閩南語語句基週軌跡產生：兩種模型之混合與比較

Min-Nan Sentence Pitch-contour Generation: Mixing and Comparison of Two Kinds of Models

古鴻炎　　黃維
Hung-Yan Gu　and　Wei Huang

國立台灣科技大學資訊工程系
Dept. CSIE, National Taiwan University of Science and Technology
e-mail: guhy@mail.ntust.edu.tw　http://guhy.csie.ntust.edu.tw

摘要

本文對兩種基週軌跡的產生模型，SPC-HMM 和 ANN，作改進與擴充然後用以建造閩南語的基週軌跡產生模型，希望以閩南語作為工作語言，而其它的語言(如國語、客家語)的基週軌跡，能夠以調適過的閩南語所訓練出模型，來直接產生。除了研究閩南語的 SPC-HMM 和 ANN 模型的建造，我們也對此二種模型作了內外部測試之比較，及嘗試作模型混合的研究，希望能藉以提升效能。對於所建造的閩南語基週軌跡 SPC-HMM、ANN、和混合模型，我們也作了聽測評估，初步的結果顯示自然度方面，混合模型的分數比 ANN 模型的好，而 SPC-HMM 模型的則居末，所以混合模型在小型訓練語料下，是一個不錯的選擇。

關鍵詞： 語音合成，基週軌跡，類神經網路，隱藏式馬可夫模型

1. 前言

語音合成的研究，最近有不少人採取 corpus based 的研究方向[1,2,3,4]。大家也許會認爲，只要 copus 夠大，就不需要作音調(基週軌跡)調整之信號處理，以保持合成單元的信號清晰度與自然度，亦即不作基週軌跡模型建造和參數值產生的動作。可是實務上，要錄音及製作夠大的 corpus，所需的人力與經費，並不是一般的研究者能夠負擔得起，即使以程式自動作標音(labeling)和斷音(segmenting)，也仍然需要作人工的校正。然而當 corpus 不夠大時，會遇到的一個和基週軌跡有關的問題是，合成出的句子的語調，缺乏人類說話時的自然下傾(declining)的表現[5]，這會讓聽者較難掌握說話的節奏，而另一個可能發生的情況是，在合成單元的邊界，音調未能平順銜接。此外，還存在的一個明顯問題是，合成語句的說話速度會忽快忽慢地不一致，原因應是來自不同原始語句的合成單元，發音的速度不匹配。基於前述的問題，corpus based 的語音合成作法，除非是 corpus 夠大且錄音者能夠一直維持說話的平穩，不然還是有需要建造模型來

產生基週軌跡，用以檢驗所合成的語句，其語調是否有下傾的表現，相鄰音節間的基週軌跡是否有不平順銜接的地方。

其實本文研究閩南語語句基週軌跡之產生，是從另一個方向來思考台灣地區的語音合成研究的問題，就是強勢語言的文句翻語音(text-to-speech, TTS)研究，匯聚了大部分的研究資源(人力、經費)，而當要對另一弱勢語言作 TTS 之研究時，仍然需要再投入大量的資源，這樣的情況對於 corpus based 之研究方向應是很明顯的。因此我們開始思考，如何讓一種語言的語音合成之研究成果，能夠很經濟地轉移給另一個語言使用，如此弱勢語言面臨的存續問題，就可獲得至少一些些的舒解。對於同樣是以音節爲組成單位的聲調語言來說，例如國語(北京話)、閩南話、客家話、廣東話...等，我們可先選擇其中一個作爲工作語言(working language)，來研究它的韻律參數產生模型，然後透過模型調適，以工作語言訓練的模型，來產生出標的語言(target language)的韻律參數。這樣的想法，先前我們已曾經以基週軌跡參數爲例，選擇閩南語作爲工作語言，來建造它的 SPC-HMM 模型[6]，然後調適此模型以產生出國語和客家語語句的基週軌跡，初步實驗結果顯示，我們的想法是可行的[7]。

