時間比例基週波形內差 -- 一個國語音節信號合成之新方法

Time-Proportionated Interpolation of Pitch Waveforms -- A New Method for Mandarin Syllable-Signal Synthesis

古鴻炎　　許文龍
Hung-yan Gu and Wen-lung Shiu

國立臺灣工業技術學院 電機系
台北市基隆路四段43號
Department of Electrical Engineering, National Taiwan Institute of Technology
Taipei, Taiwan, R.O.C
e-mail: root@guhy.ee.ntit.edu.tw

摘要

在許多文句翻國語語音的系統裡，都採用音節爲語音合成之單位，因此本論文針對文句翻國語語音系統研究了一個音節信號合成之新方法，稱爲時間比例基週波形內差法。其特色是，除了具有和其它時域合成方法一樣的清晰音質之外，還提供了較多的信號控制之自由度，包括音調(或基週軌跡)之控制，以便由第一聲音節去合成其它聲調的音節；音長之控制，以調整說話速度及反映其它因素對音節長度之影響；以及聲道(vocal track)長度之控制，以便使男生原音合成之女生聲音較爲自然，並可用以合成卡通人物的聲音，這是一個新的嚐試。雖然其它時域合成方法也有提供音調、音長之控制，但是我們的合成方法提供的自由度較高，且已讓這兩控制因素間的相互干擾降低很多。

國科會補助專題研究計畫，編號：NSC 85-2213-E-011-046

1、導言

一個中文文句翻國語語音的系統，可看成是由兩個主要的組件串接而成，其中一個我們稱爲注音與韻律(prosodic)處理單元，它負責將輸入文句所對應的注音查出，然後設定各個語音合成單位的韻律控制參數，包括音調(基頻軌跡)、音長(duration)、音強(intensity)、與音前停頓(pause)，對國語語音合成來說，音節是最常被採用的語音單位。另外一個組件我們稱爲語音信號(或波形)合成單元，它必須能夠依據前一個單元輸出的語音單位編號及相隨的韻律控制參數去合成出語音信號，除了要忠實地接受韻律參數的控制外，也要合成出具有清晰音質的信號，本論文所研究的主要就是在語音信號合成這個單元上，實際上則提出了一個國語音節合成的新方法，它可在具有一定清晰度的條件下，提供更多控制上的自由度。

過去，雖然已有一些中文文句翻語音的系統被提出〔1,2,3,4〕，但是許多系統的研究重點是在韻律處理單元，關於語音信號合成單元則採用既有的技術、或作部份的修改，因此，合成出來的語音信號常受所用技術的限制而有一些缺點，如 LPC 技術〔5,6〕合成的語音信號不清晰、音質不好，共振峰合成技術〔7,8〕雖可獲得較佳的音質，但是相鄰音素(phoneme)間信號轉移(transition)之模擬仍不理想，並且需以人工來調整、設定大量的音響(acoustic)控制參數(即未有一套理想的自動化的參數值估計方法)。最近，一種直接在時域波形上操作的技術(稱爲 PSOLA, pitch-synchronous overlap and add)被提出〔9,10,11〕，它不但可合成出清晰音質的語音信號，並且控制參數數值(即基週之時間位置)較容易決定，不過它只提供侷限的音長控制之自由度，即不接受音長比值(合成音長除以原始音長)被任意設定成合理範圍內(如 0.5~2.0)的一個數值，因此，後來一種變形的技術稱爲 LP-PSOLA 被提出〔11,12〕，以提高音長控制之自由度，可是它顧此失彼，反而導入雜訊使合成的語音信號變得不清晰。再者，PSOLA 技術對國語音節

之合成還存在有缺陷，一個例子如以第一聲/ai/音節去合成第四聲之/ai/時，原本第一聲/ai/之/a/與/i/若各佔一半的時間(這裡爲了說明才用二分法，實際上是逐漸轉移的)，則合成之第四聲/ai/裡，/i/部份會比/a/部份長，因爲第四聲在信號波形上的特徵是週期由小變大，而 PSOLA 技術爲了改變音調的一種作法是單純地把週期逐漸拉大，這樣就造成了/i/部份會比/a/部份長之副作用，實際上就是在時間軸上扭曲(time warping)頻譜變換之速度，另外一種作法是捨棄幾個信號週期後再把剩下的週期逐漸拉大，以使音長維持固定，但是直接丟棄信號週期必然會使頻譜走勢變成不連續(特別是在一個雙母音音節中)，這樣爲了控制一個因素而導致另一個因素失控(反觀人本身並無此種失控情形)，並不能算是一個好的解決方法。

爲了改進前述語音合成技術的缺點，本論文提出了一個國語音節信號合成的方法，稱爲時間比例基週波形內差法(Time-Proportionated Interpolation of Pitch Waveform, TPIP)，它也是在時域上直接對信號波形作處理，因此合成的語音信號具有一定的清晰度，不過它提供的音長控制之自由度卻是比 PSOLA 技術高很多，這也是當初研究此技術的一個重要動機。至於音調控制之自由度，對國語語音合成來說則是一項基本的、不可缺少的的控制因素，這樣才能接受韻律處理單元的控制去合成出國語裡的各個聲調或整句的句調，而所提出之方法不僅能配合各種音調去合成出語音信號，並且沒有 PSOLA 技術裡的因爲改變音調而破壞頻譜變換速度的副作用。除了音長、音調控制之自由度外，我們也增加了另外一個自由度，即可讓合成語音的共振峰頻率全體(F1,F2,F3,...)作相同倍率之升高或降低的控制，提供這種控制的動機是，當把音調(F0軌跡)提高以便由男生發的原始音節波形去合成女生聲音時，合成的語音信號聽起來總是有男生在假裝女生聲音的感覺，這說明男女生除了基頻的高低差異之外，還有其它先天上的差異(發音習慣可說是後天上學來的)，其中重要的一項是聲道長短的差異，一般說來男生聲道較女生的爲長，並且由聲道的音響模型可知

〔13,14〕，聲道長度和共振峰頻率值之間存在者反比的關係(聲道短則共振證頻率高)，因此我們相當於提供了聲道變長變短之控制。關於音長、音調、聲道長等三項因素之控制，我們提出的音節信號合成方法，在合理的參數值範圍內各個因素可說是獨立的、不相互干擾的，這可由頻譜分析圖及所建造的原型文句翻國語語音系統來得到驗證。

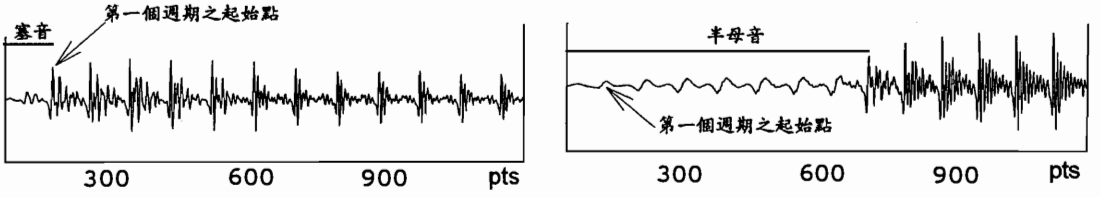
2、國語音節結構 與 無聲部份信號合成

一個國語音節的信號可看成是由無聲(voiceless)部份與有聲(voiced)部分兩部份串接組成，無聲部份的信號對應於波形無週期性之塞音(stop)、擦音(fricative)、或塞擦音(affricate)，而有聲部份的信號則對應於波形有週期性之鼻音(nasal)、滑音(glide)、流音(liquid)、或母音(vowel)，如果一個音節全由有聲之音素構成，則我們可把此音節當作無聲部分極爲短暫的音節，即把第一個週期前的信號看作是無聲的部分，如此，每一個國語音節都可說是具有 UV 之結構，U 與 V 分別表示無聲與有聲部份。

由於一個合成音節的時間長度是由韻律處理單元輸出的音長參數來決定的，因此，當要合成一個音節的信號時，需先依據音長參數去決定無聲、有聲兩部份各要佔據多少時間，然後才分別去合成無聲與有聲兩部份的信號。這裡將先介紹無聲部份在信號上的分類，然後對不同類的無聲信號去說明無、有聲兩部份的時間分配作法及無聲信號的合成方法，其重點在於如何配合韻律處理單元的音長要求但不破壞無聲音素的特性(即要清晰可辨)，至於有聲部份的合成方法則在下一節介紹。

