基於 VQ/HMM 之國語語句基週軌跡產生之方法

A VQ/HMM Based Sentence Pitch-Contour Generation Method for Mandarin

古鴻炎　　楊仲揆
Hung-Yan Gu　　Chung-Chieh Yang

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 Technology
E-mail: root@guhy.ee.ntust.edu.tw

摘要

本文提出一種以向量量化(VQ)與隱藏式馬可夫模型(HMM)為基礎的方法，來掌握一個國語語句中相鄰音節的基週軌跡的出現相關性，而使得產生出的句子基週軌跡在自然度上獲得大幅的改進，我們稱它為基於語句基週軌跡隱藏式馬可夫模型(SPC-HMM)之基週軌跡產生法。關於國語語句的 SPC-HMM 的建立，我們先對訓練語句中各音節的基週軌跡作時間與音高的正規化，及向量量化的處理，然後以各訓練語句中相鄰音節的量化碼組合之序列，來訓練 SPC-HMM 模型，其中音高的正規化，我們也提出了實際有效的方法。在合成階段，除了可使用三次元動態規劃演算法來產生基週軌跡量化碼序列之外，也可配合上游處理得到的資訊來規劃狀態轉移序列，如此可得到更具有韻律變化的合成語音。我們進行實際的聽測實驗後發現，當把一個平常人所唸出的語音的韻律喜好度定為 8 分而滿分為 10 分時，依本方法產生的語句基週軌跡來合成的語音，反而可得到更好的 8.2 分之喜好度。

關鍵詞：　文句翻語音，國語語音合成，基週軌跡，向量量化，隱藏式馬可夫模型

ABSTRACT

In this paper, a method based on vector quantization and hidden Markov model is proposed to model the co-occurrence relation of adjacent syllables' pitch-contours in a Mandarin sentence. With this method, large improvement of naturalness is obtained for the generated sentence pitch-contour. We name it as the pitch-contour generation method based on sentence pitch-contour hidden-Markov-model (SPC-HMM). To construct an SPC-HMM for Mandarin sentences, the pitch-contours of the syllables comprising each training sentence are normalized on both time and frequency axes first. The method for frequency normalization is effective and newly proposed here. After normalization, the pitch-contours are vector quantized. Then, the quantization codes of adjacent syllables are combined and treated as the observation symbols for hidden Markov modeling. In the synthesis phase, the quantization code sequence representing a sentence's pitch-contour can be generated with a three-dimension dynamic programming algorithm. In addition, the state transition sequence can be prepared according to the information from the text-analysis stage. By such preparation, the synthesized speech is perceived as having better prosodic presentation. We had conducted practical perception tests. The prosody-preference score of the speech uttered by an ordinary speaker is defined as having 8 points while the perfect prosody has 10 points. Under this definition, it is unanticipated that the speech synthesized by using the sentence pitch-contour generated from our method is evaluated to have 8.2 points.

Key Words:　text-to-speech, Mandarin speech synthesis, pitch contour, vector quantization, hidden Markov model

1. 前言

國語文句翻語音系統的應用相當廣泛，如電腦輔助教學系統、有聲書、多媒體資訊服務系統等，都可加入文句翻語音的功能，來使得系統變得更爲人性化。一般說來，一個國語文句翻語音系統可看成是由三個處理單元所組成，即文句分析(text-analysis)、韻律(prosodic)參數值產生、語音信號波形合成等三個單元 [1,2]。其中，文句分析單元要對輸入的文句作分析，以得到一個句子中各中文字的發音資訊，如音節的編號、聲調等，此外也包括斷詞處理、呼吸群決定、破音字發音、變調處理等。接著，先簡介語音信號合成單元，再來說明和本文主題較相關之韻律參數值產生單元，語音信號合成的方法可粗略分爲三類：(1)時域方法，如 PSOLA (pitch synchronous overlap and add) [3]，TIPW (time-proportioned interpolation of pitch waveform) [4]，(2)頻域方法，如幅峰合成(Formant Synthesis) [5]，(3)混合域方法，如混合使用時域與頻域的方法 [6]。其中 TIPW 法是我們的原型文句翻語音系統中所使用的語音信號合成方法，它的特色是，可以免除 PSOLA 法裡會發生的迴音(reverberation)現象與雙聲調(chorus or dual tone)現象，並且，經由調整音調高低及聲道長度，可合成出許多的音色來，而使合成出的語音信號更生動活潑。

韻律參數值產生單元，其功能是依據文句分析單元所得到的資訊來產生各個韻律參數的數值，如各音節的基週軌跡(pitch contour)、音長(duration)、音量(amplitude)、及音節前停頓(pause)等。過去已有許多關於韻律參數值產生之方法被提出，大致上可將它們分成三類：(1)規則依循(rule-based)，由專家來從大量的文句中分析出有效、完整的規則 [6,7]。(2)類神經網路(neural network)，先藉由類神經網路的學習程序，將韻律變化的規律性以類神經元的權重記錄下來，然後在合成階段用以產生出韻律參數的數值 [8]。(3)統計式(statistical)，使用統計的方法來建立語言特徵和韻律參數值之間的對應關係，過去被提出的研究成果包括統計式基週軌跡合成方法 [9]，階層式詞語韻律之樣式樹 [10]，及音量和音調參數值的迴歸(regression)分析 [11]。

在各種韻律參數中，基週軌跡是最重要的一個，因爲它對合成語音的自然流暢度的影響是最大的，因此過去有許多研究都把焦點放在基週軌跡的產生上 [7,9]。雖然本文也是研究基週軌跡的產生，但是我們是以不同的方法來研究這個問題，特點是，透過新提出的音高正規化作法，可大幅減少準備訓練語句的人力花費，並且以

隱藏式馬可夫模型來對整個句子的基週軌跡作考慮，可免除規則依循方法的規則擷取的問題。我們希望用此方法來取代我們先前使用的規則依循的作法，而使合成語音的自然度(naturalness)能獲得大幅的改進。事實上，本文提出的方法經實驗驗證得知，已讓自然度、韻律喜好度獲得大幅的提升了，若有疑問或感興趣，歡迎隨時至 http://guhy.ee.ntust.edu.tw/gutts 進行線上測試。

關於句子基週軌跡產生的問題，本文提出的解決方法是以向量量化(VQ)與隱藏式馬可夫模型(HMM)爲基礎的，稱爲基於句子基週軌跡隱藏式馬可夫模型(SPC-HMM)之基週軌跡產生法，這裡的 SPC-HMM 模型是本文新提出來的。考慮一個音節的基週軌跡，除了會受音節本身的組成音素、聲調、及相鄰音節的影響之外，也必須考慮整句話的行進所帶來的影響。根據前人的研究可知道，直述句大體上的趨勢是，在句末音量會減少、音調會下降，因此我們想至少可以"句首"、"句中"與"句尾"等三個狀態來區分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，因此我們決定使用語音辨識領域常被採用的 HMM 技術 [12]來處理這個問題，也就是令"句首"、"句中"與"句尾"等三個韻律狀態對應到 HMM 的隱藏狀態，接著，以向量量化技術來對音節的基週軌跡作量化編碼，據以建立基週軌跡量化碼和 HMM 的觀測符號(observation symbol)的對應關係。除了考慮觀測符號序列的發生機率外，我們也將相鄰音節基週軌跡之音高差因素含納進來考慮。此外，對於 HMM 的訓練資料不足的問題，如某些聲調組合未出現過，我們也設計了一個降階的機制，以使 HMM 模型在訓練語句不足時，仍能表現出一定的效能。