台灣的三種主要語言：國語、閩南語、客家語，我們所以會選擇閩南語作爲工作語言，是因爲它的聲調數量(7 個)和基本音節數量(785 個)都是最多的，並且它的聲調類型可含蓋客家語的(四縣腔 6 個、海陸腔 7 個)，及國語裡的前四個，而國語裡的輕聲也可以閩南語的低入聲來近似。雖然先前我們曾研究建造閩南語基週軌跡的 SPC-HMM 模型，不過有不少研究成果指出類神經網路(ANN)模型具有不錯的效能[8, 9]，因此我們覺得對於工作語言的基週軌跡模型，再嘗試以 ANN 模型或以 SPC-HMM 和 ANN 之混合來提升效能，是值得去研究、探討的，如此對於標的語言的效能也將會獲得改進。所以，本論文先在相同訓練語料的情況下，比較個別的 SPC-HMM 和 ANN 模型的基週軌跡預測誤差，再嘗試以模型混合之作法來降低預測誤差。由於訓練語料的大小、來源不一樣，所以本文得到的基週軌跡效能，並未和他人的研究成果作比較。

2. 訓練語句錄音與基週軌跡之前處理

基週軌跡模型的訓練與測試語料，都是由本文第二作者在實驗室以麥克風發音直接錄音到電腦中，取樣率爲 22,050Hz，樣本值寬度 16bits，總共錄了 643 句(3,696 音節)閩南語句作爲訓練之用，65 句(437 音節)閩南語句作爲外部測試之

用，各句分別存成一個音檔。發音用的文句主要是取自閩南語歌曲的歌詞，部分則是自行造句，聲調大部分是採變過調的以方便唸讀錄音，在錄音之前我們已對文句作過分析篩選，以確保訓練語句裡前後三音節的聲調組合，能夠含蓋所有可能的聲調組合，實際情形是各種組合最少都出現三次，而出現次數在 3 至 5 次之間的有 194 種組合，在 6 至 10 次之間的有 126 種組合。

錄音後，接著進行基週頂點標記和音節邊界標記的動作，先以程式作自動偵測，再由人工更正錯誤的標記。由於錄得的各音節的時間長度不相同，因此我們先作時間正規化之處理，讓各個音節的基週軌跡都以相同的 16 維度(dimensions)的頻率向量 $<f_0, f_1, \ldots, f_{15}>$來表示，$f_k$ 表示在音節有聲部分時間比例 k/15 地方的基頻頻率值取對數，該頻率值可依基週頂點標記資料去內差得到[6]。

由於訓練及測試語句的錄音是分散在很多天裡，錄音者很難一直維持在同一種精神狀況與心情下來錄音，而使得錄到的語句有的音調較高有的音調較低，若不作音高正規化之處理，則模型產生出的句子基週軌跡會有高低起伏的不平順銜接的現象。因此作完時間正規化處理之後，接著還需進行音高正規化的處理。一種簡單且有效的正規化方法是[6]: (a)求各個音節的平均音高，亦即求取頻率向量的 16 個維度數值的平均；(b)求第 k 個語句的平均音高 HS_k，亦即求取該語句的組成音節的音高的平均值；(c)求總體語句的平均音高 HT，亦即將各個語句的音高加總再除以語句數量；(d)依據 $HD_k = HS_k - HT$，將第 k 個語句的各個組成音節的音高減去 HD_k，亦即頻率向量的各個維度都要減去 HD_k。

3. 句子基週軌跡 HMM 模型

由於整句話的行進(語調)對於音節基週軌跡的影響是不易精確描述的，因此先前我們對國語作了這樣的研究[10][6]，即以 HMM 裡的隱藏狀態來描述，音節基週軌跡在一句話行進中的不同時間位置所受的不同影響，稱爲 SPC-HMM(sentence-pitch-contour HMM)。其觀念是以 HMM 的三個隱藏狀態來對應一個句子(或呼吸群)內的”句首”、”句中”與”句尾”等三個隱含的韻律狀態，如圖 1 所示。至於 HMM 的觀測(observation)值，我們採用離散式的觀測符號，且令一個語句的各個音節分別產生出一個觀測符號。由於考慮時刻 t 時的音節基週軌跡至少會受到前、後及本音節聲調的影響，所以我們定義觀測符號爲，前一個音節聲調、本音節聲調、下一個音節聲調及本音節基週軌跡量化碼等四個因素

