基於 ANN 之頻譜演進模型及其於國語語音合成之應用

An ANN based Spectrum-progression Model and Its Application to Mandarin Speech Synthesis

古鴻炎　　吳昌益
Hung-Yan Gu　and　Chang-Yi Wu

國立台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@mail.ntust.edu.tw

摘要

考量合成語音的流暢性不佳的問題，本文提出以動態時間校正(DTW)來匹配目標(句子發音)音節與參考(單獨發音)音節之間的頻演(頻譜演進)路徑，再將頻演路徑轉換成固定維度的頻演參數，用以去訓練頻演參數類神經網路(ANN)模型。之後，將文句分析、頻演參數、韻律參數、和信號合成模組的程式作整合，而成爲可實際運轉的系統。當把此系統合成出的語音，拿去作聽測評估，所得到的平均分數顯示，頻演參數 ANN 模型的確可明顯地改進合成語音的流暢性。

關鍵詞: 頻譜演進, 流暢性, ANN, DTW, 語音合成
Keywords: spectrum progression, fluency, ANN, DTW, speech synthesis

一、前言

由前人的研究成果可知，要合成出自然、流暢的國語語音，韻律(prosody)參數的塑模(modeling)及數值產生扮演重要的角色[1,2,3]。一般被歸屬爲韻律參數的語音特性，包括：音節的基週軌跡(pitch-contour)、時長(duration)、音強(amplitude)、及音節前停頓(pause)等。我們依據過去的研究經驗發現，當採取 model based 的研究方向時，也就是韻律參數產生和信號波形合成分開處理的作法，就算是我們的韻律模型已經可以產生出相當自然的韻律參數值，但是合成出的語音信號，聽起來就是不像人講的那麼順暢。所以會這樣地具有不錯的自然度(naturalness)而欠缺流暢度(fluency)，我們先前檢討時，認爲是因爲相鄰的合成單元(音節)串接時，邊界上的共振峰軌跡(formant trace)沒有平順轉移所造成，因此我們便研究了一種解決共振峰軌跡平順轉移問題的作法[4]。使用此作法後，由聆聽合成的語音發現，流暢性是可以獲得一些改進，但是距離人講話的流暢性，仍然存在著明顯的差距。

最近回顧一些文獻後發現，我們所關心的流暢性不足的問題，其實已經有其他研究者注意到了[5,6,7]，他們提出的一種作法是，以 HMM(hidden Markov Model)模型的數個狀態，來切割一個音節的時長成爲數個時間片斷，再分別去掌握各片段上的頻譜特性(例如頻譜包絡, spectrum envelope, 的形狀)，並且以特定的狀態駐留(state staying)機率分佈來掌握在各個狀態上所應停留的時間長度。這樣的作法，以我們的觀點來看，就是在於作更細緻的規劃，把一個音節的時長以某一種非均勻(或非線性)的方法作切割，而讓不同的狀態分配到不等的時間長度，造成不同的頻譜包絡形狀會佔據不同長度的時長，以便更細緻地模仿真人發音(articulation)時的頻譜隨著時間變化的關係。

前述頻譜(包絡形狀)隨著時間演變的關係，在本文裡簡稱之爲頻譜演進(spectrum progression)，而頻譜演進路徑(簡稱爲頻演路徑)指的是，當把欲合成的音節放在橫軸上，而把相同拼音的原始錄音音節放在縱軸上，此時橫軸上各時間點所應對應的縱軸時間點，需要一條曲線來描述此對映(mapping)關係，一個例子如圖 1 所示，這樣的對映曲線就是本文所謂的頻演路徑。過去很多的國語語音合成系統，

