查 T_i 與 T_{i+1}，採取步驟(a)或(b)的處理方式，直到整個緩衝區處理完畢。處理完成之後便得到如圖 7 所示的基週位置偵測結果。

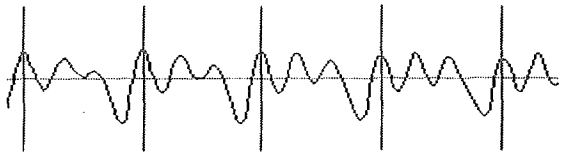


圖 7　基週修正後的結果

2.2 基週位置偵測之評估

加入基週位置修正之處理後，我們進行了評估實驗，其結果如表 1 所示，整體來說修正之處理可將基週位置選取正確率提高至 95 %以上。另外，由表 1 可知，對於正確率影響最

表 1　基週位置選取之正確率

	基週選取正確個數	基週選取錯誤個數	沒選到的基週個數	總基週個數	正確率
實驗一	122	0	6	128	95.3%
實驗二	283	3	11	297	95.3%
實驗三	1279	8	60	1347	95.0%

大的是沒有被選到的基週個數太多了，根據我們的觀察，這些沒被選到的基週全部集中在當聲音由靜音區轉換到有週期性時的前一至二個週期，以及從有週期性進入靜音區間之最後的一至二個週期，歸納其原因，是由於前述這兩個交換區間週期長度變化較大，且能量小於我們所設的能量門檻，因此被判斷爲非週期信號，而表 1 中的基週選取錯誤，主要是發生在爆破音上，如/ㄊ/及/ㄅ/。

3. 音調與聲道長調整

這裡的音調與聲道長調整指的是音高調整以及聲道長度調整等兩個部分，大體而言，影響音色最大的因素是音調(基頻)高低與聲道的共振頻率(formant frequency)高低，其中共振頻率高低與聲道長短有著密切的關係，當聲道變短時，共振頻率會提高，反之則會降低，即呈現反比的關係，因此，我們可以經由改變共振頻率的高低來達成控制聲道長短改變的目的。以往一種控制音高及共振頻率的簡單作法是，將錄音機的放音速度調快或調慢，這樣的做法的確可同時將音高與共振頻率在相反方向成比例的升高與降低，不過發音的時間長度卻也跟著會減少或延長，所以這並不是一個好的方法。因此，本文採用我們先前提出的 TIPW 音節信號合成法，TIPW 可以獨立的控制音長(duration)、音高以及共振頻率的高低，因此可以解決前述的問題，而達到分別控制音高與聲道長的要求。以下就對 TIPW 法的處理步驟作一簡單描述，詳細情形請參考原始文章[7]。

3.1 基頻調整

首先介紹基頻(音高)的調整，基頻的物理意義是指聲帶振動的頻率，通常小孩子的基頻比較高，其次是成年女性，而成年男性的基頻較低，我們藉由調整基頻便可以達到初步的音色轉換，以下是基頻調整的處理步驟：

(STEP 1)　依據基週調整比率及輸入信號裡的基週長度，求取目前欲合成之基週的長度。
(STEP 2)　依據合成週期的時間位置，找尋兩個對應的在輸入信號裡的原始基週，然後依線性時間比率計算兩原始基週波形的加權。
(STEP 3)　兩原始基週波形各乘上自己的加權。
(STEP 4)　依據合成週期的長度及原始週期的長度來共同決定餘弦窗(cosine window)的長度，然後將原始基週波形乘上兩個半邊的餘弦窗，對齊放在合成基週的左右兩邊界並疊加。
(STEP 5)　將兩個處理過的原始基週波形相加。

3.2 聲道長度調整

聲道長度的調整是透過調整聲道共振頻率之方式來達成，而共振頻率的調整是經由再取樣(resampling)來達成，例如當要把共振頻率全體調高爲原來的 1.25 倍時，就相當於在原始的信號波形上，每 1.25 個取樣點設定一個新的取樣點(但取樣率不變)，因此新的取樣點可能會落在兩個舊取樣點之間，這裡基於實作的考量，我們採取二項式內插的方法來求取新取樣點上的振幅值，二項式內插的公式爲

$$y = f(x) = A \cdot x^2 + B \cdot x + C \tag{8}$$

當在連續三個舊取樣點 x_0、x_1、x_2 上的樣本值爲 y_0、y_1、y_2 時，並且要內差的取樣點 x 是介於 x_1 與 x_2 之間，則我們可令 x_0、x_1、x_2 之值爲 0、1、2，如此帶入公式(8)可得 A、B、C 的解爲

$$\begin{cases} A = (y_2 - 2y_1 + y_0)/2 \\ B = (4y_1 - y_2 - 3y_0)/2 \\ C = y_0 \end{cases} \tag{9}$$