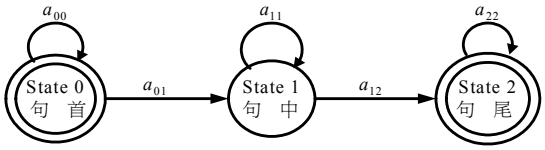

圖 1　韻律狀態轉移圖

之組合，即令時刻 *t* 時的觀測符號爲:

$$O_t = G_{t-1} \times G_t \times G_{t+1} \times V_t \tag{1}$$

G_t 表示句子中第 *t* 個音節的聲調, $0 \le G_t \le 6$

V_t 表示句子中第 *t* 個音節的基週軌跡量化碼, $0 \le V_t \le 7$

由於公式(1)中用到的V_t，是音節基週軌跡之向量量化碼，因此我們必須先訓練閩南語七個聲調各自的碼書(code book)，這裡採用了 K-means Clustering 演算法 [11]，向量量化之距離量測是均方根(root mean square)距離，其公式爲

$$d(X,Y) = \sqrt{\frac{1}{16}\sum_{k=1}^{16}(x_k - y_k)^2} \tag{2}$$

至於向量量化碼書大小的選擇，如果碼字越多的話，則量化誤差會比較小，但是對於後面的 HMM 的訓練則需要有更多的訓練語句來訓練，才能使 HMM 內部測試之誤差改進。所以，在有限的訓練語句之下，我們選擇將各個聲調的音節基週軌跡都量化成 8 類。

HMM 模型的訓練語句當然是愈多愈好，如果訓練語句不足，將使某一種音節聲調之組合沒有出現過，此時可用降階機制和資料分享來解決語料不足的問題。降階機制的第一層表示沒有降階，就是使用公式(1)來產生出觀測符號；第二層是，只考慮前一個音節的聲調、本音節的聲調和本音節的基週軌跡量化碼的組合。然後各層各自訓練，以得到該層所對應的 a_{ij}、$b_j(k)$等參數。解決訓練語料不足的方法，除了降階機制外，資料分享是另一方法，其主要觀念是將一個觀測符號的出現機率分享給距離量測上最近的另外幾個觀測符號，來提升未出現或出現機率較少的觀測符號的機率值。我們的作法是，一個音節組合出的觀測符號，它的出現機率分享給距離最近的另外兩個，相同聲調組合但是由不同基週軌跡量化碼字所形成的觀測符號，分享的比率爲 0.99, 0.005, 0.005。

以 SPC-HMM 模型來作一個語句的基週軌跡的合成，輸入的是聲調序列而沒有基週軌跡量化碼字的資料(這和訓練階段的情況不同)，因此當對於第 *t* 個音節的聲調，要以公式(1)來組合出觀測符號時，會有八種觀測符號可供選擇，此時必須將訓練階段的最佳路徑搜尋之維特比演算法作修正，將原本由時間軸和狀態軸構成的搜尋平面，擴充成時間軸、狀態軸和軌跡量化碼字軸之三度搜尋空間，再配合 HMM 模型的 a_{ij}、$b_j(k)$等參數，以找出最佳的三度搜尋空間裡的路徑，再作回朔(back tracking)，以確定各音節所選到基週軌跡量化碼字。實作上，可依上游文字分析所得的呼吸群、詞邊界訊息來直接設定各音節所停留的狀態，如此可產生出更自然的句子基週軌跡。

4. 句子基週軌跡 ANN 模型

本文依據前人研究 ANN 來產生國語語句基週軌跡的經驗[12]，作進一步的研究改進，以用來產生閩南語語句的基週軌跡。這裡所採用的是遞迴式類神經網路，其結構如圖 2 所示，分爲四層:輸入層、隱藏層、隱藏遞迴層、輸出層。輸入層有 28 個輸入單元，來輸入語境參數；輸出層有 16 個單元，以表示一個音節之基週軌跡(即 16 維度的頻率向量)；隱藏層的單元數，依實驗結果設定爲 30。內部鏈結的權重值需經由學習來決定數值，這裡使用的是遞迴學習演算法[13]。