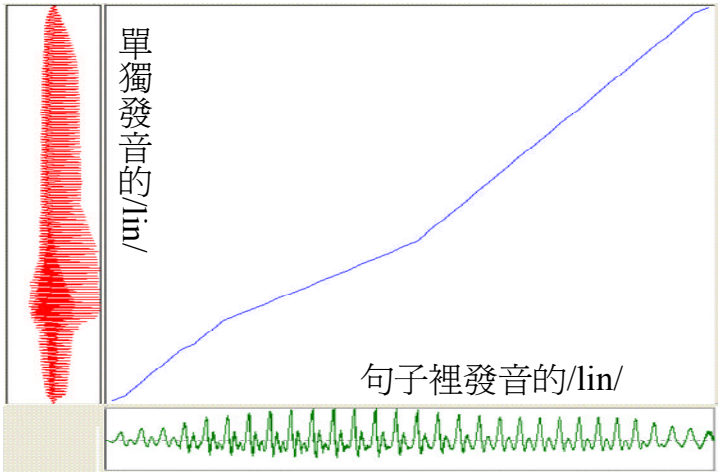


圖 1 頻譜對映曲線之例子

其合成出的語音的流暢性不佳的一個主要原因，我們認爲是因爲它們直接把頻演路徑設定爲直線，而沒有特別考慮頻演路徑的塑模(modeling)，再據以產生出逼近真人講話方式的頻演路徑。因此，我們便開始研究頻演路徑塑模及產生的問題，在此我們不追隨前人採取 HMM 來建立頻演路徑的模型，原因是 HMM 未去掌握時間上相鄰的觀測(observation)向量之間的依存(dependency)性，這相當於假設時刻 t 的觀測向量 O_t 和 O_{t+1} (或 O_{t-1})之間沒有依存關係，而只有去掌握 O_t 和它所停留的狀態之間的關連性，這樣的 modeling 方式令我們懷疑其是否可以滿足語音合成上的需求；此外，一個合成音節的頻演路徑並不會是只有固定的一條而已，而是會隨著左右鄰接

音節的不同，去行走不同的路徑(也就是 context dependent)，在此情形下，一個 HMM 的各個狀態如果只是各自去考慮 state duration 的機率分佈，而沒有考慮鄰接狀態和鄰接音節之間的相關性，則不免讓我們懷疑其完善性。

基於前述的考量，我們逐決定以 ANN (artificial neural network)來建立頻演路徑的模型，而模型的訓練步驟是: (a)逐一將整句發音裡的音節信號放在橫軸，而把相同拼音的單獨發音音節信號放在縱軸，再以 DTW(dynamic time warping)來匹配出一條頻演路徑；(b)將橫、縱軸上的音節信號的時間範圍各自正規化成 0 至 1 之間，然後在橫軸音節上均勻放 32 個正規化的時間點，各點再依頻演路徑對映至縱軸而得到介於 0~1 之間且隨著橫軸作非線性漸增的 32 的數值；(c)將各個句子發音裡的音節對映出的 32 個正規化的時間值(在本文裡稱爲頻演參數)作爲 ANN 模型學習的目標，並且把該音節及其前、後鄰接音節的資訊(也就是語境資料)作爲 ANN 的輸入資料，去訓練頻演參數的 ANN 模型。

得到頻演路徑參數的 ANN 模型後，就可將此模型和韻律模型、文句分析模組、及信號合成模組整合成一個文句翻語音系統，其結構如圖 2 所示。由圖 2 可知我們採取的是 model based 而非 corpus based 的研究方向，並且本文的焦點是在頻演參數模型的建構，而圖 2 裡其它的模組都是直接使用先前研究的成果[8,9]。我們傾向於不採取 corpus based 的研究方向，其考量是，corpus 的錄音、整理(標音、切音)需花費很大的人力、金錢(如購買 corpus)，並且 corpus 若不夠大，依然會發生音節之間基週軌跡銜接得不順暢，並且一個句子裡的音節時長會有太長及太短者，而造成發音速度忽快忽慢的不流暢情形。

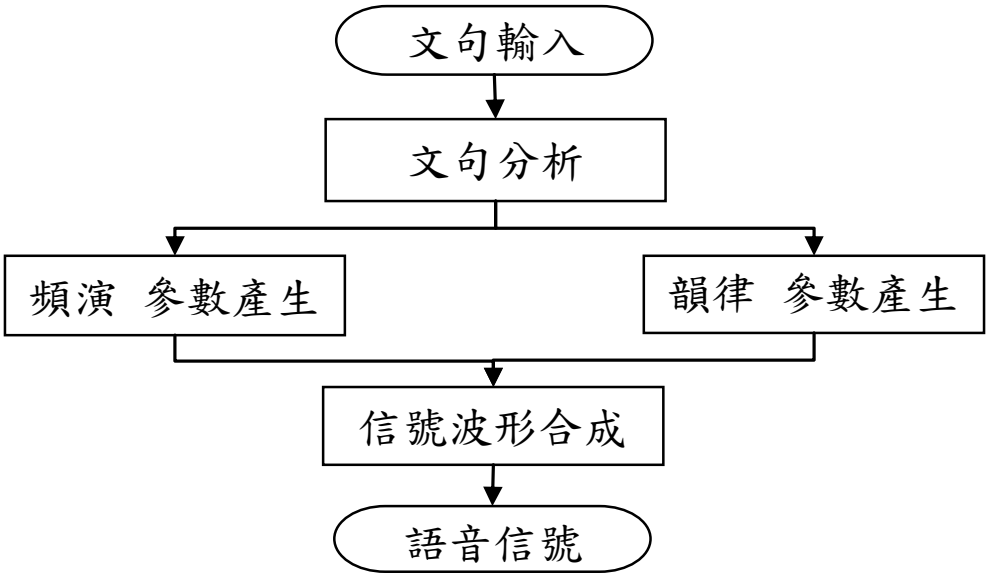


圖 2 整合頻譜演進之文句翻語音系統

二、頻演模型建造

2.1 訓練語料

我們所使用的訓練語料是由一位女性發音，先錄了 409 個單獨發音的國語基本音節，另外再錄 375 個句子的發音，總共 2,926 個音節，取樣率是 22,050 赫茲(Hz)。接著以音訊處理軟體 Wavesurfer 對語音檔進行標音(labeling)，在時間軸上標示出各個音節的音標、聲調和邊界點。標音後，我們寫了一個程式依據標籤檔來取出各音節的資訊，並將句子發音裡的各個音節切割成爲分別的音節檔案。