求得 A、B、C 之後，再將 x 以 x - x_0 之值帶入 $f(x)$ 便可以得到新的樣本值，如此繼續可得到新的週期波形，然後再將 resampling 後的波形取代 3.1 節中音高調整(STEP 3)裡所用的原始基週波形，就可以得到經過音高調整以及聲道長度調整的新合成的週期波形了。

4. 聲源訊號調整

聲源訊號指的是藉由聲帶震動所產生的氣流訊號，聲源訊號的頻率(即聲帶開關的頻率)，就是前面所說的基頻，關於基頻的調整，上一節已經討論過了，而本節所指的聲源訊號調整是指，調整聲源訊號的波形，如圖 8 所示，將一個聲源訊號的週期由聲帶張開至最大的時

刻 a，往兩旁分別進行拉長與縮短 T_1,T_2 的處理，使其仍維持一個週期的長度。根據前人的研究結果[9]，這樣的改變可使語音的音色發生變化，當 T_1 增加時可以使聲音具有 jitter 的效果，而當 T_2 增加時則具有 shimmer 的效果。

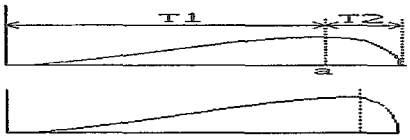


圖 8　T_1 增加時的聲源波形

我們考慮實作上可能遇到的問題，並未經由積分 LPC 分析之剩餘訊號來求得聲源訊號波形，而是採用近似的做法，如 4.1 節裡的說明。這樣做是因爲，實際上經由 LPC 分析求得的聲源信號，有許多時候，不像圖 8 所示那麼理想，只有一個波峰而已，那麼程式如何去分辨哪一個波峰才是所要的？

4.1　聲源訊號求取

在說明聲源信號最高點之時刻 a 如何決定之前，我們先說明剩餘信號是如何求出的，就是先取得一個週期的語音信號，對它進行LPC分析而建立聲道的全極(all-pole)模型，然後讓分析用的語音信號通過反全極模型，就可求得剩餘信號[11]。接著，由比較剩餘訊號與聲源訊號我們知道，當剩餘訊號振幅達到最大時通常就是圖 8 裡聲源訊號的 a 點位置，而再觀察剩餘訊號與原始訊號之間的對應關係，則可發現當原始語音信號到達基週端點(pitch peak)時，剩餘訊號振幅也會達到最大，因此我們推論，當原始語音信號位於基週端點時，通常就是聲源訊號的振幅最大值位置。

前面所指的一個週期的語音信號，是指圖 9 中由 b 至 c 之間的信號。b 與 c 的定義爲 b = (S1+S2) / 2，c = (S2+S3) / 2，並且，我們令圖 8 中的時間點 a 等於 S2，即原始語音信號中的基週邊界點。我們以 12 階的 LPC 來分析一個週期的語音信號，對於男性的聲音而言算是滿適合的，而對於女性的聲音，可能會因樣本點數過少而會有分析不準確的情況。

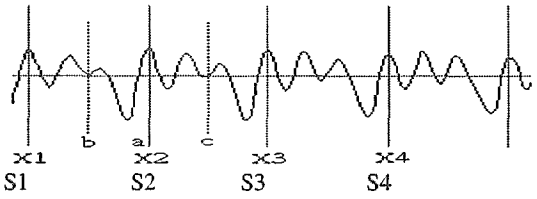


圖 9　求 LPC 分析剩餘信號之週期設定方式

4.2　聲源訊號調整方法

由 4.1 節的推論，我們依據圖 9 裡的 a、b、c 參數來訂定圖 8 裡的 T_1 與 T_2 的數值，即令

$$T_1 = a - b, \quad T_2 = c - a \tag{10}$$

由於我們令 a 點在週期邊界點 S_2 上，因此圖 8 中的 T_1 通常不等於 T_2，但是爲了維持與原週期長度相同，因此 T_1 與 T_2 的調整是有連帶關係的，我們的調整方式爲