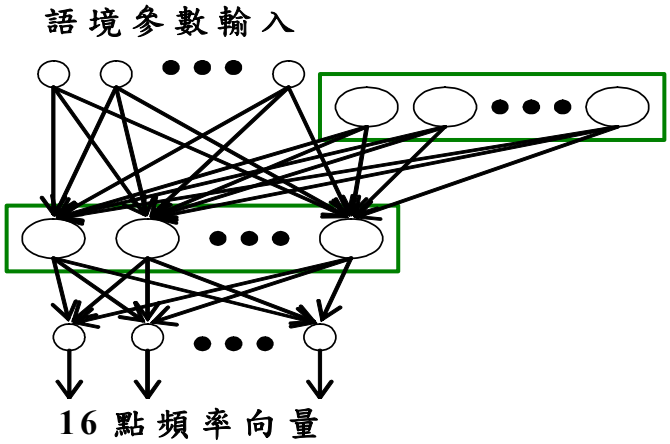

圖 2　遞迴式類神經網路之結構

關於 ANN 模型的輸入，本文使用以音節爲單位的語境參數：聲母、韻母、聲調、時間比例等參數。由於語音是時序性的資訊傳遞，所以除了要考慮本音節的聲調種類、聲母類型、韻母類型，也要考慮到前一個音節的聲調種類和韻母類型，以及後一個音節的聲調種類和聲母類型，此外使用了一個時間比例參數，以

代表本音節在整句話中之時間進度，此參數相當於傳入韻律狀態的資訊。上述的語境參數共使用了 28 個 bits，詳細情形如表 1 所示，表中從左起依序爲輸入層

表 1 ANN 輸入層的資料項目及 bit 數

項目	上個音節聲調	上個音節韻母類別	本音節聲調	本音節聲母	本音節韻母	$\frac{t}{N+1}$	下個音節聲調	下個音節聲母類別
bits	**3**	**4**	**3**	**5**	**6**	**1**	**3**	**3**

項目和其所佔的 bit 數。其中前一個音節的韻母類別佔 4 bits，因爲它被**粗**分類成 12 類，下一個音節的聲母類別佔 3 bits，因爲有 6 個聲母之**粗**分類類別，本音節聲母佔 5 bits，因爲有 18 個聲母之細分類，本音節韻母佔 6 bits，因爲有 60 個韻母之細分類。本音節及前後兩音節的聲調三個項目都各佔 3 bits，因爲閩南語有 7 個聲調。表中另一項目 $t/(N+1)$ 用以指示本音節在句子裡的位置，N 代表句子的總字數，這個項目用一個浮點數值表示，不像其他項目是用二進位數值表示。

聲母、韻母在細分類(原始分類)之外，這裡還另外作粗分類的原因是，聲母和韻母的種類太多，使得本音節和前後音節之聲母、韻母組合的數量過多，這意味我們必須準備目前人力難以達成的龐大訓練語料，否則不能讓所有聲母、韻母組合有足夠的出現次數，而必然會影響到 ANN 模型的效能[14]。所以我們採取前後音節的聲母、韻母先作粗分類再作組合的方式，以減少組合的數量，讓每個組合在訓練語料中有足夠的出現次數。粗分類僅用於前後音節之聲母、韻母，本音節之聲母、韻母仍採用細分類。關於閩南語聲母的粗分類，這裡依據前人研究國語粗分類的觀念[12, 14]，分類成如表 2 所示的方式，國語和閩南語都有的聲

表 2 閩南語聲母之粗分類(通用拼音符號)

類別	聲母
1	零聲母, m, n, l, r, ng, q, v
2	h, s
3	b, d, g
4	z
5	p, t, k
6	c

母放在相同的類別上，而閩南語才有的/ng, q, v/，因爲都是有聲的子音，所以將它們放在第一類上。關於閩南語韻母的粗分類，詳細的分類方式如表 3 所示，雖然基本上參考了國語韻母粗分類的觀念，但我們也依據發音口形作了一些修

表 3 閩南語韻母之粗分類(通用拼音符號)

類別	韻母
1	空韻母
2	-a, -ia, -ua
3	-o, -io
4	-er
5	-e, -ue
6	-ai, -uai, -au, -iau
7	-i, -u, -ui, -iu
8	-ing, -eng, -in, -un, -en
9	-ang, -iang, -uang, -ong, -iong, -an, -uan
10	-am, -iam, -im, -om
11	-ah, -eh, -ih, -oh, -uh, -auh, -erh, -iah, -ierh, -ioh, -uah, -ueh
12	-ap, -iap, -ip, -op, -at, -et, -it, -uat, -ut, -ak, -iak, -ik, -iok, -ok