雖然過去也有人提出以 HMM 爲基礎的方法 [13,14]，來研究基週軌跡產生的問題，不過，Ljolej 和 Fallside 研究的焦點是放在分離(isolated)音節的基週軌跡的產生上，使用 1 個混合(mixture)的連續(continuous) HMM 的 4 個狀態來記錄音節中不同段落的音高，而 Fukada 研究的焦點是在日本語多音節詞的基週軌跡的產生上，基本上是以前者的音節基週軌跡 HMM 模型加以串接的作法來達成。本文的研究和他們不一樣的是，要先將音節的基週軌跡做向量量化，然後以離散(discrete)HMM 模型的狀態，來模化(modeling)整句話發音的韻律狀態，而不是著眼於音節的 HMM 模型。很明顯的在我們的 HMM 裡，時間軸的刻度是以音節來計數的，而前人的 HMM 裡，時間軸的刻度是以基週來計數的。

SPC-HMM 建立的工作流程如圖 1 所示，在圖中最左邊的虛線方塊，代表音節基週軌跡碼書(codebook)的建立流程，首先訓練語句的語音波形先以人工切割出音節，接著作半自動的基週頂點(pitch peak)偵測，再經過時間正規化及句子音高正規化的處理，而得到正規化的基週軌跡資料。接著，訓練語句中各聲調的音節經由碼書訓練程序就可得到各聲調的基週軌跡碼書。有了各聲調的基週軌跡碼書及經過正規化的基週軌跡資料，訓練語句的組成音節就可依據聲調來選碼書去作向量量化的處理，而得到各訓練語句的音節基週軌跡量化碼序列。上述基週軌跡的時間正規化、句子音高正規化及碼書建立等的詳細作法，將在第二節中說明。在圖 1 中央的虛線方塊是，SPC-HMM 的訓練流程，由之前得到的音節聲調資料、量化碼字序列、正規化基週軌跡等資料來訓練 SPC-HMM 模型。在圖 1 中最右邊的合成階段，則根據所訓練出來的模型參數，使用 3D(dimension)動態規劃演算法來產生出無狀態駐留限制的最佳基週軌跡，或經過規劃狀態轉移序列之最佳基週軌跡。關於 SPC-HMM 的訓練和合成部份，我們將在第三節作詳細的說明。在第四節則對本文提出的方法做一試聽實驗和評估。第五節爲結論，對本文之研究做一總結。

2. 音節基週軌跡向量量化

2.1 音節基週軌跡之時間正規化

要對音節基週軌跡作向量量化，首先會遇到基週軌跡的表示問題。在前人的研究裡曾提出正交轉換(Orthogonal Transform)的方法 [9]，以四個正交多項式做基底來將音節基週軌跡轉換成四個係數，然後在這四個係數上定義距離量測，但是這樣的距離量測是否能準確量出兩基週軌跡間的聽覺距離呢？所以這裡我們考慮使用直接的表示法，實際上是以時間比例內差來將不同

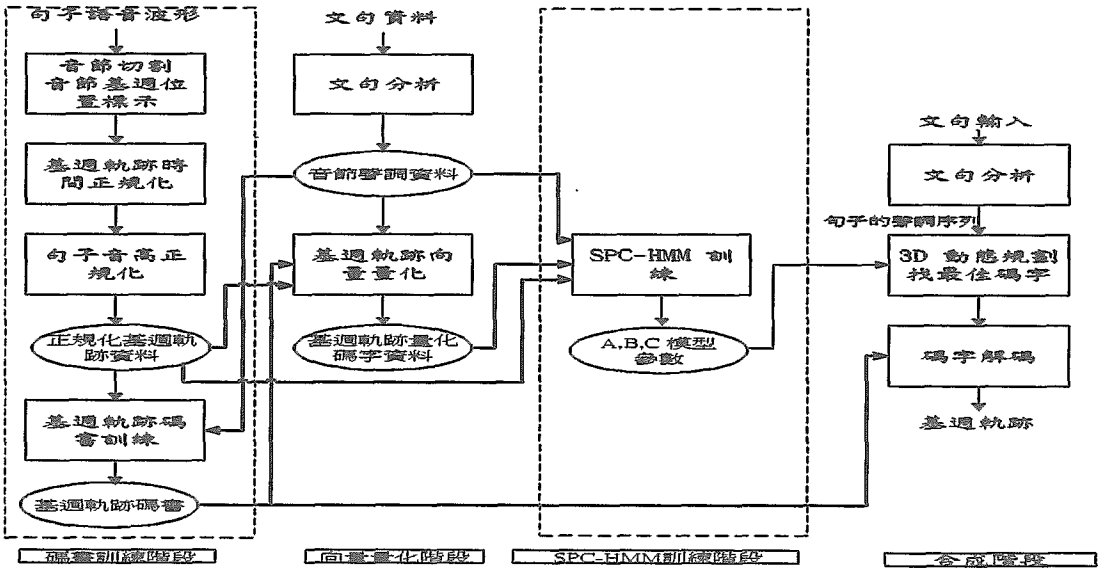
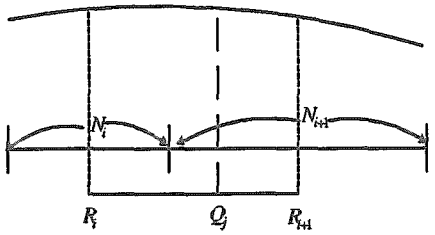


圖 1 建立 SPC-HMM 模型之流程圖

長度的基週軌跡，都正規化成時間軸上具有固定的點數(或固定的週期個數)。詳細的正規化方法，我們以圖 2 來說明，圖中曲線代表某個音節的部分基週軌跡，N_i 表示該音節中第 i 個週期的樣本點數，N_{i+1} 表示下一個



週期裡的樣本點數，且假設總共有 n 個週期，而各週期是以中心點 R_i 來代表該週期的時間位置，則當要在時間點 Q_j 上求此點上對應的週期長度時，我們以時間比例爲權重來內差出新的基週長度 L，即令

$$Q_j = (\frac{j}{m-1}) \sum_{i=1}^{n} N_i , \quad j = 0, 1, \ldots, m-1 \tag{1}$$

$$L = \frac{Q_j - R_i}{R_{i+1} - R_i} \times N_{i+1} + \frac{R_{i+1} - Q_j}{R_{i+1} - R_i} \times N_i \tag{2}$$