2.2 譜演路徑求取

動態時間校正(DTW) 是一種傳統的語音辨識方法[10]，尤其是在語者相關的語音辨識方面。DTW 的功用就是，它可以快速地找出參考音和測試音之間的一條具有最短距離的匹配路徑，當使用以頻譜差異爲依據的距離量測時，就可用 DTW 來找出頻譜上最匹配的路徑。如果我們將前述的測試音換成句子發音裡的音節，而把參考音換成單獨發音的相同拼音音節，則用 DTW 找出的頻譜上最匹配的路徑，就是本文所謂的譜演路徑。

令 $X=X_1、X_2、\ldots、X_n$ 表示，測試音(句子裡的音節)切割成音框後再求取特徵向量而得到的特徵向量序列，而 $Y=Y_1、Y_2、\ldots、Y_m$ 表示，參考音(單獨發音音節)切割成音框後再求取特徵向量而得到的特徵向量序列，在此使用的特徵向量包含 13 維度的 MFCC 係數和 13 維度的相鄰音框係數差值[10, 11]。在使用 DTW 來對 X、Y 兩序列作頻譜匹配之前，必須先選擇適當的局部路徑限制(local constrain)，過去被提出使用的局部路徑限制至少包括如圖 3 所示的三種[10]，其中 α 和 β 限制並不適合用於作頻演路徑的 DTW 匹配，原因是它們允許行走水平方向，這會使得放在橫軸的句子發音音節，要依據頻演路徑對映至縱軸放的單獨發音音節時發生混淆，也就是發生多對一(多個橫軸時間點對映同一個縱軸時間點)的情況。因此，我們選擇 γ 局部限制，以避免發生混淆的情況。

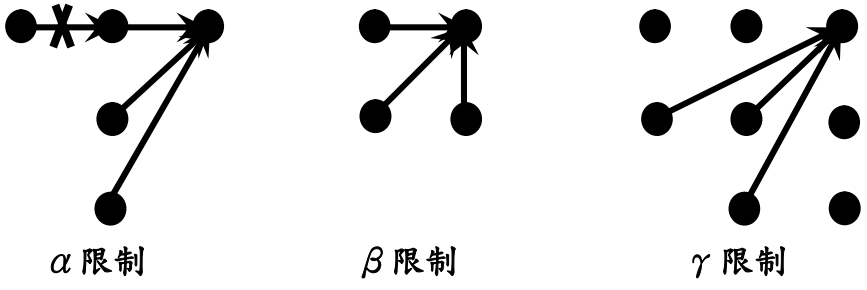


圖 3 DTW 之局部路徑限制

當採取圖 3 的 γ 局部限制時，DTW 的累積距離 Da(X,Y)的遞迴計算方式，就如公式(1)，

$$D_a\left(X_i,Y_j\right)=\min\begin{Bmatrix}D_a\left(X_{i-1},Y_{j-2}\right)+3\cdot D\left(X_i,Y_j\right)\\D_a\left(X_{i-1},Y_{j-1}\right)+2\cdot D\left(X_i,Y_j\right)\\D_a\left(X_{i-2},Y_{j-1}\right)+3\cdot D\left(X_i,Y_j\right)\end{Bmatrix} \quad (1)$$

其中，$D(X_i,Y_j)$表示以幾何距離量測特徵向量 X_i 和 Y_j 的距離，常數 3 和 2 則是我們爲了消除路徑偏好而設定的局部路徑權重值。

實際製作 DTW 程式後，進行初步測試時我們發現，如果作頻譜匹配的音節是含有無聲(unvoiced)聲母的(如/s,h,p/)，則時常會發生一個現象，就是某一軸(橫、縱軸)的聲母結尾部分會對映到另一軸的韻母起始部分，也就是匹配出的路徑不會如預期的聲、韻母的邊界剛好相互對應。因此，對於以無聲聲母開頭的音節，我們先作基週偵測[12]以找出無、有聲之邊界點，不過基週偵測也不保證會 100%正確，所以所發展的程式介面上，允許使用者去作邊界點的調整，有了邊界點之後，再將音節分割成兩段，分別去作 DTW 頻譜匹配。一個例子如圖 4 所示，橫軸放的是目標(句子發音)音節/siang/的波形，縱軸放的是參考(單獨發音)音節/siang/的波形，橫軸格線表示目標音節的音框，縱軸格線則表示參考音節的音框。進行 DTW 頻譜匹配時，分別對區域 A 作無週期性聲母的頻譜匹配，及對區域 B 作週期性部分的頻譜匹配。