改，例如把/-au, -iau/改放於第 6 類，把/-ing, -eng/改放於第 8 類，把/-an, -uan/改放於第 9 類。另外，第 11、12 類是國語所無的閩南語入聲韻母，第 10 類是國語所沒有的其它閩南語韻母。

5. 模型混合之方法

這裡的模型混合不是要把兩個模型合併成一個模型，再用以產生基週軌跡，而是將兩個模型各自產生的輸出作組合，以得到一個新的輸出。包括國語的漢語方言的電腦語音合成上，過去研究者提出的 HMM 和 ANN 模型，各有其優缺點，作模型混合就是希望能夠取長補短以提升合成語音之品質，例如兩種模型分別在基週軌跡的穩定性和活潑性上佔優勢，則混合此兩種模型有可能產生出兼具穩定性和活潑性的語音。

關於混合的方法，本文研究後只找到兩種比較有實際效果的方式，在此稱爲: 簡單加權方式和 16 維度標準差加權方式。簡單加權方式就是將 HMM 和 ANN 模型所產生的基週軌跡值，直接依照各種加權比例去組合出新的基週軌跡值；16 維度標準差加權方式，則是針對兩個模型產生的基週軌跡之 16 個維度來分別作加權，音節的各個維度分別當成一個集合來求出標準差，因此共有 16 個標準差，再利用兩個模型各自的 16 個標準差，作爲權重比例。簡單加權方式的公式和 16 維度標準差加權方式的公式分別如下所列:

$$F^S = W_H \cdot F^H + (1 - W_H) \cdot F^A \tag{3}$$

$$f_j^M = \frac{W_H \cdot \frac{\sigma_j^A}{\sigma_j^H} \cdot f_j^H + (1 - W_H) \cdot f_j^A}{W_H \cdot \frac{\sigma_j^A}{\sigma_j^H} + (1 - W_H)} \quad , \quad j=1, 2, \cdots, 16 \qquad (4)$$

其中 F^S 表示簡單加權方式混合後的基週軌跡頻率向量，W_H 表示 HMM 模型的權重，F^H 與 F^A 分別表示 HMM 和 ANN 模型產生出的頻率向量；f_j^M 表示 16 維度標準差加權方式混合後第 j 維度的頻率值，f_j^H 與 f_j^A 分別是 F^H 與 F^A 的第 j 維度的頻率值，σ_j^A 表示 ANN 模型產生的頻率向量第 j 維度之標準差，σ_j^H 表示 HMM 模型產生的頻率向量第 j 維度之標準差。公式 4 裡，我們依 σ_j^A 與 σ_j^H 的比率來個別調整各個維度中 ANN 與 HMM 兩模型的加權值 W_H 與 $1-W_H$。

爲了分析混合後模型的效能，在此我們先以模型之訓練語句作爲輸入，來計算兩模型輸出的頻率向量之間的相關係數，結果發現 16 個維度各別的相關係數 R_j 都在 0.95 左右(0.944 至 0.972)，對於如此高的相關性，我們大約已經知道，混合後模型的預測能力，應不會有很大幅度的改變。我們量測相關係數的公式如下:

$$R_j = \frac{\frac{1}{N}\sum_{k=1}^{N}(f_j^H(k) - g_j^H) \bullet (f_j^A(k) - g_j^A)}{\sqrt{\frac{1}{N}\sum_{k=1}^{N}(f_j^H(k) - g_j^H)^2} \bullet \sqrt{\frac{1}{N}\sum_{k=1}^{N}(f_j^A(k) - g_j^A)^2}} \quad , \; j = 1,2,...,16 \qquad (5)$$

其中 N 表示訓練語句裡的總音節個數，$f_j^{H(k)}$ 與 $f_j^{A(k)}$ 分別表示第 k 個音節的 f_j^H 與 f_j^A，而 g_j^H 與 g_j^A 分別表示 HMM 和 ANN 模型產生出的頻率向量的第 j 維度的平均值，即

$$g_j^H = \frac{1}{N}\sum_{k=1}^{N} f_j^H(k) \quad , \quad g_j^A = \frac{1}{N}\sum_{k=1}^{N} f_j^A(k) \quad , \quad j = 1,2,...,16 \qquad (6)$$