其中 m 表示正規化後時間軸上的點數，如此，就可以將原來具有 n 個週期的基週軌跡經由正規化後，用 m 個點來表示，當然使用越多的點數來表示基週軌跡會越精確，可是考慮到實際上音節的長度會有少於 16 個基週的短音節，於是我們就對 16、14 和 12 等三種 m 值來作考慮，分別去量測不同 m 值時的向量量化誤差和 HMM 模型內部測試的誤差，結果發現三種 m 值下量到的誤差值差距不大，但仍以 16 點時的誤差較小一些，所以我們就採用 16 點來對訓練用音節的基週軌跡進行時間正規化。至於時間正規化後的 16 點音節基週軌跡，將來如何用以合成出不同時間長度的音節信號波形，請參考另一篇論文 [4]。

本文中，關於向量量化誤差的計算是，在自然對數尺度上量測經時間正規化的訓練音節基週軌跡和對應的量化碼字(codeword)基週軌跡之間的平均均方根誤差；而 HMM 模型內部測試(inside test)的誤差是，量測 HMM 模型所產生出來的句子基週軌跡量化碼字序列所對應的基週軌跡，和原始訓練語句的時間正規化之基週軌跡在自然對數尺度上之平均均方根誤差。兩個經時間正規化之音節基週軌跡間的均方根誤差的計算公式爲

$$D(x, y) \equiv \sqrt{\frac{\sum_{i=1}^{m} (\ln(x_i) - \ln(y_i))^2}{m}} \tag{3}$$

其中 x_i 與 y_i 表示時間正規化基週軌跡的第 i 點上的音高(單位 Hz)。

2.2 句子音高正規化

由於訓練語句的錄製需要一段時間，而發音者的情緒很難維持一致，使得在不同情緒下所說出來的同一句話，句子基週軌跡也會有高低之差別，而錄音時的身體狀態(如感冒鼻塞)也會對句子基週軌跡造成影響。如果把未作音高正規化處理之訓練語句，直接拿去訓練 SPC-HMM，其結果將是，合成出來的句子基週軌跡會有忽高忽低之異常起伏的現象，因此，訓練語句的基週軌跡必需先作音高正規化的處理。

正規化的方法，如前人的研究裡所用的 [9]，是將同一句話發音數次，如此一個訓練語句中的各個中文字，都可得到數個基週軌跡之量化碼，然後從中選取出現次數最多的一個作爲音高正規化的結果。在本文裡，爲了降低準備訓練語句的人力花費(或者說是人力有限)，只對一個訓練語句發音一次，因此，我們便研究了不同的音高正規化方法，第一種方法是，先計算所有訓練語句的平均句子音高，再以句子爲單位，比較一個句子的音高和平均句子音高的差異，然後對一個句子的所有音節，進行相同的音高調整，詳細的步驟爲：

(1) 求出第 i 個訓練語句中各個音節的對數尺度音高 E_j，再計算此語句的句子音高 S_i，即令

$$E_j = \frac{1}{16} \sum_{k=1}^{16} p_{jk} , \quad j \text{ 表示第 } j \text{ 音節,} \tag{4}$$

p_{jk} 表示第 j 個音節之第 k 個時間點上的對數尺度音高

$$S_i = \frac{1}{n} \sum_{j=1}^{n} E_j , \quad \text{設第 } i \text{ 句裡有 } n \text{ 個音節,} \tag{5}$$

(2) 將所有訓練語句的句子音高做算數平均，以得到所有訓練語句的平均句子音高，即令平均句子音高 SA 爲

$$SA = \frac{1}{ST} \sum_{i=1}^{ST} S_i , \; ST \text{ 表示訓練語句的總句數} \tag{6}$$

(3) 將各個訓練語句的句子音高和平均句子音高相減，就可以得到各語句的句子音高調整值，即令

$$\Delta_i = S_i - SA \tag{7}$$

(4) 根據第 i 句的句子音高調整值 Δ_i，調整此語句中各個音節的音高，即令

$$\overline{p}_{jk} = p_{jk} - \Delta_i , \; k = 1,2,\ldots,16, \; j = 1,2,..,n, \tag{8}$$

這樣的作法可說是一種簡單的音高正規化方法，並不是最有效的解決方法，所以後來我們又作了另一種嘗試，就是把經由上述的音高正規化處理後的訓練語句，拿去訓練具有三個狀態之 SPC-HMM 模型，然後依訓練語句的狀態序列去求取模型的每個狀態上各聲調的平均音高，當作另一種音高正規化方法的參數，詳細的正規化步驟爲：

(1) 將每個訓練語句裡的音節序列平均切割成三段，依序分配給 SPC-HMM 模型的 0、1、2 三個狀態。

(2) 將一個訓練語句中的各個音節的音高 E_j，和該音節在相對應狀態上應有的聲調平均音高相減得到 d_j，再取此一語句中各 d_j 的算數平均 $\overline{d}$，即令

$$d_j = E_j - H_{ku} , \; j = 1,2,..,n, \tag{9}$$

$$\overline{d} = \frac{d_1 + d_2 + \ldots d_n}{n} \tag{10}$$

設此語句共有 n 個音節，E_j 表示第 j 個音節的音高，依公式(4)計算，
設第 j 個音節的聲調爲 u，且被分配到狀態 k，H_{ku} 表示 SPC-HMM 模型狀態 k 上，聲調 u 之平均音高，

(3) 根據音高調整值 $\overline{d}$，將此語句中各個音節的 16 點基週軌跡，進行如下之調整

$$\overline{p}_{jk} = p_{jk} - \overline{d} , \; k = 1,2,\ldots,16, \; j = 1,2,..,n, \tag{11}$$

p_{jk} 表示第 j 個音節之第 k 個時間點上的對數尺度音高,

如上的第二種正規化方法是比較精細的，因爲它根據所有訓練語句訓練 SPC-HMM 模型後，所統計出的各狀態上各聲調的平均音高爲比較依據，但是以語句爲單位來作音高調整。經過這樣的音高正規化處理後，向量量化誤差變得更小了，而且在實際試聽後發現更沒有異常的音調起伏了。表 2 裡所列的是，量測向量量化誤差和

表 1 句子音高正規化方法與誤差量測

	未正規化	正規化方法 1	正規化方法 2
向量量化誤差	0.0398	0.0330	0.0308
HMM 內部測試	0.0720	0.0502	0.0426

HMM 模型內部測試誤差所得的數據，由此表可知，用音高正規化方法一後的誤差已經有不少的改進，而音高正規化方法二的效果更好，平均每個音節在對數尺度上之向量量化均方根誤差爲 0.0308(相當於 3.7Hz 於 120Hz)，而 SPC-HMM 模型內部測試得，平均每個音節在自然對數尺度上之均方根誤差爲 0.0426(相當於 5.1Hz 於 120Hz)。