若 $T_1 \geq T_2$，則令 $T_2' = T - T_1'$，

若 $T_1 < T_2$，則令 $T_1' = T - T_2'$ (11)

其中 $T = T_1 + T_2$，而 R 是調整比率，R 的數值範圍是 0 < R < 2，有了新的 T_1', T_2'之後，接著以如同 3.2 節中的 resampling 的作法來改變聲源訊號上升區間與下降區間的樣本點數，而整個聲源訊號調整的流程如圖 10 所示。在聽覺上，經過聲源訊號調整的語音信號，當 R 值比 1.0 大許多時(如 1.5)，其音色聽起來的感覺像是當

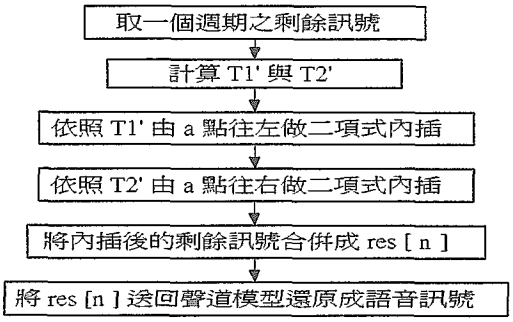


圖10　調整聲源訊號之流程圖

喉嚨乾燥(發炎)時所發出來的聲音，而當 R 比 1.0 小許多時(如 0.5)，聽起來的語音有種閃爍不清的感覺。

5. 聲道內部比例調整

人類的聲道是由聲帶以後的咽腔、口腔、鼻腔等所構成，由於每一個人的生理結構都不盡相同，因此發音出來的音色也就不一樣。這裡我們考量的是，即使聲道長度相同的兩個人，他們的音色聽起來仍然是有差別的，造成這種現象的原因，我們認爲聲道內部比例(即咽腔、口腔的長度比例)是其中一個具有重大影響力的因素，因此這一節就說明我們實現聲道內部比例調整的一種方法。

依據前人的研究，我們知道經由 LPC 分析所得到的反射係數，其代表的意義是聲道內部各相鄰區間(section)的截面積比 [15]，而由反射係數所建立的格狀(lattice)濾波器，如圖 11 所示，便可當作是聲道的一種模型。

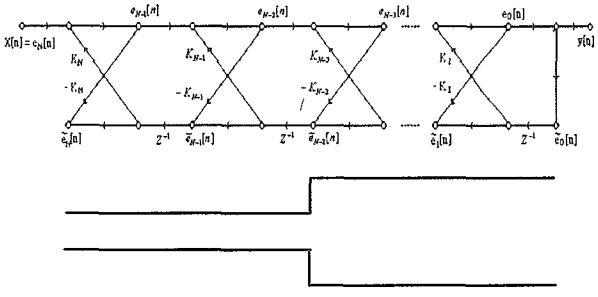


圖 11　格狀濾波器與聲道的對應圖

由於 LPC 分析得到的反射係數 K_1 至 K_n，分別表示聲道從聲帶到嘴唇之間的 n 個相鄰區間的截面積比，因此我們便想到以此作爲聲道內部比例調整的依據，詳細情形如下。

5.1 聲道內部比例調整方法

本文研究的聲道內部比例調整之方法是，先依據反射係數去求出聲道各區間的相對截面積大小，再以內插的方式來調整聲道內各區間的截面積大小，然後再將改變後的截面積反轉成新的一組反射係數，用以建立聲道內部比例調整後的新的聲道模型，詳細處理步驟為：

(STEP a) 對各個語音信號週期進行 LPC 分析而求出反射係數 $K_1 \sim K_L$ 與剩餘訊號，設 LPC 分析階數為 L；

(STEP b) 將 $K_1 \sim K_L$ 帶入公式（12）求出聲道截面積 $Area_1 \sim Area_L$；

$$Area_{i+1} = \left(\frac{1-K_i}{1+K_i}\right) \times Area_i, \text{ 令 } Area_0 = 100 \qquad (12)$$