6. 模型效能之比較

依據公式(3)與(4)，接著我們進行模型混合之實驗，在內部測試(inside test)時，使用的測試語句就是模型訓練所用的 643 個語句，而在外部測試(outside test)時，則使用另外的 65 個未參加模型訓練的語句。一個音節的原始基週軌跡和模型預測該音節所輸出的基週軌跡，兩者之間以公式(2)來量測誤差距離。當採簡單加權方式時，我們得到如表 4 所示的誤差數值；而當採 16 維度標準差加權方式時，我們得到如表 5 所示的誤差數值。

表 4 簡單加權式模型混合之預測誤差

權重 W_H	Inside test			Out test		
	AVG	STD	MAX	AVG	STD	MAX
-0.1	0.0396	0.0193	0.1579	0.0545	0.0384	0.4299
0	0.0386	0.0188	0.1563	0.0530	0.0386	0.4324
0.1	0.0380	0.0185	0.1601	0.0528	0.0394	0.4350
0.2	0.0377	0.0185	0.1683	0.0539	0.0407	0.4376
0.25	0.0377	0.0186	0.1725	0.0549	0.0415	0.4390
0.3	0.0377	0.0188	0.1767	0.0561	0.0425	0.4403
0.4	0.0380	0.0195	0.1852	0.0592	0.0450	0.4431
0.6	0.0395	0.0217	0.2023	0.0674	0.0517	0.4488
0.8	0.0421	0.0248	0.2197	0.0776	0.0600	0.4548
1.0	0.0456	0.0282	0.2373	0.0891	0.0694	0.4826

表 5 十六維度標準差加權式模型混合之預測誤差

權重 W_H	Inside test			Out test		
	AVG	STD	MAX	AVG	STD	MAX
-0.1	0.0396	0.0193	0.1579	0.0543	0.0384	0.4301
0	0.0386	0.0188	0.1563	0.0530	0.0386	0.4324
0.1	0.0380	0.0185	0.1600	0.0528	0.0394	0.4348
0.2	0.0377	0.0184	0.1682	0.0537	0.0405	0.4373
0.25	0.0377	0.0186	0.1724	0.0546	0.0412	0.4386
0.3	0.0377	0.0188	0.1766	0.0557	0.0422	0.4399
0.4	0.0380	0.0194	0.1850	0.0586	0.0445	0.4426
0.6	0.0395	0.0217	0.2021	0.0666	0.0510	0.4483
0.8	0.0420	0.0247	0.2196	0.0770	0.0595	0.4544
1.0	0.0456	0.0282	0.2373	0.0891	0.0694	0.4826

在表 4 和 5 裡，AVG 表示所有參加測試的音節的平均預測誤差，STD 表示預測誤差的標準差，而 MAX 表示所有音節的預測誤差的最大值。首先比較 W_H 權重值爲 0 和爲 1 的兩列，W_H=0 代表只使用 ANN 模型，W_H=1 代表只使用 SPC-HMM 模型，由表 4 和 5 可看出，ANN 模型的基週軌跡預測誤差在 AVG 項分別是 0.0386 (內部測試)與 0.0530 (外部測試)，都比 SPC-HMM 模型的 0.0456 與 0.0891 來得小許多，0.0386 相當於 120Hz 音高時線性的 4.72Hz 的差異，而

0.0530 則相當於線性 6.53Hz 的差異；另外在 STD 和 MAX 項，ANN 模型的誤差也都是比 SPC-HMM 模型的好很多，所以個別的模型來說，ANN 模型的確比 SPC-HMM 模型具有比較準確的預測能力。

如果再考慮各種不同的 W_H 權重值來了解混合模型的效能，則由表 4 和表 5 可看出兩種加權方式都呈現相同的趨勢，以項目 AVG 來看，在內部測試方面，最好的權重值都是令 W_H = 0.25，而在外部測試方面，最好的權重值則是令 W_H = 0.1，以獲得較小的預測誤差平均值。在內部測試時，使用混合模型可以讓平均預測誤差，從 ANN 模型的 0.0386 降至 0.0377，下降幅度約爲 2.3%；在外部測試時，則可從 ANN 模型的 0.0530 降至 0.0528，下降幅度只有 0.4%。所以這裡所研究的混合方式，並不能夠大幅度提升效能，其原因之一應是，兩種模型所產生出的基週軌跡之間，已經具有非常強的相關性。