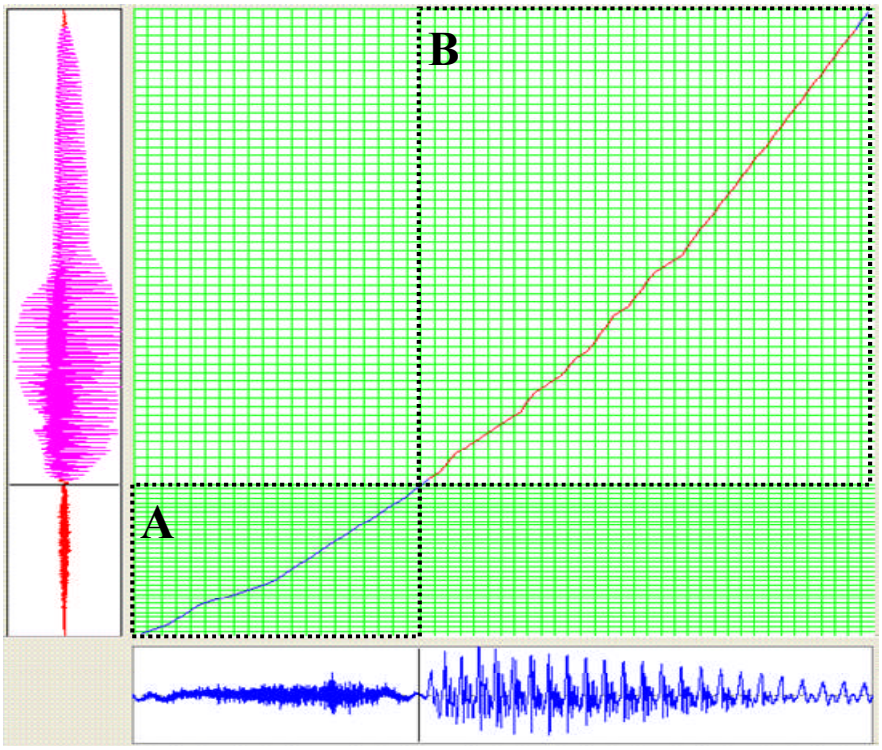


圖 4 兩段式 DTW 頻譜匹配

在音框位移(shift)和音框數量方面，原先內定的音框大小是 20ms，而音框位移是 5ms。不過我們採用的局部路徑限制的先天限制，必須將兩個音節的音框數量比例限制在 0.5 ~ 2 之間，以確保能夠滿足起點對起點、終點對終點的要求。因此當參考音音框數與目標音音框數比例超過 1.5 倍之門檻時，我們就將音框數較多的音節的音框位移作調整，也就是把音框位移乘上一個倍率，使導得出的音框數落在限制的範圍內，作法如公式(2)，

$$F_a = S_a \times \frac{2 \cdot N_a}{3 \cdot N_b} \tag{2}$$

公式(2)裡，Na 爲音框數較多之音節的原始音框數，Nb 爲音框數相對較少音節的原始音框數，Sa 爲音框數較多音節的原始音框位移，Fa 則爲調整後的音框位移。

2.3 ANN 頻演參數模型

本文採取的類神經網路結構如圖 5 所示，輸入層用以輸入 8 種語境資料，使用一層的隱藏層和一層的遞迴隱藏層，輸出層則有 32 個節點，用以輸出 32 個頻演參數。關於 ANN 權重值的訓練，採用的是最陡坡降學習法，此外遞迴隱藏層的權重值也是經由學習來決定，使用的是遞迴學習演算法。