(STEP c) 依照所需要的聲道內部調整比例 u，將

$$Area_1 \sim Area_{L/2} \text{ 內插成 } Area_1 \sim Area_{uL/2}$$

$$Area_{\frac{L}{2}+1} \sim Area_L \text{ 內插成 } Area_{\frac{uL}{2}+1} \sim Area_L$$

(STEP d) 將 $Area_1' \sim Area_L'$ 帶入公式(13)以求取調整後的反射係數 $K_1' \sim K_L'$

$$K_i = \frac{Area_i - Area_{i+1}}{Area_i + Area_{i+1}} \qquad (13)$$

(STEP e) 將圖 12 中的反射係數 K 以 K' 替代，再將(STEP a)所求得的剩餘訊號代入格狀濾波器，便可得到改變聲道內部比例之後的語音信號。

5.2 驗證與比較

5.2.1 截面積上的驗證

對於前面所提的聲道內部比例調整方法的一個驗證作法是，拿原始語音信號的 LPC 分析聲道截面積，與調整聲道截面積後的合成信號的 LPC 分析截面積作比較，藉此判斷所提出的方法是否能夠達成我們想要的效果。例如圖 12 裡，是對/a/

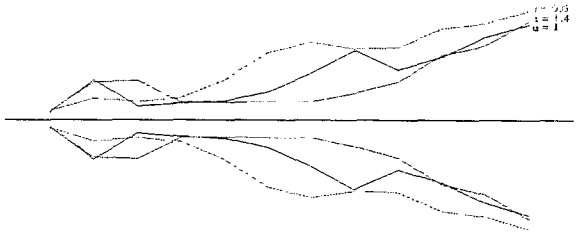
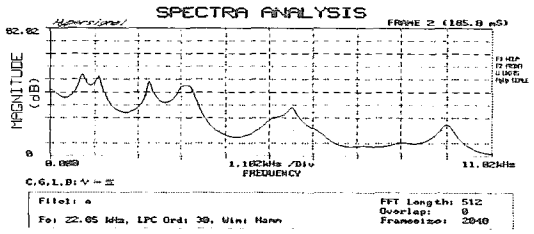


圖 12 聲道內部比例調整之/a/音截面積比較

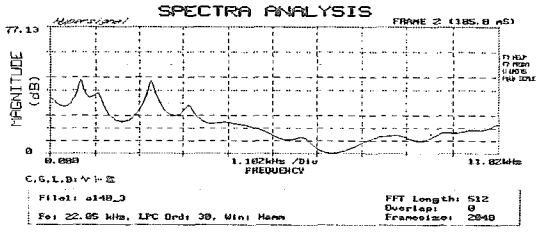
音信號在 u＝0.6、1.4 及 1 之三種調整比例時，對合成出的信號進行 LPC 分析所得到的聲道截面積之比較。由於這個方法所採用的是直接改變聲道內部前後兩段的長度比例，理論上應該只會影響聲道內部的長度比例，不過，在分析合成的語音信號後，我們發現除了聲道內部長度比例會被改變之外，截面積也會受到影響，推測其原因是，經由 LPC 分析所得到的剩餘訊號，並不是只有乾淨的聲源訊號而已，它還包含了聲道的訊息在內，這些包含聲道訊息的剩餘訊號經過調整過的聲道模型時，便會產生截面積上的改變。其他音素如/ i / 與/ u /的分析結果，也發現有類似的現象。

5.2.2 頻譜上的比較

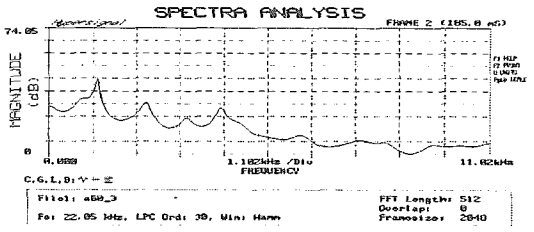
關於頻譜上的比較，在此僅列出/a/音在頻譜上的分析，其餘的/ i /以及/ u /的頻譜也都有相似的結果，圖 13 為經過 LPC 分析後得到的頻譜圖，圖 13(a)、(b)與(c)分