7. 聽測評估

由於本文的研究主題是基週軌跡之產生，所以在此初步的聽測評估裡，其它的韻律參數(如音長、音量)的產生，及信號波形的合成，都是直接沿用以前的成果[12,15,16]。音長、音量等韻律參數的產生，是採用簡單的規則式作法，而信號波形合成裡，每個合成單元(音節)只用一個固定的平調發音，合成方法則是 TIPW[16]，它可說是 PSOLA 的改進作法。基週軌跡的產生，分別使用了 ANN 模型、SPC-HMM 模型、及混合模型，混合方式採簡單加權式，權重值則設爲 W_H = 0.25，因爲它是內部測試裡 AVG 項最好的權值，在此不採外部測試的權值，因爲外部測試的句子數量不夠多，可信度我們仍存懷疑。對於模型輸出的 16 維度頻率向量，我們再於所關心的時間點附近，找出相鄰的 4 個維度的頻率值，作 Lagrange 內差，就可求得該時間點上的週期長度。至於文句分析的處理，我們可說是幾乎沒有作，因爲我們直接把如表 6 所示的拼音文句，輸入給所建造的語音合成系統，不過這些文句未參加模型之訓練。

參與聽測的測試者爲 15 人，其中 10 人爲各實驗室的研究生，年齡在 20~30 歲，另外 5 人是年齡在 30~50 歲的鄰居，我們以可辨度和自然度來衡量所合成出來的語音品質，可辨度是指測試者聽得懂合成語音的程度，自然度是指合成語音像人類語音的程度，這裡分成五個評分段供試聽者作爲評分標準：9.0 ~ 10(非常像人類語音)、8.0 ~ 8.9(很像人類語音)、 7.0 ~ 7.9(接近人類語音)、6.0 ~ 6.9(及

表6 聽測用的閩南語文句

	文句	通用拼音(聲調採傳統編號)
1	我有滿腹的心聲	qua-1 wu-3 <mua-1 bak-4> e-7 <sim-7 siann-1>.
2	有話想要對你講	<wu-3 we-7> <siunn-3 veh-8> <dui-2 li-1> gong-2.
3	無講誰人會知影	vo-7 gong-2 <siann-1 lang-5> e-3 <zai-7 yann-2>.
4	有啥麼代誌	wu-3 <siann-1 mi-1> <dai-3 zi-3>.
5	我攏總聽無	qua-2 <long-1 zong-1> <tiann-7 vo-5>.
6	請汝講卡大聲	<ciann-1 li-2> gong-1 kah-8 <dua-3 siann-1>.

格)、5.9 以下(劣等)。自然度聽測時，隨機播放三種基週軌跡模型合成出的語音檔，然後再選擇以自然度最高的語音檔作可辨度評估，試聽者每聽完一句就將該句覆述一遍，以便記錄聽錯的音節，聽完全部語句後就可計算可辨度，可辨度定義爲聽對的音節數佔總音節數的比率，最後依據 15 人的評分來算出平均值。結果我們得到如表 7 所示的數值，由此數值可知，ANN 模型的自然度 7.2 分比

表 7　合成語句的聽測評分

	混合模型	ANN模型	SPC-HMM模型
自然度	**7.4**	**7.2**	**6.9**
可辨度	**96.0%**		

SPC-HMM 模型的 6.9 分好一些，而混合模型的自然度 7.4 分又比 ANN 模型的好一些，這剛好和表 4 和 5 裡的預測誤差數值，呈現一致的走勢。另外在可辨度方面，96%可以被聽對，我們覺得是不錯的，因爲通常短的語句比較難聽得懂。

不過，SPC-HMM 模型的基週軌跡預測誤差，比另二者大許多，而混合模型的預測誤差僅比 ANN 模型的改進一些些，但是在聽測上，SPC-HMM 模型的 6.9 分也不會比 ANN 模型的 7.2 分差很多，而混合模型的 7.4 分，則也明顯的有改進。這樣的現象，我們認爲是因爲，SPC-HMM 模型和 ANN 模型的輸出之間，基本上具有非常強的相關性，而 SPC-HMM 模型的預測誤差，具有比 ANN 模型大許多的直流成分(電學觀念)，這應是肇因於向量量化處理，實際上我們計算基週軌跡預測誤差在各維度上的平均值得知，SPC-HMM 模型的誤差平均各維度都約在 0.013 左右，而 ANN 模型的誤差平均，各維度則大多在絕對值 0.002 以下。另外，ANN 模型偶而會有偏移量大許多的誤差值出現(比較不穩定)，而 SPC-HMM 模型則比較無此種情況(較穩定)，所以混合模型在聽測上，會表現得有一些改進。