對於音節之無聲部份的信號，我們歸類成兩種信號類別，以方便作時間分配與信號合成之處理，第一類稱爲短暫無聲，表示其無聲部份的時間

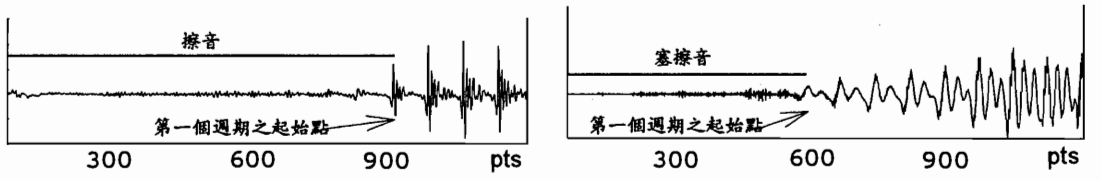
很短，第二類稱爲長帶無聲，表示其無聲部份的時間相對地長很多。我們希望能把非送氣塞音(如/ㄅ/)開頭之音節(波形例子如圖1(a)所示)、半母音(含鼻音、滑音、與流音)開頭之音節(波形例子如圖1(b)所示)、及母音



(a) 塞音開頭音節/ㄅㄚ/之波形　　(b) 半母音開頭音節/ㄇㄚ/之波形

圖1　具短暫無聲部份之音節波形

開頭之音節都歸屬爲具有短暫無聲部份之音節，而把擦音開頭(如/ㄏ/)之音節(波形例子如圖2(a)所示)、非送氣與送氣塞擦音(如/ㄐ,ㄑ/)開頭之音節(波形例子如圖2(b)所示)、及送氣塞音(如/ㄆ/)開頭之音節都歸屬爲具



(a) 擦音開頭音節/ㄏㄚ/之波形　　(b) 非送氣塞擦音音節/ㄐㄧㄚ/之波形

圖2　具長帶無聲部份之音節波形

有長帶無聲部份之音節，如此在實作上，就可以依據一個音節的第一個有聲信號週期之起始點的時間位置來對無聲部份作分類，我們使用的門檻是300個樣本點(取樣頻率爲11,025Hz)，即音節裡的第一個週期起始點在300點以內者就歸屬爲具有短暫無聲，否則就歸屬爲具有長帶無聲。

依據原始音節波形裡的第一個週期起始點的位置，對它的無聲部份作分類後，接著就要作無、有聲兩部份的時間分配及無聲部份之信號合成。如果原始音節具有短暫無聲部份，則將原始音節裡第一個有聲週期起始前

的樣本點直接拷貝到合成音節的起始部份，以作爲合成音節的無聲部份，然後將剩餘時間(從音長參數扣除無聲部份之時間)分配給有聲部份；如果原始音節具有長帶無聲部份，則要依據原始音節裡無、有聲兩部份的時間比例去分配音長參數所給定的時間，假設原始音節的長度爲300ms,且無、有聲兩部份的時間比例是4:6，再令音長參數給定的時間是 R 毫秒，此時，若 R*(4/10) > 300*(4/10)*1.5，則只分配 300*(4/10)*1.5 毫秒給合成音節之無聲部份，否則分配 R*(4/10) 毫秒給無聲部份，剩餘時間自然就分給有聲部份，如此之時間分配，相當於限定無聲部份最多只能佔據原始音節無聲部份的時間長度的1.5倍，這是考慮人自己將一個音節唸得很長時，主要是將有聲部份延長而無聲部份並非等比例延長，得知無聲部份的時間長度後，接著將原始音節裡開頭的300個樣本直接拷貝到合成音節的起始部份，以保存塞擦音的起始塞音特性，接著按照原始音節與合成音節之無聲部份的時間比例去內差出其它的無聲部份之樣本值，內差的方法以圖3來說明，

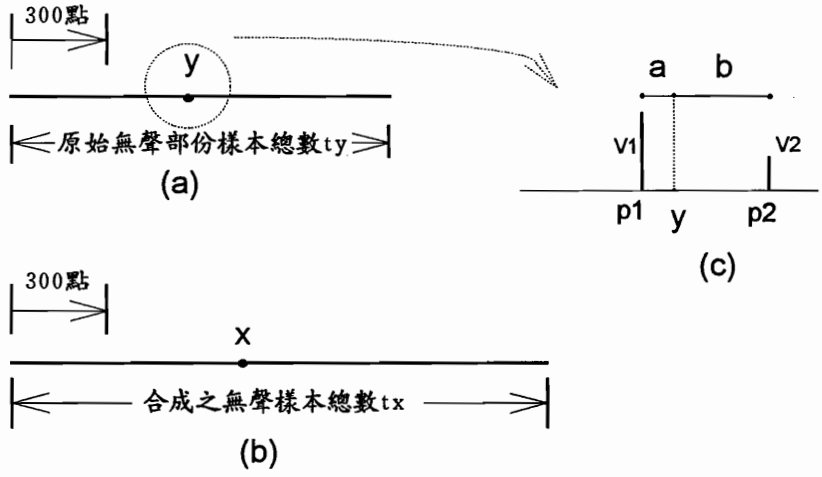


圖3　合成無聲樣本值之內差法示意圖

圖3(b)裡的 x 表示欲合成之無聲部份的一個樣本點，tx表示總共要合成的點數，圖3(a)裡的ty表示原始音節無聲部份的總點數，y是x依時間比例所對應之點，則

$$y = \frac{x-300}{tx-300} * (ty-300) + 300 \qquad (1)$$

很明顯的，由(1)式算出的 y 值並不會剛好爲整數，設介於圖3(c)中的p1和p2兩整數點之間，a爲y點到p1的距離，b爲y點到p2間的距離，v1、v2爲樣本值，我們就簡單地以線性內差來計算x點上的振幅值，如下式所示

$$x\_SampleValue = v1 * \frac{b}{a+b} + v2 * \frac{a}{a+b} \qquad (2)$$

雖然只應用簡單的線性內差來算時間位置及樣本值，但實際聽測時並未發現無聲部分有誤辨之情形。

3、有聲部份信號合成

前一節裡已說明如何將音長參數給定的時間分配給無、有聲兩部份，當知道有聲部份的時間長度後，接著就要依據韻律處理單元輸出的音調參數去計算出一序列的週期長度值 $L_1, L_2, L_3, \ldots$，即基週軌跡之計算，然後才去計算各個週期應具有的波形(或樣本值)。在觀念上，週期長度序列 $L_1, L_2, L_3, \ldots$ 的求取，和各個週期內的樣本值的求取是兩件獨立的工作，因此，在3.1節提出的週期長度值的計算方法，只代表我們的原型文句翻國語語音系統所使用的一種簡單、可行的作法，並不表示我們非常推薦它，至於3.2節提出的一個週期內各樣本值的求取方法，即本文題目所稱的時間比例基週波形內差法，則是我們非常推薦採用的。

3.1 基週軌跡計算

韻律處理單元爲了讓一個音節擁有特定的聲調(當然也可把句調考慮進來)，那麼它可透過音調參數之輸出去告知音節信號合成單元，在我們的原

型系統裡，音調參數是指用以逼近基週軌跡之6個線段的7個端點頻率值，這6個線段各分配音節有聲部份的 1/6 時間，當使用更多的線段時，基週軌跡自然會更平滑。

關於週期長度之計算，我們以圖4裡的一個線段爲例來說明，圖中Freq1和Frq2表示此線段的兩個端點頻率值，我們的目的是要求取在此線段

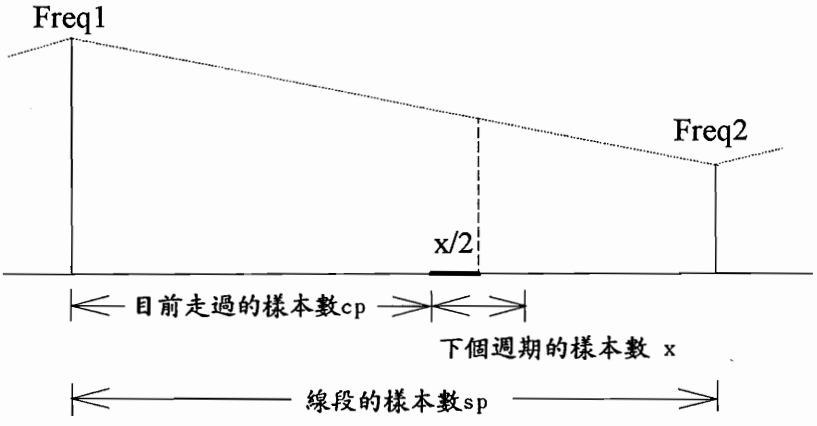


圖4　週期長度求取之示意圖

內各個週期以樣本點數來算的時間長度，設圖中目前要加入的週期，其長度爲 x 個樣本點，則 x 的數值可依據如下的線性時間比例關係來求取：

$$\left(\frac{cp+\frac{x}{2}}{sp}\right)\left(\frac{11025}{Freq2}-\frac{11025}{Freq1}\right)+\frac{11025}{Freq1}=x \qquad (3)$$