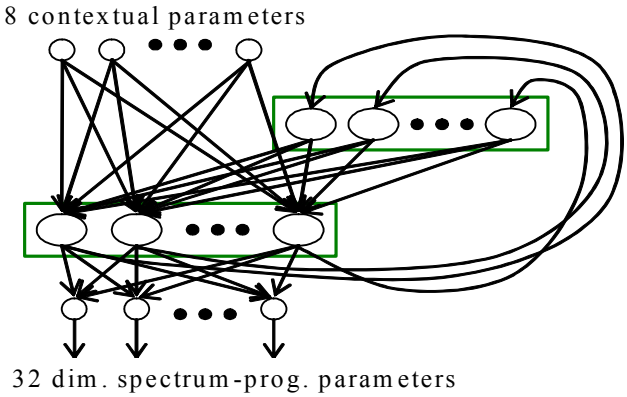


圖 5 頻演參數 ANN 之結構

輸入給 ANN 的語境資料，本研究使用“音節”作爲分析單位，由於語音是時序性資訊的傳遞，所以除了本音節的聲調種類、聲、韻母類別以外，也要考量到前一個音節的聲調和韻母類別，以及後一個音節的聲調和聲母類別。此外，考量到音節在句中的位置（例如句首、句中以及句末）也會對音節的韻律狀態產生影響，因此我們也使用一個句子時間比例之數值，來代表本音節在整句話中的時間位置。本研究所用到的 8 種語境資料，共需要以 27 個 bits 及一個浮點數來表示，詳細的配置情形如表 1 所列。

表 1 ANN 輸入之語境資料表示

項目	前音節聲調	前音節韻母類別	本音節聲調	本音節聲母	本音節韻母	句中位置	後音節聲調	後音節聲母類別
bits 數	3	4	3	5	6	浮點數	3	3

由於一個國語音節有 5 種聲調，因此聲調都以 3bits 表示。對於本音節的聲、韻母，由於國語有 22 種聲母和 39 種韻母，因此分別以 5bits 和 6bits 來表示。在前音節的韻母與後音節的聲母方面，我們考量到所準備的訓練語料較少，可能會因分類太多而造成 ANN 模型訓練語料嚴重不足，因此我們根據音節發音上的特性，將國語音節的聲母粗分爲 6 類，而韻母則粗分爲 9 類，詳細的分類方式如表 2 和表 3。依據粗分類類數 6 及 9，所以在表 1 中的前音節韻母和後音節聲母，分別使用 4 和 3bits。

表 2 國語聲母之粗分類

類別	聲母	類別	聲母
1	空聲母、ㄇ、ㄋ、ㄌ、ㄖ	4	ㄐ、ㄓ、ㄗ
2	ㄒ、ㄕ、ㄏ、ㄈ、ㄙ	5	ㄆ、ㄊ、ㄎ
3	ㄅ、ㄉ、ㄍ	6	ㄑ、ㄔ、ㄘ

表 3 國語韻母之粗分類

類別	韻母	類別	韻母
1	空韻母	6	ㄞ、ㄧㄞ、ㄨㄞ、ㄧ、ㄟ、ㄨㄟ
2	ㄚ、ㄧㄚ、ㄨㄚ	7	ㄠ、ㄧㄠ、ㄨ、ㄩ、ㄡ、ㄧㄡ
3	ㄛ、ㄧㄛ、ㄨㄛ	8	ㄢ、ㄧㄢ、ㄨㄢ、ㄩㄢ、ㄣ、ㄧㄣ、ㄨㄣ、ㄩㄣ
4	ㄜ、ㄦ	9	ㄤ、ㄧㄤ、ㄨㄤ、ㄥ、ㄧㄥ、ㄨㄥ、ㄩㄥ
5	ㄝ、ㄧㄝ、ㄩㄝ		

關於隱藏層節點數的設定，我們分別實驗了 14, 16, 18, 20 等四種數值，ANN 模型訓練的誤差值，量測後的結果如表 4 所示，其中 RMS 誤差表示 2,926 個音節的均方根誤差值的平均值，STD 誤差表示音節均方根誤差值的標準差，而 MAX 誤差表示最大的音節均方根誤差值。依據表 4 裡的誤差數值，我們最後選擇設定隱藏層的節點數爲 16。

表 4 不同節點數之 ANN 訓練誤差值

隱藏層 節點數	RMS 誤差	STD 誤差	MAX 誤差
14	0.05063	0.02443	0.18564
16	0.04906	0.02330	0.18038
18	0.04891	0.02343	0.18227
20	0.04946	0.02405	0.20055

三、系統製作

依據圖 2 所示之系統結構圖，當輸入一個中文文句後，首先會進行文句分析的處理，它經由長詞優先之查詞典過程，把文句中各個字的國語音節拼音查出，再去作三聲變調和”一”、”不”變調的處理。得到各個字的音節拼音、聲調、聲韻母、及詞邊界等資料後，接著就進行頻演參數的產生，及分別進行各項韻律參數數值的產生。由於頻演參數之 ANN 模型，已在第二節裡說明其建造過程，所以在合成階段就可以分別送各個字的語境資料給 ANN 模型，來得到各個字的頻演參數。至於韻律參數的產生和信號波形的合成，則分別在 3.1 和 3.2 節作說明。