8. 結論

本文將過去研究國語基週軌跡產生的 SPC-HMM 模型和 ANN 模型作更改與擴充，以用來訓練及建立閩南語的基週軌跡模型，考慮了更多的閩南語聲調，及聲母、韻母的粗分類問題。此外，我們也比較了兩種模型的效能，及嘗試作模型的混合，希望能夠藉以提升效能，所嘗試的兩種混合方式是，簡單加權方式、和十六維度標準差加權方式。經由內、外部的測試實驗後，結果顯示混合模型與 ANN 模型的效能都比 SPC-HMM 模型的好很多，並且混合模型的又比 ANN 模型的好一些。

此外，我們也把模型產生出的基週軌跡拿去合成出語音信號，再作主觀的聽測評估，結果在自然度方面，混合模型得到 7.4 分，比 ANN 模型的 7.2 分好一些，而 SPC-HMM 模型也可得到 6.9 分，雖然說 SPC-HMM 模型的基週軌跡預測誤差，在內、外部測試裡都比另二者差很多。在可辨度方面，結果顯示 96% 的字可被聽對。由於本文的研究裡，訓練的語句共只有 3,696 個音節，算是小型語料的情況，所以在小型語料的情況下，混合模型可說是建造基週軌跡模型的不錯選擇。

9. 致謝

感謝國科會計畫的支援，計畫編號 NSC 92-2213-E-011-078.

參考文獻

[1] Sagisaka, Y., *et al.*, "ATR v-talk Speech Synthesis System", ICSLP'92, Canada, pp. 483-486, 1992.

[2] Chou, Fu-chiang, Corpus-based Technologies for Chinese text-to-Speech Synthesis, Ph. D. Dissertation, Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, 1999.

[3] Min Chu, *et al.*, "Microsoft Mulan - a bilingual TTS system", ICASSP '03, Vol. 1, pp. I264-I267, 2003.

[4] 張唐瑜，以大量詞彙作爲合成單元的中文文轉音系統，碩士論文，國立中興大學資科所，2005。

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

[6] Gu, H. Y. and C. C. Yang, "A Sentence-Pitch-Contour Generation Method Using VQ/HMM for Mandarin Text-to-speech", ISCSLP’2000, Beijing, pp. 125-128, 2000.

[7] Gu, H. Y. and H. C. Tsai, "A Pitch-Contour Model Adaptation Method for Integrated Synthesis of Mandarin, Min-Nan, and Hakka Speech", 9’th IEEE Int. Workshop on Cellular Neutral Networks and their Applications (Hsinchu, Taiwan), pp. 190-193, 2005.

[8] 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.

[9] Lin, C. T., R. C. Wu, J. Y. Chang, and S. F. Liang, “A Novel Prosodic-Information Synthesizer Based on Recurrent Fuzzy Neural Network for the Chinese TTS System”, IEEE trans. Systems, Man, and Cybernetics, Vol. 34, No. 1, pp. 309-324, Feb. 2004.

[10] 楊仲捷，基於 VQ/HMM 之國語語音合成基週軌跡產生之研究，碩士論文，國立台灣科技大學電機所，1999。

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

[12] 曹亦岑，使用小型語料類神經網路之國語語音合成韻律參數產生，碩士論文，國立台灣科技大學電機所，2003。

[13] Lee, S. J., K. C. Kim, H. Y. Jung, and W. Cho, "Application of Fully Recurrent Neural Networks for Speech Recognition", ICASSP’91, pp. 77-80, 1991.

[14] 郭威志，使用語者辨認做前處理之國語 TTS 系統發展，碩士論文，國立交通大學電信系，2000。

[15] 李雪貞，客語語音合成之初步研究，碩士論文，國立台灣科技大學資工所，2001。

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