式子裡的 11,025 是取樣頻率，等式的觀念是以兩端點頻率對應的週期長度來內差出 x 的樣本點數，並且一個週期是以其中心點爲代表，所以cp要加上x/2。當一個週期跨越到下一個線段時，我們就看週期中心點是否已超出本次的線段，如未超出則保留此週期，否則就將剩餘時間轉給下一個線段。

3.2 週期內樣本值求取

在計算出一序列的週期長度值 $L_1, L_2, L_3, ...$之後，接著就要去合成各個週期的波形，即計算出週期內的各個樣本值，這裡我們提出一種新的求取方法，就是前面提到的時間比例基週波形內差法，它的詳細作法可以如下之四個處理步驟來說明：

(Step 1) 找尋兩個對應之原始信號週期及決定加權:

首先依據欲合成之信號週期的中心點c，如圖5的上方所示，去找出原始第一聲音節波形中兩個對應的信號週期，尋找方法是依據線性時間比例

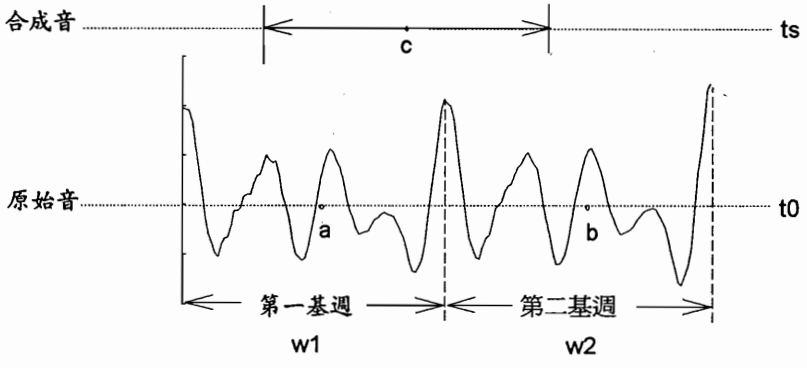


圖5　合成週期與對應之兩原始信號週期

之關係，從原始音節波形中找出兩個相鄰週期的中心點a與b，如圖5的下方所示，使得不等式

$$\frac{a}{to} \le \frac{c}{ts} \prec \frac{b}{to} \qquad (4)$$

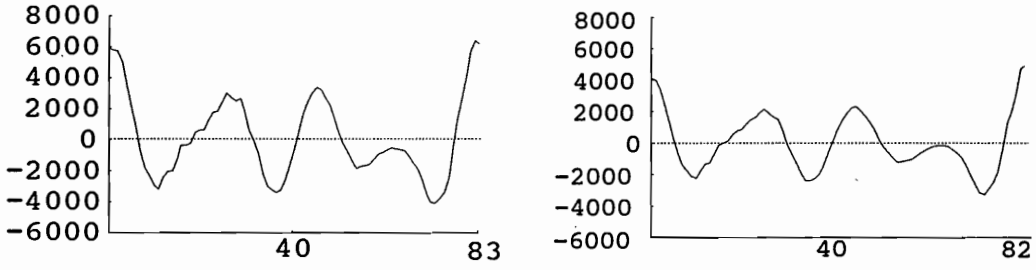
成立，其中ts表示合成音節之有聲部份的總點數，to表示原始音節裡有聲部分之總點數。找到兩個相對應的原始音基週之後，接著要計算二個原始週期波形當被內差組合以合成新的基週波形時，各自的加權值w1與w2要設爲多少，這裡我們仍是以線性時間比例來決定w1、w2的值，如下式:

$$令\ \alpha=\frac{a}{to}\ ,\ \beta=\frac{b}{to}\ ,\ \gamma=\frac{c}{ts}$$

$$設\ w1=\frac{\beta-\gamma}{\beta-\alpha}\ ,\ w2=\frac{\gamma-\alpha}{\beta-\alpha} \qquad (5)$$

(Step 2) 乘上加權值:

此步驟的動作是將找出之第一個原始週期的各個樣本點乘上w1，再將相鄰的第二個原始週期的各個樣本點乘上w2，圖6顯示乘上加權值後的結果，因爲欲合成週期的中心點較靠近第一個原始週期，所以圖6(a)裡的波形的振幅較圖6(a)裡的大。



第一週期乘w1後的波形　　第二週期乘上w2後的波形

圖6 乘上加權值後的兩原始週期波形

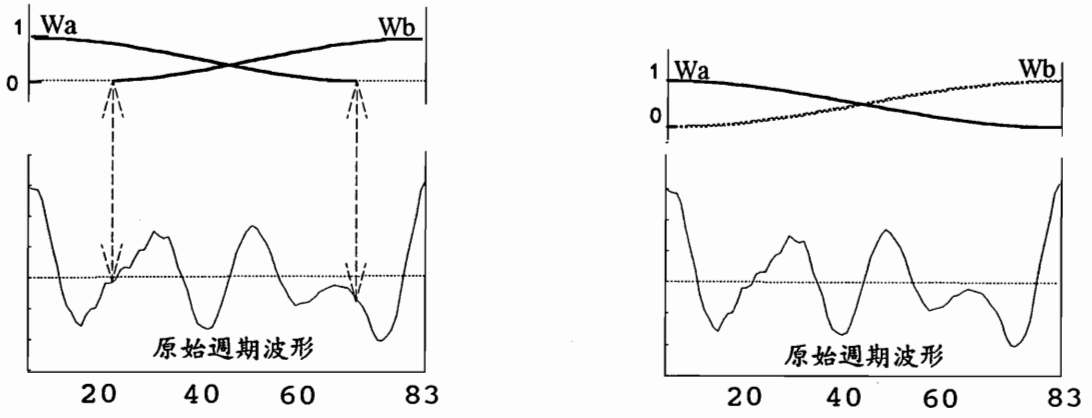
(Step 3) 乘上餘弦窗:

在把兩個原始週期乘上餘弦窗之前需先決定餘弦窗的長度，餘弦窗的長度由原始週期長度和新合成週期長度共同決定，如果原始週期的長度大於新合成週期的長度，則設定餘弦窗長度爲新合成週期長度的兩倍，如圖7(a)所示，圖中橫軸是樣本點數，否則以原始週期長度的兩倍作爲餘弦窗長度，如圖7(b)所示。餘弦窗之函數如(6)式所示：

$$w(n)=0.5+0.5*\cos\left(\frac{2\pi n}{N}\right),\quad 0\le n\le N-1 \qquad (6)$$

決定餘弦窗的長度後，接著就把兩個對應的原始週期個別乘上二個半邊的餘弦窗，如圖7裡的 Wa 和 Wb 表示二個半邊的餘弦窗，左邊的右半餘弦窗

(Wa所表示者)乘上後得到的信號波形就放在新合成週期的左邊，而右邊的



(a) 合成週期長度小於原始週期長時之餘弦窗　(b) 合成週期長度大於原始週期長時之餘弦窗

圖7　餘弦窗長度設定

左半餘弦窗(Wb所表示者)乘上後得到的信號波形就在放在右邊，此時就會得到如圖8(a)與8(c)所示的波形，圖8(a)表示對第一個原始週期之處理，而圖8(c)是第二個原始週期的，然後作疊加的動作，就會得到如圖8(b)與