2.3 音節基週軌跡向量量化

在向量量化的分類過程中，必須選擇有效的距離量測公式來對基週軌跡作分類，不過距離量測是否能反應人類的聽覺距離的問題也必須納入考慮，例如交錯的兩條基週軌跡，在聽覺上其距離應比兩條相鄰且平行的軌跡來的大，過去也有其他的研究者持這樣的看法 [11]，如果只採用幾何距離，將很難區分出這一點，因爲可以推想得知的是，當以幾何距離作爲基週軌跡向量量化的距離量測時，所得到的碼字(codeword)之基週軌跡會比較平緩，因爲具有交叉的基週軌跡也會被分到同一類，而當以平均的方式來取此類別的中心點基週軌跡時，中心點基週軌跡必然會趨於平緩。爲了克服這個困難，我們曾經研究並設計了一個距離量測的公式，稱爲交叉加權距離公式 [15]，使用交叉加權距離量測來作基週軌跡向量量化之分類，就比較能夠維持同類基週軌跡原先的斜率，並且實際上試聽時也發現，使用交叉加權量測所得到的碼書比較有更明顯的抑揚頓挫，因此，如果原先錄音者的抑揚頓挫不是很明顯的，建議使用交叉加權之距離量測公式。不過，本文爲了評估上的方便，仍採用了幾何距離來進行實驗。

本文中， 基週軌跡向量量化碼書(codebook)的訓練，採用了 K-means Clustering 演算法 [12]，而整體的處理流程如圖 1 左邊虛線部分所示。關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，這可從我們所作的實驗結果觀察出來，數據如表 2 所示，所以，在有限的訓練語句及參考前人的研究成果之下，我們選擇將各個聲調的音節基週軌跡都量化成 8 類，即國語五個聲調的每個聲調的碼書大小都爲 8。

表 2 碼書大小與誤差之關係

碼書大小	向量量化誤差	HMM 內部測試
4	0.0386	0.0468
6	0.0335	0.0439
7	0.0324	0.0429
8	0.0308	0.0426
9	0.0301	0.0432
10	0.0296	0.0435
12	0.0283	0.0453

3. 基週軌跡隱藏式馬可夫模型

3.1 SPC-HMM 之建立

由於整句話的行進(句調)對於音節基週軌跡的影響是不易精確描述的，因此，我們便有這樣的想法，以 HMM 裡的隱藏狀態來掌握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，基本上我們想以 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，由於韻律狀態的轉移是只能前進而不能後退的，所以韻律狀態的轉移情形就如圖 3 裡的情況，其中 a_{ij} 表示由狀態 i 轉移到狀態 j 的機率。

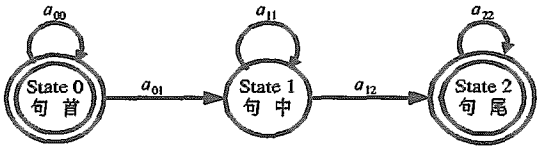


圖 3 韻律狀態轉移圖

關於離散 HMM 的觀測符號，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(observation)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之組合，即令時刻 t 時的觀測符號爲

$$O_t = F_{t-1} \times F_t \times F_{t+1} \times V_t \ , \qquad (12)$$

F_t 表示句子中第 t 個音節的聲調, $0 \le F_t \le 4$

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

然後就可以 $b_j(k)$ 參數來描述在狀態 j 上產生觀測符號 k 之機率。不過，當訓練語句不夠多，而發生某一種聲調組合之觀測符號沒出現時，要如何處理？對於這種狀況，我們考慮了三層的降階作法，然後將各層上所定義的觀測符號看作是獨立的，並且各層各自去訓練出該層所對應的 a_{ij} 、 $b_j(k)$ 等參數。

在第三層的觀測符號，如公式(12)所示，考慮前一個音節聲調、本音節聲調、下一個音節的聲調及本音節基週軌跡量化碼的組合，當 t 不是句子的第一個及最後一個字時，觀測符號的數值範圍設爲 $0 \le F_{t-1} \times F_t \times F_{t+1} \times V_t < 5 \times 5 \times 5 \times 8 = 1000$，當 t 是句子的第一個字時，觀測符號的數值範圍設爲 $1000 \le F_t \times F_{t+1} \times V_t + 1000 < 5 \times 5 \times 8 + 1000 = 1200$，當 t 是句子的最後一個字時，則設爲 $1200 \le F_{t-1} \times F_t \times V_t + 1200 < 5 \times 5 \times 8 + 1200 = 1400$。在第二層裡的觀測符號，考慮前一個音節聲調、本音節聲調及本音節基週軌跡量化碼的組合，觀測符號的數值範圍設爲 $1400 \le F_{t-1} \times F_t \times V_t + 1400 < 5 \times 5 \times 8 + 1400 = 1600$。在第一層裡的觀測符號，考慮本音節聲調、下一個音節聲調及本音節基週軌跡量化碼的組合，觀測符號的數值範圍設爲 $1600 \le F_t \times F_{t+1} \times V_t + 1600 < 5 \times 5 \times 8 + 1600 = 1800$。 如此在合成階段裡，當前後音節的聲調組合沒

有看過，就往下降一階，而在 SPC-HMM 的訓練階段，各層的觀測機率是獨立去訓練的，不需要考慮降階的情況。

另外，由於訓練語句並不充足，對於由相同聲調組合但不同的基週軌跡量化碼所組合出的觀測符號，我們也使用了機率分享的方式，來提升出現次數較少的觀測符號的機率值，就是將一個出現的基週軌跡量化碼的機率分享給同一個聲調中最接近它的另外兩個基週軌跡量化碼，分享的比率爲 0.0009 和 0.0001。

對於 SPC-HMM 模型的訓練，我們採用了 Segmental K-means 之訓練方法 [12]，這個方法中的一個重要步驟是，以維特比(Viterbi)動態規劃演算法將每一句訓練語句目前對應的最佳狀態序列找出，然後依據狀態序列將觀測符號序列切割、分給各個狀態，再去重新估計各狀態上的機率參數的數值。在訓練階段裡，各個訓練語句所對應的觀測符號序列，一定可以找到一條狀態序列，使得各個時刻的 $b_j(O_t)$ 參數都有機率值，也就是沒有降階的問題，因此，三個降階層的 SPC-HMM 的訓練可以各層獨立地去進行，訓練完後再將三個階層的觀測符號機率參數 B，及狀態轉移機率參數 A 一併儲存起來。