在此值得一提的是，依據頻演模型產生出的頻演參數，我們可用以估計一個以無聲聲母(如/s/)開頭的欲合成的音節內，該無聲聲母所應分配的音節時長之比例。方法是，先將頻演參數作片段線性(piece-wise linear)內差，來形成如圖 4 裡所示的對映函數，然後以縱軸所放的參考音節的聲、韻母邊界點作爲參考點，再經由對映函數來找出橫軸上的對應點，而此點之正規化時間值，就是聲母的時長分配之比例。

3.1 韻律參數產生

韻律參數中的時長、音強、和基週軌跡等參數，我們同樣是使用如圖 5 所示的 ANN 模型結構，來對這三項韻律參數分別作訓練，也就是這三項韻律參數各自有一個獨立的 ANN 模型。我們所以必須對各項韻律參數和頻演參數分開作訓練，主要是因爲訓練語料的數量不夠多，如第 2.1 節裡所說的只有 375 句的 2,926 個音節。雖然訓練語料不是很足夠，但是經由適當地對語境資料作分類，如表 2 和表 3 之粗分類，所訓練出的模型仍可表現出不錯的效能，這可由實際的聽覺試聽來作驗證，我們爲此建立的一個網頁在 http://guhy.csie.ntust.edu.tw/spmdtw/。

另外可以一提的是，在本研究裡我們只使用 2.1 節所說的語料，去訓練時長和

音強的模型，而基週軌跡的 ANN 模型，則是直接使用先前研究所建造的模型[8]，也就是基週軌跡模型的建造，使用的是另一位男性所錄製的語料[13]，而語料數量是一樣的。所以會使用不同人的語料來訓練不同的韻律參數，其原因是我們想嘗試，當結合不同人的說話方式時，合成出的語音會有什麼不一樣的地方? 不過，作過實驗後並沒有感覺到什麼特殊的地方，初步的判斷是，各項韻律模型應可以使用不同來源的語料來作獨立的訓練。

3.2 信號波形合成

關於信號波形的合成，我們選擇採取 HNM (harmonic-plus-noise model)爲基礎的合成方法[14]，其原因是，一般熟知的 PSOLA 合成法，它所合成出的語音信號品質並不穩定，尤其是當音高(pitch)或時長(duration)作較大幅度的改變時。這裡所指的信號品質是，愈少迴音(reverberant)和愈清晰，則品質愈好。爲了能夠在大幅度改變韻律特性的情況下，仍然能夠合成出高信號品質的語音，我們在二年前就開始研究以 HNM 爲基礎的國語語音合成方法[15]，並且對它作了一些改進[9]。

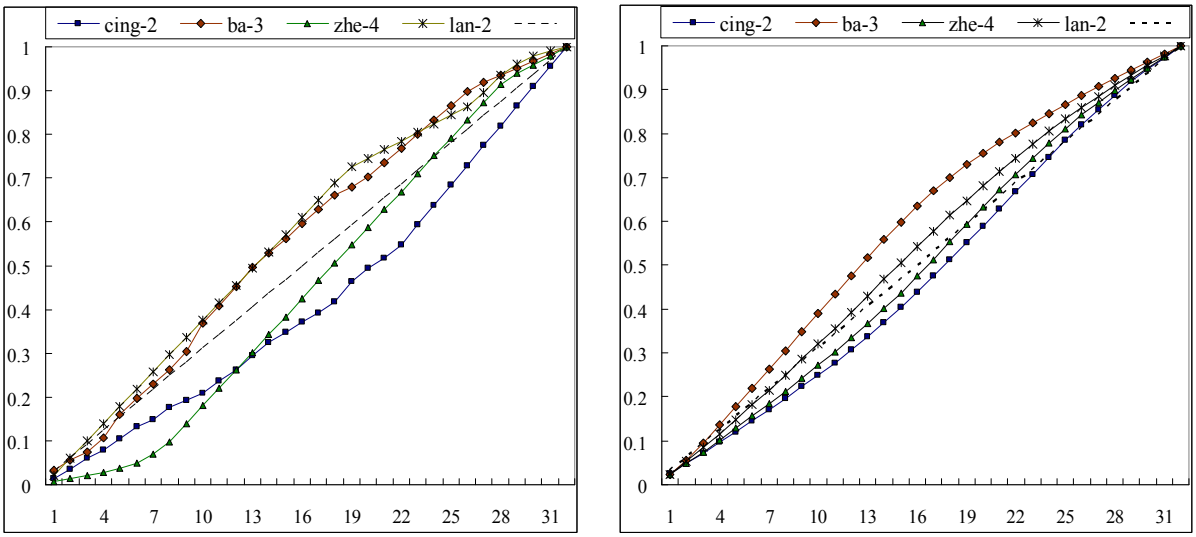
在參數分析階段，我們先對 2.1 節提到的單獨發音音節作 HNM 分析，以求得國語各音節各自的 HNM 參數，包括各個音框內表示雜音頻譜包絡的 cepstrum 係數、和各個諧波的頻率、振幅、相位參數。由於各個國語音節都只錄、存一次原始信號，所以在此只有一份 HNM 參數可供使用，也就是不能作單元選擇(unit selection)。在 HNM 的信號合成階段，首先要依據欲合成音節的時長數值(由韻律單元產生)，來決定在合成音的時間軸上要佈放多少個控制點(control point)。然後，對於各個控制點，要依據它所在的時間位置，去決定它上面的 HNM 參數數值，一個簡單作法是先以線性的時間軸對映方式，來找出原始音節上的對應的音框，再將該音框的 HNM 參數複製到該控制點上。由於本研究的重點是，探討頻演參數模型對於合成語音的流暢性的影響，系統製作上就是要去控制合成音和原始音時間軸的對映(mapping)關係，所以，我們要先把頻演參數模型產生的 32 個正規化時間的頻演參數，作片段線性內差以形成一個對映函數，然後依據這個對映函數，就可找出一個控制點所應對應的原始音上的兩個相鄰音框，再將此二音框的 HNM 參數作線性內差，並且複製到該控制點上。