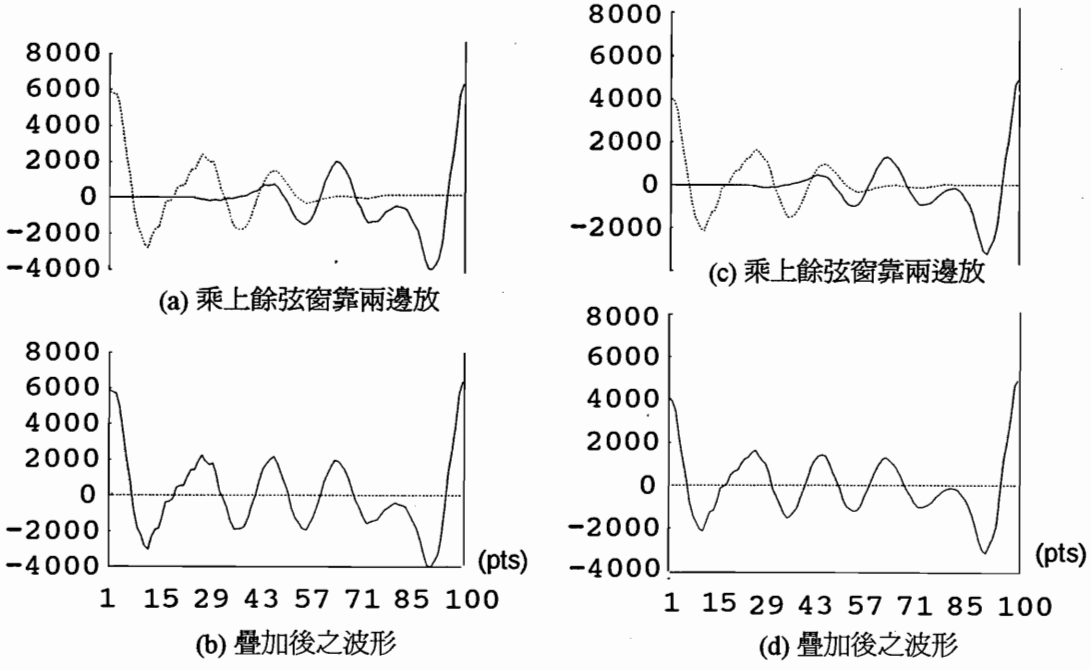


圖8 乘上餘弦窗後靠兩邊放、疊加之兩原始週期波形

8(d)所示的經過處理的原始週期波形(這裡假設新合成週期長度大於原始週期長)。

(Step 4) 相加兩處理過的原始週期波形:

把圖8(b)與圖8(d)所示的兩個經過處理的原始週期波形相加，最後就會得到如圖9所示的新合成之波形，比較圖9和圖5可發現，新合成之波形裡有5個波峰，而兩個原始週期之波形各只有4個波峰，這是因爲這裡的新合成週期長度被設爲100點，比兩個原始週期的長度都長，當週期長度變長，波峰數也要變多，這樣才能維持相同的共振頻率值。

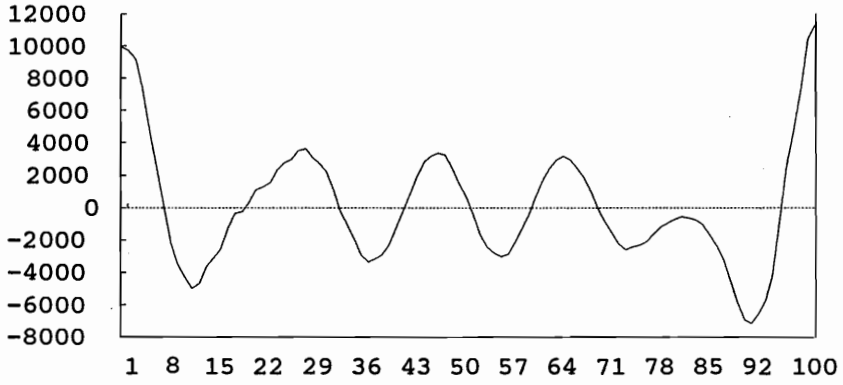


圖9 新合成的信號週期

如此，重覆(Step 1)至(Step 4)的步驟去處理基週軌跡裡的各個週期，就可將音節之有聲部份的信號合成出來，再加上無聲部分的合成處理，則整個音節的信號就可被合成出來了。

3.3 聲道長度調整

由於原始音節信號是由男生錄製的，因此當只把基頻軌跡整個提高來模仿女生聲音時，合成的語音聽起來會像一位男生在假裝發女生聲音，顯

得非常不自然。男、女聲除了發音習慣和基頻高度不同外，聲道的長度也不相同一般來說，男生的聲道較長，女生的聲道較短，因此，我們增加了聲道長度這個控制參數，希望在合成女生聲音時，聽起來較自然。

由聲道的音響(acoustic)模型〔13,14〕可知，聲道長度和共振頻率位置有密切關係存在，即聲道變短共振頻率會提高之反比關係，所以，調整聲道長短，就相當於把各個共振頻率按比例提高或降低。過去，在爲卡通人物配音時，常以錄音機快速放音之方式來將成人聲音轉成小孩聲音，錄音機快放，除了會把基頻提高外，也會把各個共振峰頻率按比例提高，不過，聲音的時間長度卻縮短了。我們希望設計一個具有相當彈性的國語音節合成器，能夠對音長、基頻(F0)軌跡、共振峰頻率(F1、F2...)整體等三項因素幾乎獨立地去控制，在3.2節我們已說明一個可獨立去控制音長與基頻軌跡的方法，這裡我們將說明一個可對共振峰頻率全體獨立去調整的作法，它可被簡單地嵌入3.2節的(Step 1)與(Step 2)之間，而不會破壞其它處理步驟，詳細說來就是應用錄音機快放的道理去作 resampling 的處理，不過是對3.2節(Step 1)找出的兩個原始基週波形去作 resampling，例如當要把共振峰頻率全體調高1.3倍時，就對應於在原始基週波形上，每次要走1.3個取樣點(若取樣頻率不變的話)，如此做，新的取樣點可能會落在原始週期裡的兩個樣本之間，此時，我們可以利用數位訊號處理中的取樣理論將其新樣本值求出，但此種做法太過於費時(計算量大)，在實作上是不可行的，所以我們使用一種以二次多項式去內插的作法，就是把新的取樣點 x 的左邊兩個舊取樣點 x0, x1 和右邊一個取樣點 x2 代入一個二次多項式，而建立如下的方程式：

$$\begin{cases} y_0 = f(x_0) = Ax_0^{\,2} + Bx_0^{\,1} + C \\ y_1 = f(x_1) = Ax_1^{\,2} + Bx_1^{\,1} + C \\ y_2 = f(x_2) = Ax_2^{\,2} + Bx_2^{\,1} + C \end{cases} \tag{7}$$

以圖10來說明，我們可以令x0、x1、x2之值爲0、1、2，而讓y0、y1、y2表示三個樣本值，如此式(7)便形成一個三元一次方程組，而其解爲：

$$\begin{cases} A = y_2 - 2y_1 + y_0 / 2 \\ B = -y_2 + 4y_1 - 3y_0 / 2 \\ C = y_0 \end{cases} \tag{8}$$

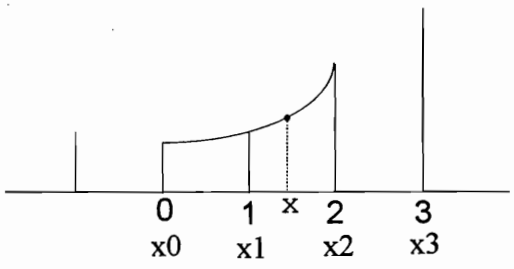


圖10 二次多項式內差之示意圖

求得A、B、C後，再將圖10的 x 座標值換成 x-x1+1 代入，如此繼續，最後便可得到如圖11所示的波形，明顯地週期長度已改變爲原先長度的 1/1.3 倍，接著再依據圖11的兩個經過 resampling 的原始週期去進行3.2節裡的

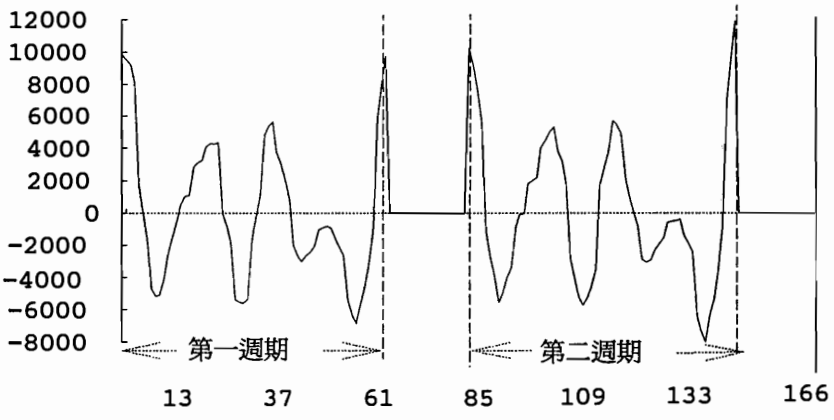


圖11 一次走1.3個取樣點之 resampling 後的兩個原始週期

(Step 2)至(Step 4)之處理，此時，餘弦窗的長度也必需依據 resampling 過的原始週期長度及合成週期長度來決定，如此，處理完後便會得到如圖12所示的波形，將此圖與圖9比較可發現，在相同週期長度的條件下，圖12