3.2 基週軌跡連接之音高差

由於隱藏式馬可夫模型假設，在同一個狀態上依序被產生的數個的觀測符號，相鄰的觀測符號間是獨立的而沒有互動關係的，但是實際上在一個語句裡，一個音節的基週軌跡是會受到韻律狀態及前後音節的影響的，所以 3.1 節裡的 SPC-HMM 只藉由觀測符號的發生機率來選擇音節的基週軌跡，而沒有直接考慮音節與音節之間基週軌跡連接的因素，因此這裡我們考慮一種增進 SPC-HMM 模型的模化(modeling)能力的作法，就是將相鄰兩音節的基週軌跡之間的音高差考慮進來，實際上的作法是，新增加一個 C 參數來記錄相鄰音節之基週軌跡重心的差值的期望值，也就是以 $c_j(k)$ 來記錄狀態 j 上觀測符號 k 的相鄰兩音節的基週軌跡重心差值的期望值，當行進到第 t 時刻 (音節)時，是指時刻 t-1 與 t 上的兩個相鄰音節，至於重心差的標準差(standard deviation)值就不記錄了，因爲我們的訓練語句並不夠多。

有了音高差的觀念後，那麼音節基週軌跡的重心要如何定義呢？爲了避免基週軌跡重心無法描述基週軌跡斜率的問題，我們定義一個音節的基週軌跡有前後兩個重心，前重心 WF 的定義爲，經過時間、音高正規化的基週軌跡，在對數尺度上取前八點上的音高作算數平均，而後重心 WB 的定義爲，在對數尺度上取後八點上的音高作算數平均，即令

$$WF_t = \frac{1}{8}\sum_{j=1}^{8} p_{t\,j}, \qquad WB_t = \frac{1}{8}\sum_{j=9}^{16} p_{t\,j}, \tag{13}$$

$p_{t\,j}$ 表示第 t 個音節的第 j 點上的對數尺度音高

如此，音高差的定義爲，本次音節(t 時刻)的前重心減掉前一個音節(t-1 時刻)的後重心，即令

$$W_t = WF_t - WB_{t-1} \tag{14}$$

在加入了 C 參數後，原先維特比演算法中關於累積距離 $\tilde{\delta}$ 的定義與計算 [12]，必須做一些修正，詳細情形將在 3.3 節中說明。在訓練完 SPC-HMM 模型得到 A、B、C 參數後，爲了瞭解模型所合成出的基週軌跡的效果，我們分別採用內部測試(inside test)和外部測試(outside test)來評估模型的好壞，內部測試是以參加模型訓練的語句來測試模型，將模型產生出來的基週軌跡量化碼所對應的軌跡，和經時間、音高正規化的訓練語句基週軌跡相比較，而外部測試則是拿沒有參加過模型訓練的語句來測試模型，將模型所產生出來的基週軌跡量化碼所對應的軌跡，和經時間、音高正規化的測試語句基週軌跡相比較。比較的方式是，量測對應音節的兩條基週軌跡之間的均方根誤差，再將所有音節的均方根誤差取平均。這裡做內部測試的句子共有 375 句，外部測試的則有 45 句，並且，SPC-HMM 模型的輸入是語句的聲調序列，而每一種聲調都有 8 個量化碼可供選擇，這裡是以 SPC-HMM 的 Model_B 合成模式來選擇，詳細作法將在 3.3 節裡說明。結果我們得到如表 3 裡所示的誤差值，第一欄是只使用 A、B 參數時的 SPC-HMM 模型的平均音節均方根誤差，而第二欄是加入 C 參數後的平均音節均方根誤差，比較此二欄可看出，加入音高差之 C 參數，約可得到 5%的改進幅度。

表 3　SPC-HMM 模型的內、外部測試

	A,B 參數	A,B,C 參數
內部測試	0.0449	0.0426
外部測試	0.0553	0.0524

3.3 語句基週軌跡合成

SPC-HMM 模型的訓練程序和語音辨識裡的訓練程序類似，但是合成階段裡的處理，就不同於語音辨識的辨識階段，因爲 SPC-HMM 模型的合成階段的輸入是一個語句的組成音節的聲調串成的序列，而不是訓練階段裡所用的音節基週軌跡量化碼序列，並且反過來要輸出的卻是音節基週軌跡量化碼序列，這就如圖 1 右邊部分所顯示的情況。

當使用 SPC-HMM 模型來爲一個聲調序列產生對應的基週軌跡量化碼序列時，一者我們可以完全信賴 SPC-HMM 模型，依模型的定義去找一個模型機率上最佳的基週軌跡量化碼序列，這種找法在本文中稱爲 Mode_A 合成模式，其實行作法將在 3.3.1 節中說明；不過從另一個角度來看，SPC-HMM 模型在訓練時及 Mode_A 合成模式時，都未用到呼吸群及詞邊界的資訊，而呼吸群及詞邊界對基週軌跡的影響是很強烈的 [10,16]，那麼我們又如何能期望在 Mode_A 合成模式下可產生出生動、逼真的基週軌跡呢？因此，我們另外設計了二種以 SPC-HMM 模型爲基礎的基週軌跡合成模式，稱爲 Mode_B 與 Mode_C，在這兩個模式裡，HMM 的狀態轉移序列被加上了限制，如 Mode_C 裡是依據文句分析階段得到的呼吸群及詞邊界資訊來規劃、設定 HMM 的狀態轉移序列，而不像 Mode_A 裡是完全無狀態駐留時間之限制的 [17,18]，詳細情形在 3.3.2 節中說明。

3.3.1 基週軌跡合成模式 Mode_A

這一節說明如何依模型的定義去找一個模型機率上最佳的基週軌跡量化碼序列。由公式(12)可知，時刻 t 時的觀測符號是由前後音節的聲調和本音節的聲調、量化碼等四個因素所組合出來的，然而在合成階段，時刻 t 時的音節，除了聲調是輸入進來的、是已知的之外，本音節聲調可能對應的 8 種基週軌跡卻是等待我們從其中選出一個來，在不知如何選取的情況下，我們可先把 8 種基週軌跡和前、本、後音節的聲調組合出 8 個候選

的觀測符號，如此，在時間軸和狀態軸交織的每一個格子點上，都有 8 個候選的觀測符號，這和訓練階段裡找最佳狀態序列的演算法中，每一個時間軸與狀態軸的格子點上只有一個觀測符號的情況比起來，等於是增加了一個基週軌跡的次元，所以，我們必需把訓練階段裡的時間與狀態之二次元搜尋空間，擴充成時間、狀態、與基週軌跡的三次元搜尋空間，然後以 3 次元的動態規劃演算法來找出一條統計上最佳的路徑，而經由此路徑就可知道各個時刻上該選擇那一個觀測符號，再由選出的觀測符號也就可以知道對應的基週軌跡量化碼了。

在說明三次元動態規劃演算法之前，先定義如下之符號：

T 與 N：表示語句中的音節總數，及 HMM 模型的狀態個數。

π_i：在狀態 i 上的初始機率。

$O_t(k)$：表示在時刻 t 時，本音節聲調 F_t 的第 k 個基週軌跡量化碼所對應的觀測符號向量 $(O_t(k,3),O_t(k,2),O_t(k,1))$，這個向量的三個組件，分別代表降階的第三層、第二層與第一層之觀測符號。聲調 F_t 有 8 個軌跡可供選擇，即 $0\le k\le 7$。