(a) 原始 /a/音的頻譜圖



(b) 設 u=1.4 時的頻譜圖



(c) 設 u=0.6 時的頻譜圖

圖 13 聲道內部比例調整之頻譜分析

別為調整比例 u 設為 1.0、1.4 與 0.6 時所得到的分析結果。觀察圖 13 可發現，當 u = 1.4 時，F1 的振幅顯現增加的趨勢，F2 的則減少，在頻率位置上，F1 與 F2 兩者都沒改變，我們可以把這種現象解釋為低頻部分的能量增加；另外當 u = 0.6 時，F1 的振幅顯現減少的趨勢，F2 的則增加，在頻率位置上，F1 有往 F2 移動的趨勢，而 F2 則沒改變，我們可以把這種現象解釋為高頻部分的能量增加。

5.2.3 音色上的比較

關於音色上的比較，我們還是以/ a /、/ i /、/ u /三個母音來實驗，經過實際試聽後，得到如表 2 裡記錄的結果，即設定調整比例 u=0.6 時，三個母音都一致地變得較為浮悶(音高感覺升高了)，而設定調整比例 u=1.4 時，三個母音也都一致地變得較為厚實(音高感覺降低了)，這樣的現象可由 5.2.2 節裡觀測到的頻譜變化來加以解釋；並且，這裡的聲道內部比例調整方法，對於各種音素的音色改變都能展現出一致性，而沒有南轅北轍的音色改變情形。

表 2 三個母音在 u 設為 1.4 與 0.6 時的比較

	u = 0.6	u = 1.4

/a/	比原音浮悶	比原音厚實
/i/	比原音浮悶	比原音厚實
/u/	比原音浮悶	比原音厚實

6. 系統評估

我們的半自動人聲配音系統是在個人電腦之Linux 作業系統上建造、發展的，由於舊式音效卡無法同時進行錄音與放音的動作，所以我們使用了兩張音效卡來同時進行錄音與放音的處理；在軟體的寫作上，採取了即時處理的作法，圖 1 裡的各個處理步驟，是以信號週期爲單位，循序、獨立地去處理；所用的個人電腦 CPU 是 AMD K6-2 / 333MHz 。

6.1 音色測試

我們採取聽測的方式來進行音色評估，事先請一位男性與一位女性各錄一句話並加以存檔，然後拿這兩句話給我們的系統作音色轉換處理，每句話經由不同參數設定之音色轉換處理，選取較具代表性的十句不同音色的合成語句給試聽者聽，這裡參與聽測的人數有 18 人，分別就十句合成語音的清晰度、自然度與辨別度作評估。在此所謂的清晰度是與原始錄進來的語句作比較，一樣清晰無雜訊則得滿分十分，比原始語句差則酌量扣分，而自然度是指，合成出來的語句是否有男生假裝女聲說話或是女聲假裝男生說話的情形，如果沒有則得滿分十分，否則依照假裝程度酌量扣分。

測試時我們系統的相關設定是，語音信號取樣頻率爲 44,100Hz，LPC 分析階數爲 24，測試語句爲"請把這籃兔子送走"，其他參數的設定如表 3 裡所列的，包括音高升降、聲道長度、聲源調整、聲道內部比例等參數的設定。對於播放的十個語句，試聽者依其感覺逐句給予清晰度與自然度兩項評分，結果得到如表 4 裡所列的分數資料。觀察表 4 男聲部分，可以發現第五句、第九句的清晰度與自然度都相對的較低，對照表 3 的設定，皆爲將音高降低，

表 3　音色轉換之 10 種參數設定

	音高%	聲道長度%	R%	U%		音高%	聲道長度%	R%	U%
第一句	125	100	100	140	第六句	200	60	120	100
第二句	140	70	100	100	第七句	80	110	100	170
第三句	400	50	100	100	第八句	140	70	100	80
第四句	60	60	100	100	第九句	50	130	80	80
第五句	50	150	100	100	第十句	70	90	100	30

聲道總長度拉長，也就是將聲音往男低音的方向改變，與第七句比較起來，算是調整滿大的，因此效果上較差，反過來，第六、八句則是自然度與清晰度最高的，對照表 3 的設定，爲將音高升高，聲道總長度減少，也就是往女高音的方向改變，由於原本是男性的聲音，因此這樣的改變效果還算不錯。再由表 4 的女聲部分來看，第