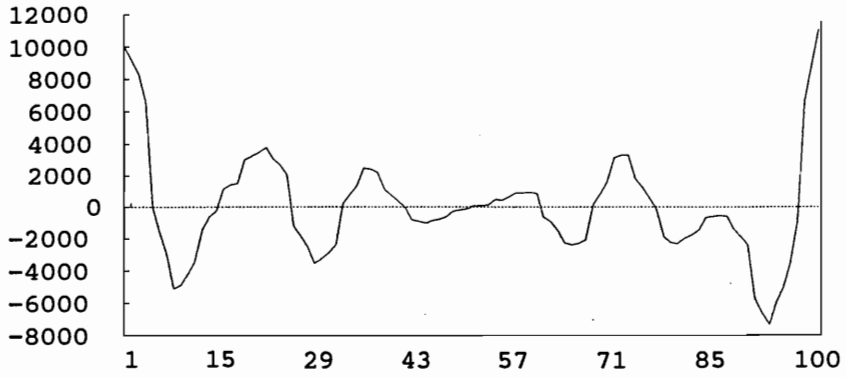


圖12 一次走1.3個取樣點之新合成週期

裡有6個波峰，而圖9只有5個波峰，所以，圖12裡的波形具有較高的共振頻率值，這正是我們我們原先所想要的。

一般來說女生的聲道比男生的短，所以共振頻率會較高，因此在調整基準音高時，我們也希望共振峰頻率(F1,F2...)整體一起被連帶地調整，如此以避免發生男生裝女生聲音的感覺，所以，我們便依據男女生的平均基頻值和平均第一共振峰F1的比值來建立如下的調整公式：

$$WalkPoints = 1 + 0.15 * \left(\frac{NewF0}{120} - \frac{OldF0}{120} \right) \qquad (9)$$

基中，NewF0 表示新設定的基準音高，OldF0 表示原錄製音節信號的基準音高，120是男生的平均基頻值，而1.15是發/ㄜ/音時，男女生平均的F1值之比值，如此，WalkPoint就表示在原始週期波形上每次要走的取樣點數。此外，我們也可設計成讓使用者單獨去設定 WalkPoint 的值，而不改變基準音高值。利用上述作法去調整聲道長度，就可以得到如卡通人物的聲音，或是老沈的聲音。

4、本合成方法之實驗驗證

對於一個音節信號合成單元來說，所合成信號的品質或清晰度(是否混濁、是否有雜訊)是一項重要的評估項目，可是在實際使用上，音節信號合成單元並不是單獨存在的，它通常需和韻律處理單元一起工作(即構成一個文句翻語音系統)，也就是說我們懷疑單獨工作時的表現是否能代表和韻律處理單元一起工作時的表現，因此，我們就實際去建造一個原型的中文文句翻國語語音系統，原始的409個第一聲音節波形是請一位男性播音員來錄製的，使用 11,025Hz 取樣頻率及 16bits/sample 之解析度，在切除 silence 信號後，信號波形共佔 2.24 Mbytes，這顯示時域上的音節信號合成作法，並不如想像中那麼佔記憶體，關於韻律處理單元的製作，我們基本上是參考前人的 rule-based 作法[1]，但也做了一些修改，目前整個原型系統已可即時地唸出國語語音，初步聽測試驗顯示，音質相當清晰，具有和其它時域合成方法一樣的清晰特性，至於可辯度和自然流利度，則和韻律處理單元的好壞有密切關係，因此不宜以此兩項目去評估信號合成單元。

除了清晰度之外，一個好的信號合成單元尚需具備充分的信號控制上之自由度，自由度愈大，則韻律處理單元愈有發揮的空間，例如唸快或唸慢，高昂或低沈，角色扮演(男生、女生或小孩聲音)等。所提出之音節信號合成方法，提供了音調、音長、聲道長之信號控制的自由度，值得注意的是，這三項控制因素在合理的參數值範圍內，都幾乎可獨立去控制(改變數值)，下面就以聲譜(spectrogram)分析來檢查，當一項控制因素被改變數值時，是否會發生副作用。

4.1 音調變換

由於原始音節信號的聲調都是第一聲，因此，當合成出第四聲(或其它

聲調)的音節信號時，共振頻率(F1,F2,...)結構(頻率值及走勢)是否會受到影想，就成爲一個很令人關切的問題。這裡，以音節/ㄚ/爲例，去分析原始之第一聲/ㄚ/音節，其波形如圖13所示，和合成之第四聲/ㄚ/音節，其波形如圖14所示，使用了 Hyperception 公司出品之 Hypersignal 信號分析軟體〔15〕，結果得到如圖15和圖16之聲譜圖，由圖14可看出信號的週期長度一直在增加中，使得對應的圖16之聲譜圖上，基頻及其諧波隨著時間在下降，但是共振頻率則維持水平方向前進，並且具有和圖15之共振頻率一樣的垂直高度。