$b_j(O_t(k))$：表示在時刻 t 時，狀態 j 上產生觀測符號向量 $(O_t(k,3),O_t(k,2),O_t(k,1))$ 的機率值，其計算方式爲，先檢查第三層的 $O_t(k,3)$，當用第三層的 $b_j(\bullet)$ 參數發現 $b_j(O_t(k,3))>\Omega$，就設定 $b_j(O_t(k))$ 之值爲 $b_j(O_t(k,3))$，否則往下降一階至第二層，每下降一階的費用爲 ε (例 10^{-5})，就是將 $b_j(O_t(k))$ 之機率值乘上 ε 以區分不同的降階層，在第二層再用第二層之 $b_j(\bullet)$ 參數，檢查是否 $b_j(O_t(k,2))>\Omega$，如此繼續。本文裡設 Ω 爲 10^{-15}，ε 爲 10^{-5}。

$\tilde{\delta}_t(i,k)$：在時刻 t 到達狀態 i 並選擇聲調 F_t 之第 k 個基週軌跡量化碼之最短累加距離。

$\psi_t(i,k)$：在時刻 t 到達狀態 i 並選擇聲調 F_t 之第 k 個基週軌跡量化碼的最佳路徑，用以指示時刻 t-1 時的最佳狀態。

$\beta_t(i,k)$：在時刻 t 到達狀態 i 並選擇聲調 F_t 之第 k 個基週軌跡量化碼的最佳路徑，用以指示時刻 t-1 時的最佳基週軌跡量化碼。

$\tilde{P}^*$：到達語句最後一個音節時的最短距離。

q_t^*：最佳狀態序列。

v_t^*：最佳基週軌跡量化碼序列。

我們使用的 3 次元動態規劃演算法之處理步驟爲：

(1)前處理：首先將模型各參數值轉換到對數尺度上。

(2) 初始化：

$$\tilde{\delta}_1(i,k)=|\tilde{\pi}_i|+|\tilde{b}_i(O_1(k))|,\qquad 0\le i\le N-1,\ 0\le k<8$$

$$\psi_1(i,k)=-1,\qquad 0\le i\le N-1,\ 0\le k<8$$

$$\beta_1(i,k)=-1,\qquad 0\le i\le N-1,\ 0\le k<8$$

(3)遞回計算：

$$r_t(j,k)=W_t-c_j(O_t(k))$$

$$\tilde{\delta}_t(j,k)=\min_{0\le i\le N-1}\min_{0\le h\le 7}[\tilde{\delta}_{t-1}(i,h)+|\tilde{a}_{ij}|+|\tilde{b}_j(O_t(k))|+|r_t(j,k)|] \qquad (15)$$

$$\psi_t(j,k)=\arg\min_{0\le i\le N-1}[\min_{0\le h\le 7}[\tilde{\delta}_{t-1}(i,h)+|\tilde{a}_{ij}|]],$$

$$\beta_t(j,k)=\arg\min_{0\le h\le 7}[\min_{0\le i\le N-1}[\tilde{\delta}_{t-1}(i,h)+|\tilde{a}_{ij}|]],$$

(4)結束：

$$\tilde{P}^*=\min_{0\le k\le 7}[\tilde{\delta}_T(N-1,k)]$$

$$q_T^*=N-1$$

$$v_T^*=\arg\min_{0\le k\le 7}[\tilde{\delta}_T(N-1,k)]$$

(5)路徑回溯：找出最佳的狀態序列、基週軌跡量化碼序列。

$$q_t^*=\psi_{t+1}(q_{t+1}^*,v_{t+1}^*),\quad t=T-1,T-2,...,1$$

$$v_t^*=\beta_{t+1}(q_{t+1}^*,v_{t+1}^*),\quad t=T-1,T-2,...,1$$

在公式(15)中，我們曾考慮過將 A、B 參數乘上權重 $g1$，C 參數乘上權重 $g2$，即令

$$\tilde{\delta}_t(j,k)=\min_{0\le i\le N-1}\min_{0\le h\le 7}[\tilde{\delta}_{t-1}(i,h)+|\tilde{a}_{ij}|\times g1]+|\tilde{b}_j(O_t(k))|\times g1+|r_t(j,k)|\times g2$$

然後分別調整這二個權重的值，進行模型的內、外部測試實驗，看看是不是能得到更低的誤差，結果發現改進的量很小而可以忽略。

3.3.2 基週軌跡合成模式 Mode_B、Mode_C

我們所以會想到設計 Mode_B 與 Mode_C 之基週軌跡合成模式，是因爲在 Mode_A 模式裡，SPC-HMM 模型所找出的機率上最佳之路徑，其對應的狀態序列中，各狀態的出現次數經常是極不平均，例如 0,0,0,0,1,2 之狀態序列，在"句首"停留 4 次，卻只在"句中"與"句尾"各停一次，使得依據此路徑對應之基週軌跡所合成的語音信號，句調聽起來顯得不自然、不像人講的。這裡我們說圖 3 裡的狀態 0,1,2 就對應到句首、句中、句尾之韻律狀態，是因爲在 SPC-HMM 之訓練階段，我們曾統計各狀態上的各個聲調的訓練音節之平均音高，結果發現一個跨聲調的共同趨勢，就是狀態 0 上的音高比狀態 1 上的高，而狀態 1 上的比狀態 2 上的高，這符合一般觀察到的句調會隨著時間逐漸降低的現象。由此可知，在 Mode_A 模式裡找出的路徑，雖然是 SPC-HMM 模型定義上機率最佳之路徑，但是因爲 SPC-HMM 模型未考慮呼吸群、詞邊界等資訊，且沒有設定狀態駐留時間之限制，所以找出的路徑的韻律狀態轉移情形經常是不符合人類的說話方式的。

爲了消除 Mode_A 模式裡，所找出路徑的狀態序列中，各狀態的出現次數經常很不平均的情形，我們想到的作法是，在 3.3.1 節的路徑尋找演算法中，加入對各個時刻上的音節應該停留之狀態的限定，也就是將那些不該停留之狀態上的累加距離直接設爲無限大。那麼，狀態駐留的規則要如何設計呢？一個簡單的規則是，直接將一個語句的 n 個音節切割成三等分，然後限定前三分之一的音節要在狀態 0 上停留，而中間及後三分之一的音節各在狀態 1 與 2 上停留，在這種限定下的路徑找法，稱爲 Mode_B 基週軌跡合成模式。

除了前述的狀態駐留之簡單規則，我們還利用呼吸群、詞邊界等資訊來設計了另一個規則，稱爲 Mode_C 之合成模式，詳細情形如下，假設 X_1, X_2, X_3, X_4, X_5, Y_1, Y_2, Y_3, Y_4 是一個具有兩個呼吸群的語句，即 X 群後接 Y 群，則將第一個呼吸群的五個字平均切割成三段，令 X_1, X_2 於狀態 0 上停留，X_3, X_4 於狀態 1 上停