當各個控制點上都有 HNM 參數之後，下一步還需考慮如何調整各控制點上的 HNM 參數數值，以使合成音節的音調能夠符合基週軌跡參數的規定。對於 HNM 合成法來說，音調高低的更改，不是只改動諧波頻率的數值就好了，因爲會牽動到音色(timbre)，發生音色隨著音高(pitch)在變動的不穩定現象。要在維持音色一致性的條件下，作音節音調的更改，其詳細方法可參考我們先前的研究成果[9]。

四、頻演路徑產生及聽測實驗

4.1 頻演路徑產生

我們可以一個訓練用的文句爲例，如”請把這藍兔子送走”，把它帶入第 2 節所建造的頻演模型，來產生出頻演路徑，然後和該訓練語句原始的頻演路徑作比較，以觀察兩者之間的異同。圖 6(a)畫的是/cing2 ba3 zhe4 lan2/四個音節原始錄音的頻演路



(a) 原始音節的頻演路徑　　(b) 合成音節的頻演路徑

圖 6 頻演路徑比較

徑，而圖 6(b)畫的是相同四音節以模型所產生出的頻演路徑，圖形中橫軸表示正規化的時間點，而縱軸表示頻演參數數值。比較圖 6(a)與圖 6(b)可發現，它們的相同點是，對應音節的頻演路徑具有相同的走向趨勢，例如圖 6(a)裡/cing2/的頻演路徑行走中間線(左下角至右上角之直線)的下方，而/ba3/則行走中間線的上方，這樣的現象也可在圖 6(b)裡看到。至於不同點方面，圖 6(a)裡原始音節的頻演路徑，其斜率變化較大，而圖 6(b)裡的斜率變化較小，因此會感覺較平順；此外，圖 6(a)裡的頻演路徑會偏離中間線較遠，而圖 6(b)裡的頻演路徑則偏離得較少。所以一般來說，ANN 頻演模型所產生出的頻演路徑，可以保有路徑的走向趨勢，但是路徑有向中間線靠攏的現象。

4.2 聽測實驗

合成語音的一種評估方法是主觀的聽覺測試。在此我們選擇了一篇短文，然後令圖 2 的”頻演參數產生”方塊，先直接產生出線性比例之頻演參數，而以此種頻演參數

合成出的語音檔案，以 Va 表示；另外，再令"頻演參數產生"方塊，以 ANN 頻演模型來產生出頻演參數，再拿去作語音合成，而得到的語音檔案以 Vb 表示。以前述兩種方式產生出的語音檔案，我們也已經放在網頁上 http://guhy.csie.ntust.edu.tw/spmdtw/，而可讓有興趣者來作試聽和比較。

接著，我們將合成的語音檔案 Va 和 Vb，分別讓 9 位受測者來進行聽測評估。分數是以比較的方式來評定，由受測者就前後兩個播放的語音音檔(先播 Va 再播 Vb)，評出那一個比較順暢，在此"順暢"的定義是，整句話的多個音節聽起來，連接得很緊密而沒有顆顆粒粒(形容音節像是各自獨立地在發音)的感覺。評分的方式是給一個-2 到 2 之間的整數值，正值代表第二個播放的語音檔比第一個播放的好，1 和 2 代表程度上的差別，2 表示第二個音檔比第一個明顯的好，1 則是表示稍好一些，而-1 和-2 代表第二個播放的語音音檔比第一個播放的較差，-2 表示明顯的較差，-1 表示稍差一些，至於 0 則代表聽不出兩個語音音檔的差異。聽測實驗後，將受測者的評分作平均，結果得到了 1.33 之平均分數，這表示 Vb 的確比 Va 流暢，並且流暢度的提升是感覺得出來的，至於提升的程度還不算很大，我們推測有幾位受測者，可能還未把自然度和流暢度的定義區分出來。