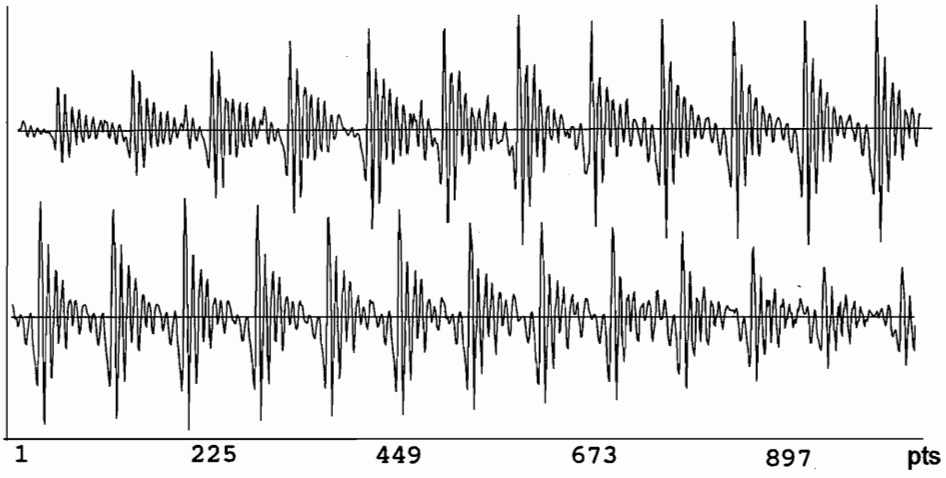


圖13 原始第一聲/ㄚ/之波形

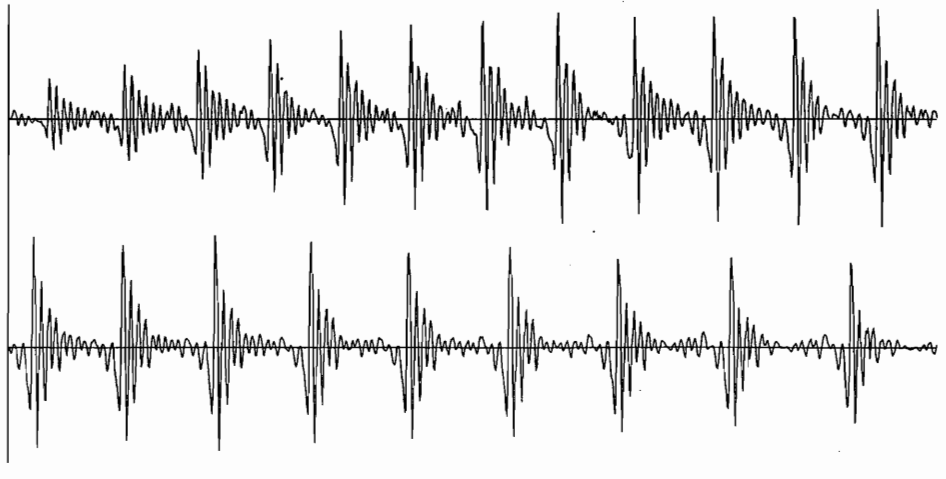


圖14 合成之第四聲/ㄚ/之波形

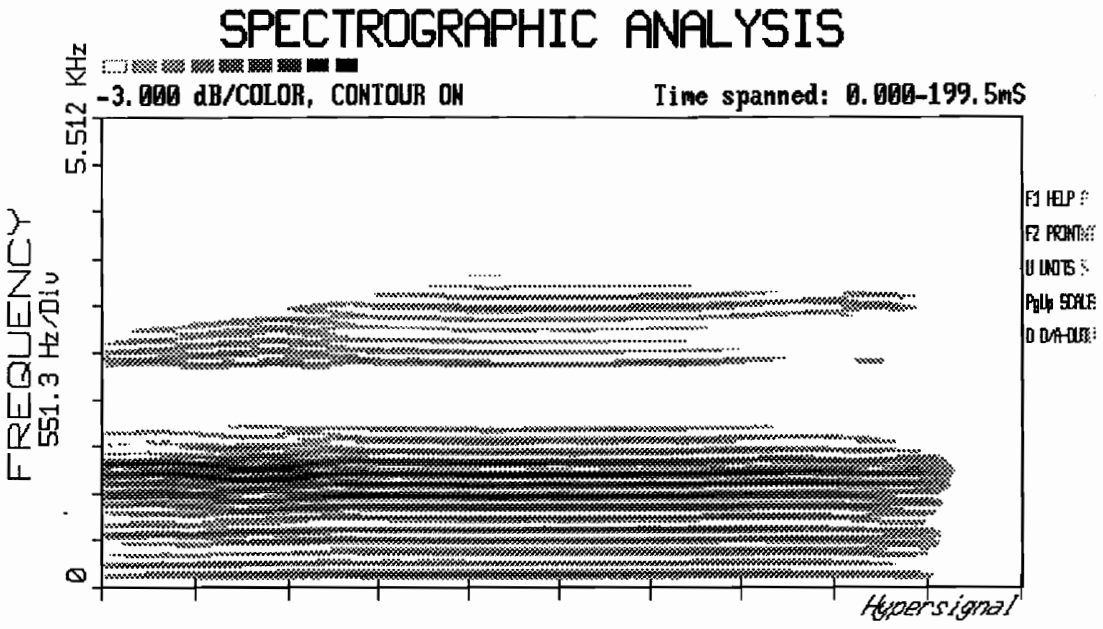


圖15 原始第一聲/Y/音之聲譜圖

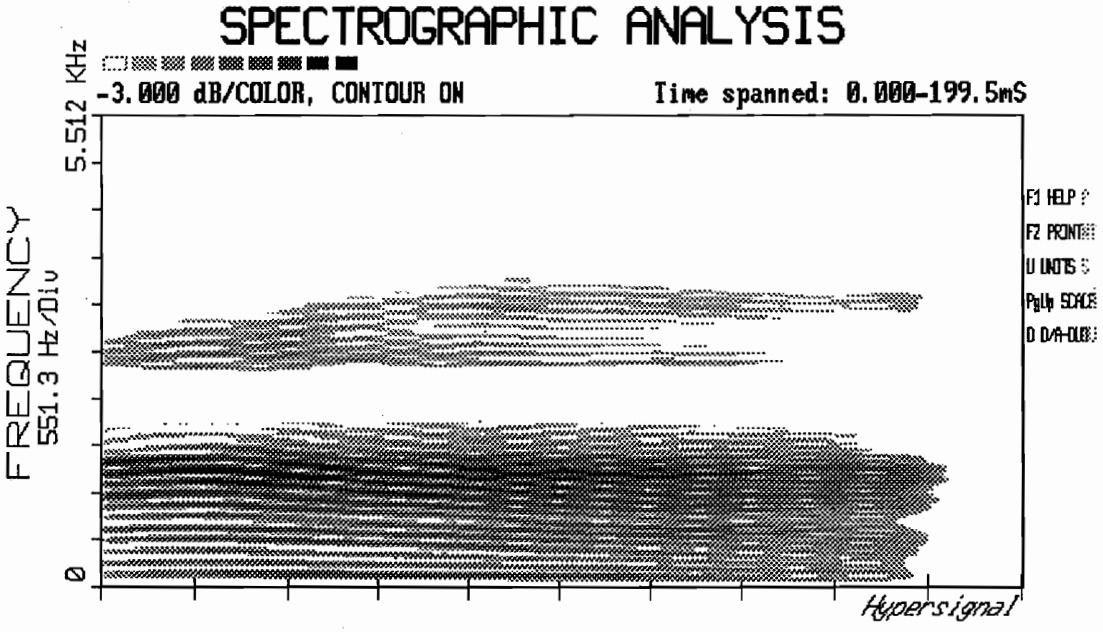


圖16 合成之第四聲/Y/音之聲譜圖

4.2 音長變換

當合成的音節信號，其音長比原始音長大(或小)許多時，共振頻率的走勢(軌跡)是否要成比例地延長(或縮短)，或是延長特定的時間部份(如頻

譜穩定或呈水平時），即要採取線性的或非線性的 time-warping，是一個值得探討的問題，這可以從觀察人自己發雙母音時唸長唸短的差異開始。我們的音節信號合成方法，理論上會採取線性方式來作時間之延長或縮短，不過，這裡仍以實驗分析方式來驗證，以音節/ㄞ/爲例，去分析原始之/ㄞ/音節，其波形如圖17所示，和合成的兩倍長之/ㄞ/音節，其波形如圖18所示，結果得到如圖19和圖20之聲譜圖，比較圖19與圖20，我們看到共振頻率F1、F2的高度及走勢並無不同之地方，不同的是時間軸的尺度，所以我們的合成方法的確是以線性方式來延長時間。