留，而 X_5 於狀態 2 上停留，至於第一群以後的第二、三群等，則都只切割成兩段，然後令 Y_1, Y_2 於狀態 1 上停留，Y_3, Y_4 於狀態 2 上停留。這樣的狀態轉移之安排，會發生由狀態 2 跳回狀態 1 的情形，而不符合圖 3 的要求，因此在合成階段使用 Model_C 模式時，我們就直接將 a_{21}, a_{22} 兩參數都設值爲 0.5。另外關於詞邊界之訊息，我們的利用方式是，在依據呼吸群來決定駐留之狀態後，再依詞邊界來檢查要不要作修正，原則是二字詞的組成字要駐留在同一個狀態上，因爲實際試聽候發現，二字詞的字若跨在不同狀態，有時會發生基週軌跡連接不自然的情形。不過，爲了讓整個語句的基週軌跡具有波浪般的韻律感，一個要確保的規則是，第一個呼吸群的最後一字要在狀態 1 或 2 上駐留，而第一群以後的其它群的最後一字要在狀態 2 上駐留。

爲了比較前述的三種基週軌跡之合成模式，我們進行了 SPC-HMM 模型的內、外部測試實驗，以量測音節基週軌跡的平均均方根誤差，結果得到如表 4 裡所示的誤差

表 4　三種基週軌跡合成模式的內、外部測試

	Model_A 模式	Model_B 模式	Model_C 模式
內部測試	0.0501	0.0426	0.0464
外部測試	0.0613	0.0524	0.0561

數值，由此表可看出，雖然 Model_A 合成模式(無狀態駐留限制)找出的路徑具有最高的 SPC-HMM 模型機率，但是它的音節基週軌跡均方根誤差，在內、外部測試裡都是最大的，造成這樣相反的情況，我們分析其原因是，雖然 SPC-HMM 模型有多於 5*5*5*8=1000 種的觀測符號，但是在 SPC-HMM 的三個狀態上，這些觀測符號的發生機率有著相當高的重疊性，如 5*5*5 種聲調組合的任何一個，幾乎都可能在句首、句中、與句尾出現，要分辨狀態原本可利用本音節的八種量化軌跡，不過量化軌跡八種仍嫌太少及音高正規化方法可能仍不夠完美，而使得八種量化之基週軌跡，在三個狀態上的發生機率仍有很高的重疊性，不像語音辨識裡，當以三個狀態之離散 HMM 來模化(modeling)音節(如/mai/)的頻譜特徵時，三個狀態上觀測符號的發生機率，不會有如此高的重疊性(例如/mai/中代表/m/頻譜的觀測符號的發生機率在狀態 0 上會很高，而在狀態 2 上會很低)。

另外，比較表 4 裡 Model_B 與 Model_C 之合成模式的音節基週軌跡均方根誤差，可知 Model_B 的誤差值比較小，我們認爲其原因是，當初唸訓練語句及測試語句的發音者，都是以直述句(一句唸到底)的方式在唸，而未注意句調的抑揚及呼吸群的停頓，因此，在 4.2 節的主觀聽測實驗裡，得到的是相反的結果，即 Model_C 比 Model_B 合成模式好。由以上描述的現象可知，這裡提出的 SPCH-HMM，模型本身(Model_A 模式)還不能掌握人類說話的結構性資訊(如呼吸群結構)，不過當把結構性資訊和 SPC-HMM 模型結合起來時(Model_C 模式)，產生出的整句話基週軌跡卻是含有韻律變化而可接受的，這也顯示人類在唸同一個句子時，並不是只有一種固定的唸法而已，而存在有多種能被接受的唸法，所以 Model_C 合成模式雖然有著較大的音節基週軌跡均方根誤差，但是在主觀測試裡，它產生出的基週軌跡的韻律喜好度卻是較高的。

4. 語音資料與聽測實驗

4.1 語音資料

在合成階段裡用來合成語音信號的波形資料，共有 409 個國語第一聲音節的波形，爲 16 bits 表示之樣本值，取樣頻率爲 11,025Hz，由一位男性在隔音間裡錄製，每次唸五個字左右的虛擬句子(無意義之音節組合)，錄好後再從句子的語音波形中將各個音節切割出來，相當於錄製虛擬的連續語音，以節省人力。雖然如此，比起過去以分離方式唸各個音節所合成出的語音信號來說，本次合成出的語音信號，的確變得平順多了。

關於 SPC-HMM 模型的訓練與測試，我們錄製的語音共有 420 句 3360 字，由同一個男性(事實上是本文作第二作者)去發音，一句念一遍。這些語句可分成三個部分來說明，第一部份是唸 112 句國語語音特性平衡句[19]，第二部分是從中央研究院平衡語料庫中考慮聲調組合後隨機選出的 263 句拿去唸，第三部份則是由報紙文章中隨機選出的 45 句拿去唸。第一和第二部分共有 375 句，我們拿來作爲訓練語句，其中唸第一聲的有 632 字，唸第二聲的有 754 字，唸第三聲的有 462 字，唸第四聲的有 867 字，唸輕聲的有 210 字。第三部分的 45 句，我們拿來作爲外部測試之用。關於訓練語句是否有含蓋所有可能的聲調組合，我們曾作了分析，在不考慮 HMM 的狀態時，所有可能的聲調組合(5*5*5=125 個)都出現過，而在分別考慮 HMM 的各個狀態時，則發現在個別狀態上某些聲調組合會較少出現或甚至沒有出現過，所以，本文在前面所提到的降階作法，就可看出其必要性。

4.2 聽測實驗和結果

關於以 SPC-HMM 模型爲基礎之語句基週軌跡產生的效能評估，在前面我們曾以模型的內、外部測試之音節軌跡均方根誤差數值當作客觀的評斷依據，以選擇相關參數的數值(如向量量化碼書大小)，這裡我們將對模型進行主觀的聽覺測試，依據 SPC-HMM 模型在 Mode_A,Mode_B 與 Mode_C 模式下找出的基週軌跡量化碼去合成出語音信號，不過，音量、音長等其它韻律參數值的設定，仍採取規則依循的作法，然後將合成的語音信號播放出給試聽者聽，看看所合成出來的語音聽不聽得懂，流暢不流暢，也就是評估可辨別度和韻律喜好度。參與測試者共 18 人，分別去聽我們的原型文句翻語音系統在三種基週軌跡合成模式下所合成出來的語音信號，以及先前一個規則依循版本所合成出來的語音信號[20]。

在可辨度的評估方面，我們先將二十句不同但難易度約略相同的句子分成四組，每組五句，然後分別用這裡的三種基週軌跡合成模式和前一個程式版本分別去合成出語音信號，並且在未告知試聽者的情況下，隨機選取版本撥放的順序，然後逐句播放給試聽者聽，而試聽者在聽完每句語音輸出後，就將所聽到的文句寫下來，然後依據測試者所寫下的文句去計算可辨度，即計算有多少百分比的國字被正確寫出。