五、結論

本文提出了一個以 DTW 加上 ANN 來建立頻演參數模型的方法。對於以 DTW 匹配目標音節和參考音節的最佳路徑時，無聲聲母開頭的音節，常常會發生兩音節的聲、韻母邊界點無法正確對齊的問題，這可經由音節分段和兩段式 DTW 的作法來解決。此外，兩個要作匹配的音節，若發生時間長度相差太多的問題，這可以調整 frame shift 長度的作法來解決。

當建立頻演參數模型之後，將它和文句分析、韻律參數產生、信號波形合成等模組作整合，用以合成出國語語音信號，再把合成出的語音拿去作聽測實驗，初步結果顯示，我們的頻演參數模型的確可用以提升合成語音的流暢度。所以，本文提出的頻演參數模型，可說是 HMM 之外的一種可行的頻譜演進模型，並且它不需要和信號波形合成模組所用的聲學特性參數(如 HNM 的諧波參數)之間有依附的關係存在，不過我們無意和 HMM 爲基礎的合成方法，去比較誰好誰壞。此外，使用本文的語音合成系統所獲得的語音流暢性，可以驗證音節內頻譜演進的掌握，會比相鄰音節之間共振峰軌跡連續性的掌握，能夠獲得超過許多的效益。

參考文獻

[1] Wu, C.-H. and J.-H. Chen, “Automatic Generation of Synthesis Units and Prosodic Information for Chinese Concatenative Synthesis”, *Speech Communication*, Vol. 35. pp. 219-237, 2001.

[2] Yu, M. S., N. H. Pan, and M. J. Wu, “A Statistical Model with Hierarchical Structure for Predicting Prosody in a Mandarin Text-to-Speech System”, *International Symposium on Chinese Spoken Language Processing* , Taipei, pp. 21-24, 2002.

[3] Chen, S. H., S. H. Hwang, and Y. R. Wang, “An RNN-based Prosodic Information Synthesizer for Mandarin Text-to-Speech”, *IEEE trans. Speech and Audio Processing*, Vol. 6, No.3, pp. 226-239, 1998.

[4] Gu, Hung-Yan and Kuo-Hsian Wang, "An Acoustic and Articulatory Knowledge Integrated Method for Improving Synthetic Mandarin Speech’s Fluency", *International Symposium on Chinese Spoken Language Processing*, Hong Kong, pp. 205-208, 2004.

[5] Qian, Y., F. Soong, Y. Chen, and M. Chu, “An HMM-Based Mandarin Chinese Text-to-Speech System”, *International Symposium on Chinese Spoken Language Processing, Singapore*, Vol. I, pp. 223-232, 2006.

[6] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Duration Modeling in HMM-based Speech Synthesis System", *International Conference on Spoken Language Processing*, Vol. 2, pp. 29–32, 1998.

[7] Yeh, Cheng-Yu, *A Study on Acoustic Module for Mandarin Text-to-Speech*, Ph.D. Dissertation, Graduate Institute of Mechanical and Electrical Engineering, National Taipei University of Technology, Taipei, Taiwan, 2006.

[8] Gu, Hung-Yan, Yan-Zuo Zhou and Huang-Liang Liau, ”A System Framework for Integrated Synthesis of Mandarin, Min-nan, and Hakka Speech”, *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 12, No. 4, pp. 371-390, 2007.

[9] 古鴻炎、周彥佐，「基於 HNM 之國語音節信號的合成方法」，*第十九屆自然語言與語音處理研討會* (ROCLING 2007)，台北，第 233-243 頁, 2007。

[10] Rabiner, L. and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, 1993.

[11] O`Shaughnessy, D., *Speech Communication: Human and Machine*, 2’nd ed., IEEE Press, 2000.

[12] 古鴻炎、張小芬、吳俊欣，「仿趙氏音高尺度之基週軌跡正規化方法及其應用」，*第十六屆自然語言與語音處理研討會* (ROCLING XVI)，台北，第 325-334 頁, 2004。

[13] Gu, Hung-Yan and Chung-Chieh Yang, "A Sentence-Pitch-Contour Generation Method Using VQ/HMM for Mandarin Text-to-speech", *International Symposium on Chinese Spoken Language Processing*, Beijing, pp. 125-128, 2000.

[14] Yannis Stylianou, *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*, Ph.D. Dissertation, Ecole Nationale Superieure des Telecommunications, Paris, France, 1996.

[15] 古鴻炎、廖皇量，「用於國語歌聲合成之諧波加噪音模型的改進研究」，*WOCMAT 國際電腦音樂與音訊技術研討會*，台北，session 2 (音訊處理 I)，2006。