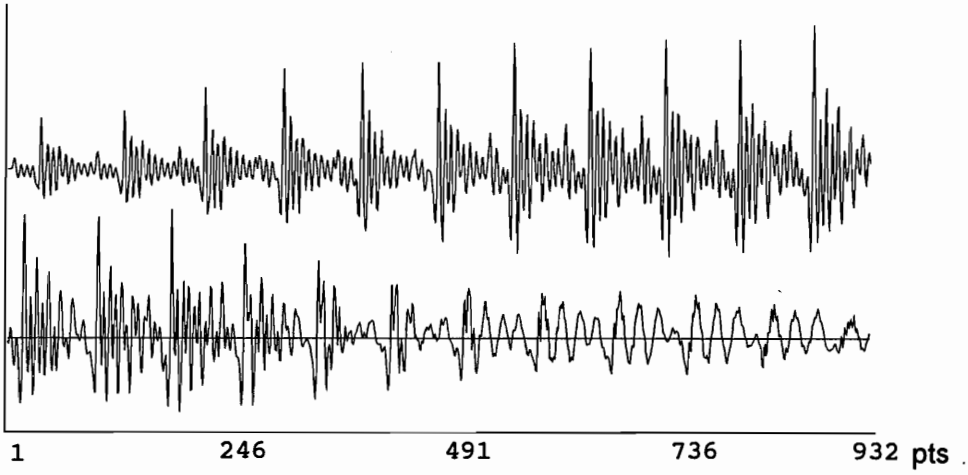


圖17 原始/ㄞ/音之波形

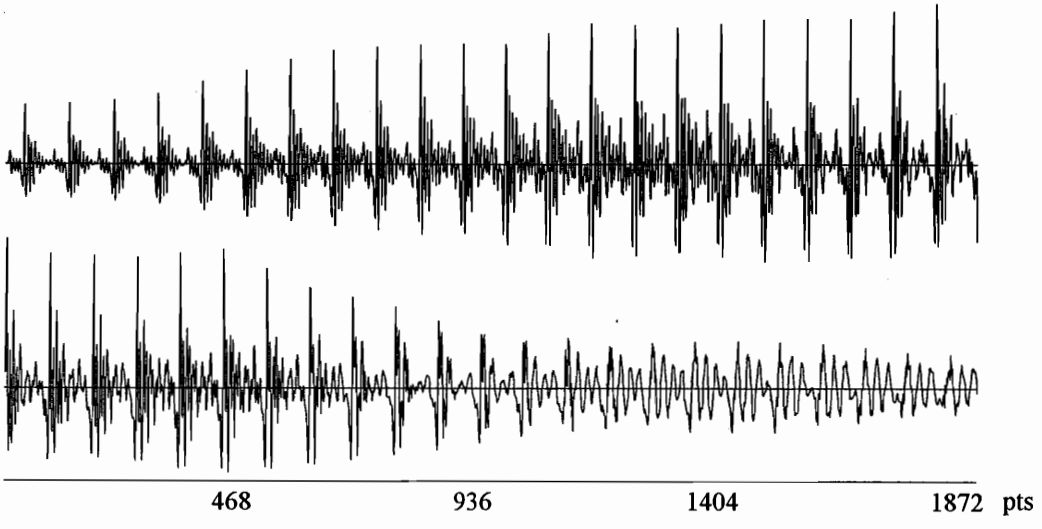


圖18 合成之二倍音長/ㄞ/音之波形

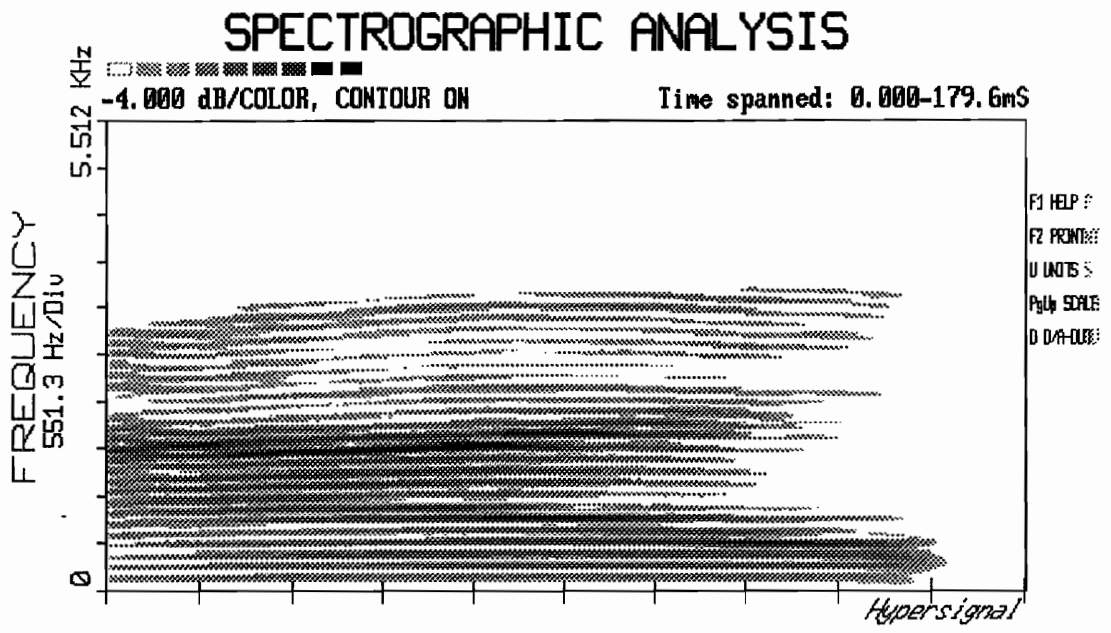


圖19 原始/ㄌ/音之聲譜圖

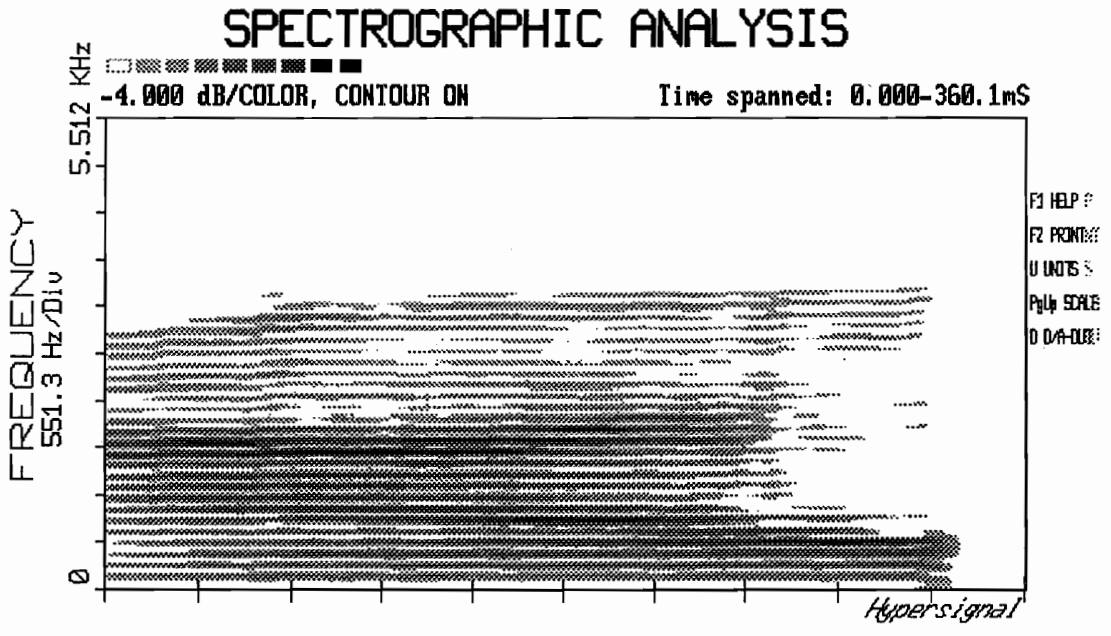


圖20 合成之二倍長/ㄌ/音之聲譜圖

4.3 聲道長變換

我們的音節合成方法以 resampling 來達到提高(或降低)共振頻率的目的，這就相當於縮短(或加長)聲道長度，不過，作 resampling 和音調控制

是兩件獨立的事情，也就是說可在相同音調的條件下去調整共振頻率高度，如圖21和圖22的波形，它們具有相同的音調、相同的週期長度，但是圖21是作一次走1.3點之 resampling 來合成的/ㄚ/音波形，而圖22則是作一次走0.7點之 resampling 來合成的/ㄚ/音波形。圖21和圖22波形所對應的聲譜圖分別如圖23和圖24所示，圖21波形每個週期裡的波峰較多，意味有較高的共振頻率，所以圖23裡我們看到的共振頻率高度要比圖24裡的高許多，事實上，圖23裡的共振頻率會是圖15的共振頻率的1.3倍高，而圖24裡的共振頻率會是圖15的共振頻率的0.7倍高。