在韻律喜好度的評估方面，我們訂定評分的範圍爲 1 至 10 分，先請一個平常人(非廣播工作者)將一篇文章唸出並且錄下來，在進行測試時就將此人唸的語音先播放給試聽者聽，作爲 8 分的評分標準，如此試聽者才有一個相對的評估標準。接者，分別以這裡的三種基週軌跡合成模式和前一個程式版本去合成出同一篇文章的語

音信號，並播放給試聽者聽，以對四種基週軌跡合成的方式作評分。

經過實驗後，我們得到如表 5 所示的結果，由表 5 之實驗結果可以看出，以 SPC-HMM 模型爲基礎所產生出的基週軌跡，的確比前一個 rule-based 的版本(每個聲調只有兩個軌跡樣板)好很多，使得可辨度可提高到 95%以上，最高者是 Model_C 合成模式之 96.5%；此外，在韻律喜好度方面，以 SPC-HMM 模型爲基礎的方法，不管用那一種基週軌跡合成模式，其評分都比前一程式版本的好許多，其中 Model_C 合成模式的韻律喜好度的評分爲 8.2，比人唸的參考標準 8 分還高，這是因爲當初唸的人，其發音的抑揚頓挫不甚明顯，而合成出來的語音信號反而有較明顯的韻律變化。

表 5 主觀聽測評估之結果

	前一版本	Mode_A	Model_B	Model_C
可辨度	81.2%	95.1%	96.2%	96.5%
韻律喜好度	5.1	7.0	7.7	8.2

5. 結論

在中文文句翻語音的研究上，要得到自然流暢的合成語音，語句的基週軌跡的產生是一個重要的因素。雖然過去已有一些基週軌跡產生的方法被提出，不過，本論文研究、提出了另外一種以句子基週軌跡隱藏式馬可夫模型(SPC-HMM)爲基礎的方法，它的確能夠產生出相當平順、自然的句子基週軌跡，而改進了我們的合成語音的自然流暢度與可理解度。此外它還有其它的特點，當訓練語句不很充足時，可透過 SPC-HMM 模型的降階機制來維持一定的效能，再者，當錄製訓練語句者心情不一致時(有時高昂、低沈、與鼻塞)，且每個語句只唸一次時，所錄得的語句信號波形都可以透過本文提出的基週軌跡音高正規化方法，來進行正規化處理，而使得 SPC-HMM 模型仍能被用以產生出平順的句子基週軌跡。

在 SPC-HMM 模型的合成階段裡，根據所訓練出來的模型參數，一者可使用 3D 動態規劃演算法來產生出模型機率上最佳的句子基週軌跡，此外，也可依據文句分析所得到的呼吸群、詞語界限的資訊，來規劃 SPC-HMM 模型裡的狀態轉移序列(即 Model_C 合成模式)，以產生出更富於韻律變化的合成語音，經由聽測實驗顯示，這樣的作法產生出的基週軌跡的確得到了較高的韻律喜好度，而比人唸的語音更被接受。所以，本文提出的 SPCH-HMM，雖然模型本身還不能掌握人類說話的結構性資訊(如呼吸群結構)，不過當把結構性資訊和 SPC-HMM 模型結合起來時，產生出的整句話基週軌跡卻是富於韻律變化而比人唸的還被接受。因此，將來可再研究 SPC-HMM 模型如何與結構性資訊作更細密的整合。另外，其它的韻律參數，如音量、音長、停頓等參數值的設定，將來也可嘗試使用隱藏式馬可夫模型來做進一步的研究。

國科會計畫編號：NSC 89-2213-E-011-058

參考文獻

[1] Wang, R., "Overview of Chinese Text-to-Speech Systems", International Symposium on Chinese Spoken Language Processing (Singapore), 1998.

[2] Shih, C. and R. Sproat, "Issues in Text-to-Speech Conversion for Mandarin", Computational Linguistics & Chinese Language Processing, Vol. 1, No. 1, pp. 37-86, 1996.

[3] Hamon, C., E. Moulines and F. Charpntier, "A Diphone Synthesis System based on Time-Domain Prosodic Modifications of Speech", ICASSP (Scotland), pp. 238-241, 1989.

[4] Gu, H. Y. and W. L. Shiu, "A Mandarin-syllable Signal Synthesis Method with Increases Flexibility in Duration, Tone and Timbre Control", Proc. Natl. Sci. Counc. ROC(A), Vol. 22, No.3, pp. 385-395, 1998.

[5] Klatt, D., "Software for a Cascade/Parallel Formant Synthesizer" , J. Acoust. Soc. Am., Vol. 67, pp. 971-995, 1980.

[6] 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.

[7] 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, pp. 287-294, 1993.

[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] Chen, S. H. and S. M. Lee, "A Statistical Model based Fundamental Frequency Synthesizer for Mandarin Speech", J. Acoust. Soc. Am., Vol. 92, No. 1, pp. 114-120, 1992.

[10 Wu, C. H. and J. H. Chen, "Prosody Generation in a Chinese TTS System based on a Hierarchical Word Prosody Template Tree", Proceedings of ROCLING X International Conference (Taipei), pp.262-266, 1997.

[11]潘能煌, 中文文句翻語音系統之音量和音調韻律研究, 中興大學應數研究所, 碩士論文, 1998.

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

[13]Ljolej, A. and F. Fallside, "Synthesis of Natural Sounding Pitch Contours in Isolated Utterances using Hidden Markov Models", IEEE trans. Acoust., Speech and Signal Processing, Vol. 34, No.5, pp. 1074-1079, Oct. 1986.

[14]Fukada, T., Y. Komori, T. Aso, and Y. Ohora, "A Study on Pitch Pattern Generation using HMM-based Statistical Information", Int. Conf. on Spoken Language Processing (Japan), pp. 723-726, 1994.

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

[16]Chou, F. C., C. Y. Tseng, and L. S. Lee, "Automatic Generation of Prosodic Structure for High Quality Mandarin Speech Synthesis", 4th Int. Conf. on Spoken Language Processing (Philadelphia, USA), pp. 1624-1627, 1996.

[17]Russell, M.J. and R.K. Moore, "Explicit Modeling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition", ICASSP (Florida), pp.5-8, 1985.

[18]Gu, H. Y., C. Y. Tseng and L. S. Lee, "Isolated-Utterance Speech Recognition Using Hidden Markov Models with Bounded State Durations", IEEE trans. Signal Processing, Vol. 39, No. 8, pp. 1743-1752, Aug 1991.

[19]Yu, S. M. and C. S. Liu, "The Construction of Phonetically Balanced Chinese Sentences", TL Technical Journal, pp. 79-86, March 1989.

[20]許文龍, 使用時間比例基週波形內差之國語語音合成器, 台灣科技大學電機系,碩士論文, 1996.