表 4　清晰度與自然度之聽測結果

女聲					
	清晰度	自然度		清晰度	自然度
第一句	9.1	7.71	第六句	7.28	6.87
第二句	9.02	7.61	*第七句*	*9.31*	*8.15*
第三句	**7.17**	**6.67**	第八句	9.24	7.33
第四句	**7.5**	**6.65**	*第九句*	*8.85*	*8.36*
第五句	*8.99*	*8.54*	第十句	7.5	7.29
平均				8.396	7.518
男聲					
	清晰度	自然度		清晰度	自然度
第一句	7.42	7.87	*第六句*	*8.4*	*7.77*
第二句	7.52	7.66	第七句	7.71	7.41
第三句	7.37	7.24	*第八句*	*8.51*	*8.51*
第四句	6.53	6.7	第九句	**6.52**	**6.06**
第五句	**6.32**	**5.75**	第十句	7.05	7.13
平均				7.335	7.21

三、四句算是清晰度與自然度較低的，而第五、七、九句爲較高的三句，對照表 3 也可以發現，將原本是女性的聲音調整成男性的聲音(第五、七、九句)的效果是比較好的，而將原本是女性的聲音調整成女高音的聲音(第三、四句)則效果較差。

再由平均值來看，在自然度上男聲、女聲兩者的差異不大，而在清晰度上明顯的是女聲部分較高，這樣的現象我們分析後發現，將兩句原始錄音（其中一句是男生唸的，另一句是女生唸的）放出來聽時，可聽出原始的男聲語句在清晰度上就比女聲語句要差一些，所以才會得到這樣的結果。不過綜合來看，不論是清晰度或是自然度上都還算是中上的水準，因爲經過調整後的語音信號，或多或少都會導入雜訊而影響其清晰度，另外試聽者也事先知道聽測語句是由男聲(或女聲)所轉換過來的，所以在自然度的評估上會給比十分低一些的分數。

另外，由合成語句的音色來辨別說話者是不是同一人的辨別度評估，我們的評估方法是：隨機從男聲(或女聲)部分的 10 個合成語句中選取出 10 個測試語句，可重複選取同一句，然後依照被選取的順序撥放測試語句給試聽者聽，讓他分辨這 10 句話是由幾個人說出來的，最後以試聽者回答的人數 P 除以真正不同音色的人數 Q，作爲辨別度的評分。舉例來說，若隨機選的 10 句中有 4 句重複，則 Q = 6，而分辨出的人數 P 有 4，5，3，7 等四種數值(設試聽者有 4 人)，此時辨別度的計算爲 (4 + 5 + 3 + 5) / (6 * 4) = 70.83 % ，請注意第四位試聽者回答爲 7 個不同的音色，但是 Q = 6，因此將第四位的 P 改爲 P'，即令

$$P = Q - (P - Q) \quad , \quad \text{if} \quad P > Q \qquad (14)$$

如此，辨別度越高，就代表各個音色越不同，我們的辨別度實驗結果如表 5 裡所示的，數值上比 85 %高，即

表 5　辨別度評估之結果

	男聲	女聲
辨別度	85.7 %	92.9 %

10 種轉換出的音色平均會被認爲是由 8.5 個人說出來的，另外也可發現由女生聲音轉換出的不同音色，具有較高一些的辨別度，不過，我們還不能做這樣的結論，就是女生聲音的音色轉換會具有較高的辨別度，因爲只實驗過一個男生與一個女生的聲音而已。

6.2 即時性測試

本系統的輸入方式分爲由檔案輸入以及麥克風輸入兩種，而取樣頻率分爲 11,025Hz，22,050Hz 以及 44,100Hz 等三種，同時又有四個參數可以調整，因此在即時性上的測試是以當四個參數都作調整的情形下進行的，即時的定義是，當連續輸入語音信號給系統處理時，由喇叭輸出的音色轉換過的語音信號，不會發生斷斷續續的情形。考慮輸入方式與取樣頻率的各種組合，測試的結果爲，當由檔案輸入而由喇叭輸出時，則前述三種取樣率下都可達到即時音色轉換，不過，當取樣率高於11,025 Hz 時且由麥克風輸入時，就無法在所有參數皆調整的情況下達到即時的要求，此時如果聲道內部比例之參數不作調整時，則以麥克風輸入且取樣率高到 44,100 Hz 仍可以達到即時的要求。