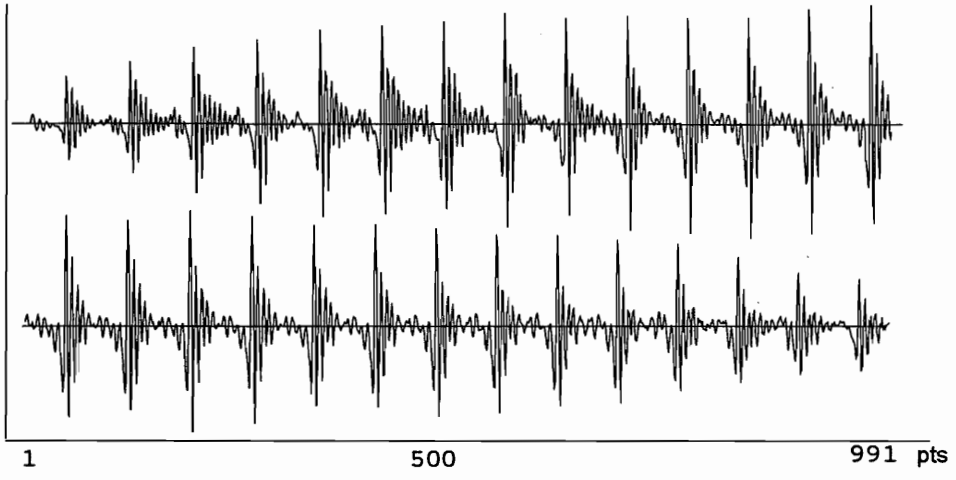


圖21 一次走1.3點之合成/ㄚ/音的波形

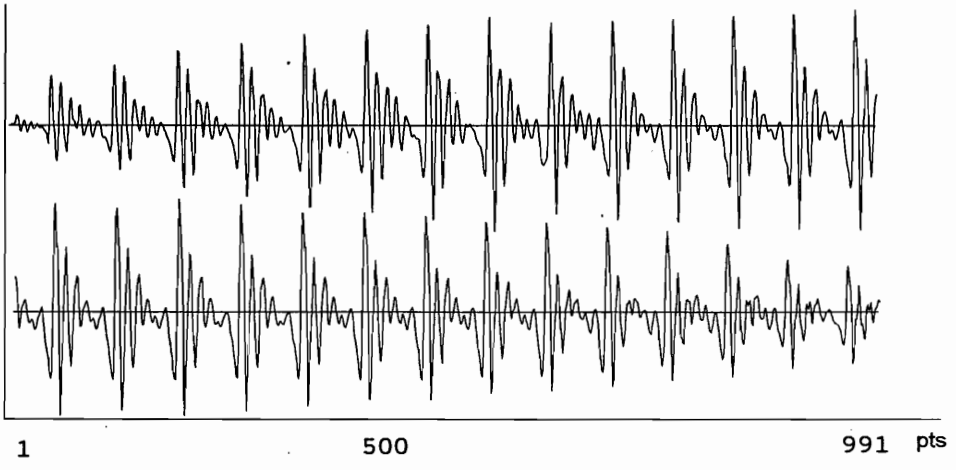


圖22 一次走0.7點之合成/ㄚ/音的波形

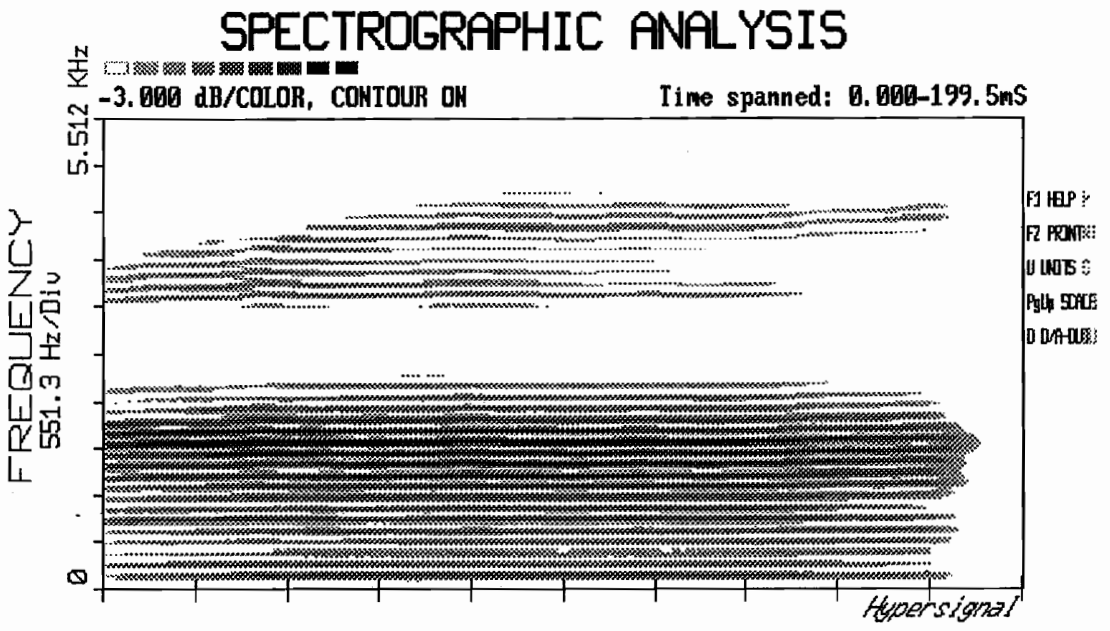


圖23 一次走1.3點之合成/ㄚ/音的聲譜圖

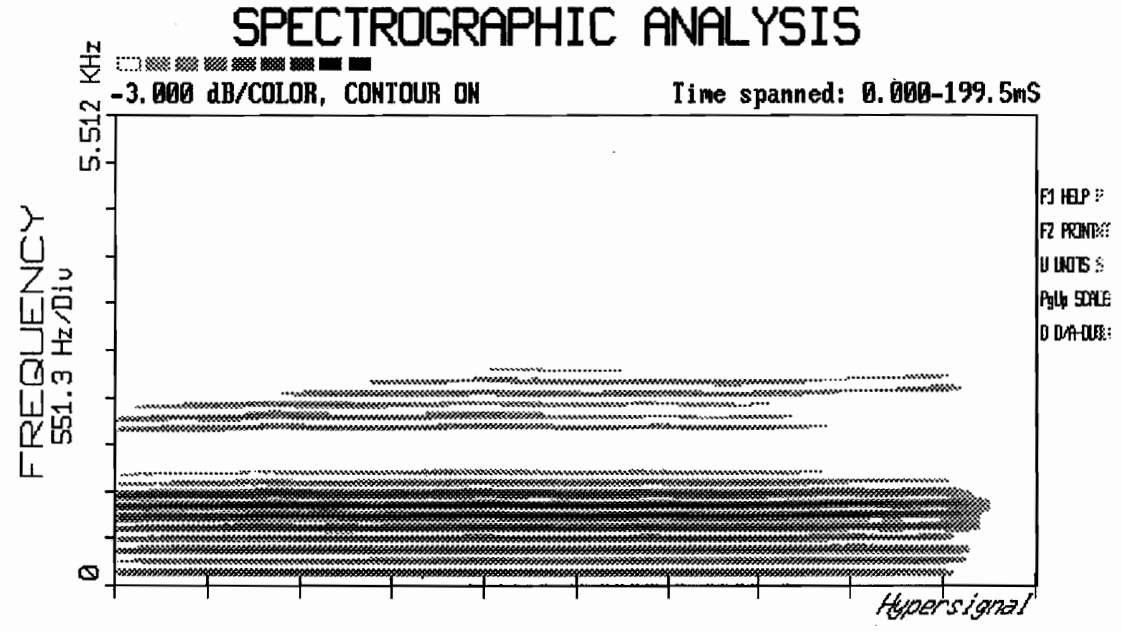


圖24 一次走0.7點之合成/ㄚ/音的聲譜圖

5、結語

本論文提出一個國語音節信號合成的新方法，它的特色在於保留時域

合成方法的清晰音質之條件下，加強了信號控制之自由度，如音調之控制，不會再導致頻譜或共振峰走勢在時間軸上被扭曲；音長之控制，較 PSOLA 技術更具有彈性；聲道長之控制，是一項新的嘗試，以前的文句翻語音系統並未提供，它使得較自然的、及更多的音色能被合成出來，而豐富的音色是拓展文句翻語音系統之應用範圍的基礎，新的應用範圍如雙(或多)主播之新聞播報、小說與故事之講述、甚至於戲劇裡的對話的合成。

除了提出音節信號合成之方法，我們也以此方法去建造了一個原型的中文文句翻國語語音系統，初步聽測合成之語音信號，顯示所提出之音節信號合成方法的確能合成出清晰的語音，並且能夠依照預先的想法，讓前述的三項控制因素獨立地去改變數值。

參考文獻

[1] Lee, L. S., C. Y. Tseng and C. J. Hsieh, "Improved Tone Concatenation Rules in a Formant-Based Chinese Text-to-Speech System", IEEE trans. Speech and Audio Processing, Vol. 1, No. 3, pp. 287-294, 1993.

[2] Chiou, H. B., H. C. Wang and Y. C. Chang, "Synthesis of Mandarin Speech Based on Hybrid Concatenation", Computer Processing of Chinese and Oriental Languages, Vol. 5, pp. 217-231, 1991.

[3] Chen, S. H., S. H. Hwang and C. Y. Tsai, "A First Study on Neural Net Based Generation of Prosodic and Spectral information for Mandarin text-to-speech", Int. Conf. ASSP, pp. 45-48, 1992.

[4] 吳宗憲、陳昭宏、莊欣中，「以 CELP 爲基處之文句翻語音中韻律訊息之產生與調整」，中華民國第八屆計算語言學研討會論文集，第233-251頁，1995.

[5] Atal, B. S. and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", J. Acoust. Soc. Am., pp. 637-655, 1971.

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

[7] Klatt, D. H., "Software for a Cascade/Parallel Formant Synthesizer", J. Acoust. Soc. Am., pp. 971-995, 1980.

[8] Holmes, J., "Formant Synthesizers - Cascade or Parallel ?", Speech Communication, pp. 251-273, 1983.

[9] Charpentier, F. and M. Stella, "Diphone Synthesis Using an Overlap-Add Technique for Speech Waveform Concatenation", Proc. Int. Conf. ASSP, pp. 2015-2018, 1986.

[10] Hamon, C., E. Moulines and F. Charpentier, "A Diphone Synthesis System Based on Time-Domain Prosodic Modification of Speech", Proc. Int. Conf. ASSP, pp. 238-241, 1986.

[11] Modulines, E. and F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones", Speech Communication, pp. 453-467, 1990.

[12] Galanes, F. M., M. H. Savoji and J. M. Pardo, "New Algorithm for Spectral Smoothing and Envelop Modification for LP-PSOLA Synthesis", Proc. Int. Conf. ASSP, pp. I-573-576, 1994.

[13] O'Shaughnessy, D., Speech Communication: Human and Machine, Addison-Wesley, 1987.

[14] Rabiner, L. and B. H. Juang, Foundmentals of Speech Recognition, Prentice-Hall, 1993.

[15] Hyperception, Hypersignal Users Manual, 1991.