7. 結論

本文研究了利用聲道特性來轉換出多樣音色的方法，並依據此方法去實作出一個即時的半自動人聲配音系統。在研究的過程中，我們考慮了發聲器官的數項機械特性，如聲帶振動頻率，聲道長度，聲源訊號調整，聲道內部比例等，以改變這些特性來模擬聲道形狀的變化，希望以此得到多樣化的音色。由於這裡的半自動人聲配音系統，須作即時的基週偵測的處理，且基週位置偵測的準確性對於下游的音色轉換處理及語音品質有很大的影響，所以，我們花了不少心力於即時基週偵測問題的研究上，目前，基週偵測的正確率可達 95%。此外，以所製作的系統合成出的語音作聽覺測試，結果顯示我們的方法的確可以轉換出多樣化的音色。

在本文的研究裡，除了沿用我們先前提出的改變音色的方法之外，還嘗試了改變聲道內部比例的因素，將聲道內部各區間的長度作拉長或縮短的調整，然後將合成出的語音信號再拿去作 LPC 分析的驗證，結果的確可達到調整的目的，不過這種調整的方法卻也或多或少會影響到聲道各區間的原始截面積大小，因爲經 LPC 分析所得到的剩餘訊號，並不是只有乾淨的聲源訊號而已，它還包含了聲道的訊息在內，所以，建立更精確的聲道模型，讓聲源信號更乾淨地從語音信號中分離出來，是需要進一步研究的。

雖然由某甲的音色固定地轉換成某乙的音色的轉換方法，已有不少的研究成果被提出，但是，我們的目標是希望由一種音色轉換出多樣的音色，並且希望透過聲道機械特性的改變來達成，我們認爲這才是根本的解決方法，而基於統計、訓練的方法，則很難有一致性(不同的音素要訓練、建立不同的對應關係)，與多樣性(多種音色)。

國科會計畫編號: NSC 88-2213-E-011-035

參考文獻

[1] H. Kuwabara and Y. Sagisaka, "Acousti Characteristics of Speaker Individuality: Control an Conversion", Speech Communication, Vol. 16, pp 165-174, 1995.

[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, pp. 153-164, 1995.

[3] Y. Stylianou and O. Cappe, "A System for Voice Conversion Based on Probabilistic Classification and a Harmonic plus Noise Model", IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Vol. 1, pp 281-284, 1998.

[4] 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.

[5] N. Iwahashi and Y. Sagisaka, "Speech Spectrum Conversion Based on Speaker Interpolation an Mulit-functional Representation with Weighting b Radial Basis Function Networks", Speech Communication, Vol. 16, pp. 139-152, 1995.

[6] G. Baudoin and Y. Stylianou, "On th Transformation of the Speech Spectrum for Voice Conversion", Int. Conf. on Spoken Language Processing, Vol. 3, pp. 1405-1408, 1996.

[7] H. Y. Gu and W. L. Shiu, "A Mandarin-syllable Signal Synthesis Method with Increased Flexibility i Duration, Tone and Timbre Control", Proceedings o the National Science Council, Republic of China, Part A: Physical Science and Engineering, Vol. 22, No. 3 pp. 385-395, 1998.

[8] D. G. Childers, "Glottal Source Modeling for Voice Conversion", Speech Communication, Vol. 16, pp 127-138, 1995.

[9] P. H. Milenkovic, "Voice Source Model for Continuous Control of Pitch Period", J. Acost. Soc. Am., Vol. 93, No. 2, pp. 1087-1096, 1993.

[10] L. R. Rabiner, et al., "A omparative Performance Study of Several Pitch Detection Algorithms", IEE trans. Acoust., Speech, and Signal Processing, pp 399-418, Oct. 1976.

[11] J. D. Markel and A. H. Gray, Jr., Linear Prediction of Speech, New York: Springer-Verlag, 1976.

[12] Y. Medan, E. Yair, and D. Chazan, "Super Resolution Pitch Determination of Speech Signals" IEEE trans. Signal Processing, pp. 40-48, Jan. 1991.

[13] 古鴻炎、譚百華，"歌唱聲至樂器聲之即時轉換系統"，全國計算機會議論文集(中壢)，第 228-234 頁，1995。

[14] J. F. Wang, et al., "A Hierarchical Neural Network Model Based on a C/V Segmentation Algorithm for Isolated Mandarin Speech Recognition", IEEE trans Signal Processing, pp. 2141-2146, Sep. 1991.

[15] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall,1993.