結合 ANN 預測、全域變異數匹配與真實軌跡挑選之基週軌跡產生方法

A Pitch-contour Generation Method Combining ANN Prediction, Global Variance Matching, and Real-contour Selection

古鴻炎 姜愷威 王皓
Hung-Yan Gu Kai-Wei Jiang Hao Wang

國立臺灣科技大學 資訊工程系
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
E-mail: {guhy, m10015067, m10315060}@mail.ntust.edu.tw

摘要

基週軌跡(pitch contour)對於合成出高自然度的語音信號是相當的重要的，因此本論文研究提出了一種新的基週軌跡產生方法，此方法就是把 類神經網路(artificial neural network, ANN)預測模組、全域變異數匹配(global-variance matching, GVM)與真實基週軌跡挑選(real contour selection, RCS)模組作結合，用以產生基週軌跡。在此，我們先分析出各個訓練音節的基週軌跡，然後使用離散餘弦轉換(discrete cosine transform, DCT)將各個基週軌跡轉換成對應的 DCT 係數之向量，然後就可拿各個訓練語句的 DCT 向量序列、及對應的語境參數去訓練 ANN 權重值與 GVM 參數。在基週軌跡產生的實驗中，我們以量測變異數比值(variance ratio, VR)來作為客觀評估的依據，由實驗結果得知，GVM 與 RCS 模組有助於提升 VR 值；此外，主觀聽測實驗的結果顯示，ANN 加 GVM 所產生的基週軌跡，其自然度比僅使用 ANN 模組的高，並且 ANN 加 GVM 加 RCS 的基週軌跡自然度，更高於 ANN 加 GVM 的。

關鍵詞： 語音合成，基週軌跡，離散餘弦轉換，類神經網路，全域變異數，軌跡挑選

Abstract

Pitch contours are important for synthesizing highly natural speech signal. In this paper, we study a new pitch-contour generation method. The method proposed is to combine ANN prediction module with global-variance matching (GVM) and real contour selection (RCS) modules. Here, a syllable pitch contour is first analyzed and then transformed via discrete cosine transform (DCT) to a DCT-coefficient vector. Each sequence of DCT vectors analyzed from a training sentence plus contextual parameters are then used to train the ANN weights and GVM parameters. In pitch-contour generation experiments, we measure variance-ratio (VR) values for objective evaluations. The modules, GVM and RCS, are shown to be helpful to promote VR values. In addition, in subjective evaluation, the pitch-contour generation method, ANN + GVM, is shown to be more natural than the method, ANN only. Also, the method, ANN + GVM + RCS, is shown to be better than ANN + GVM.

Keywords: speech synthesis, pitch contour, discrete cosine transform, artificial neural network, global variance, contour selection.

一、緒論

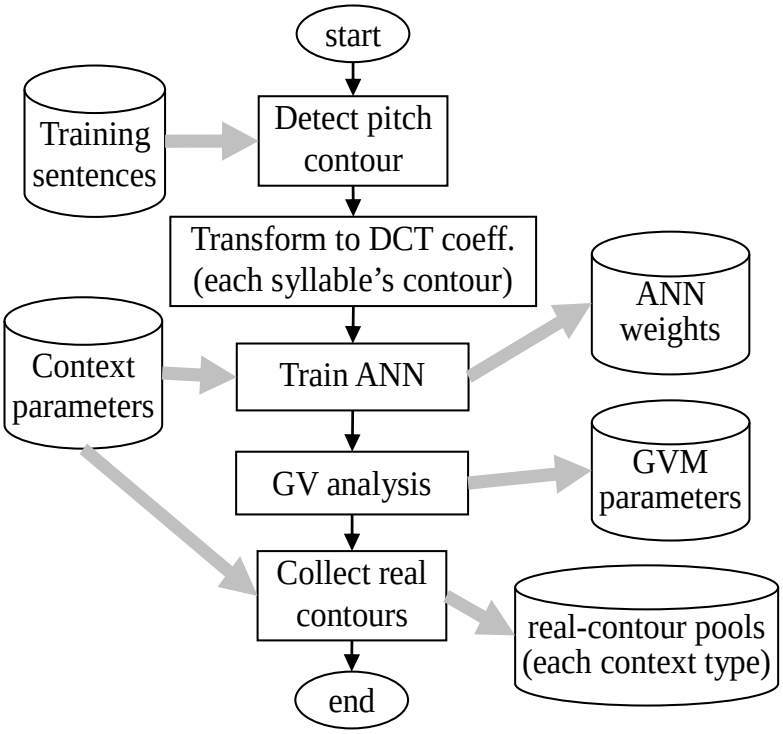
一個合成語音信號的自然度主要是由韻律參數(如基週軌跡、音長、音量等)所決定，其中基週軌跡對於自然度之提升更顯得重要，因此，過去已有許多不同的音節基週軌跡產生方法被先前的研究者所提出[1, 2, 3, 4, 5, 6]。目前，隱藏式馬可夫模型(hidden Markov model, HMM)雖然已被許多人採用於作語音合成的研究[7, 8]，然而 MSD-HMM (multi-space probability distribution HMM)產生出的基週軌跡並不十分地令人滿意，這種情形已有不少人注意到[3, 6]。

我們覺得基週軌跡之產生，並不需要和頻譜係數之產生綁在同一種機制(即 HMM)裡，並且我們想要進一步提升所產生出的基週軌跡的自然度，因此在本論文中，我們嘗試研究、提出一種把類神經網路(artificial neural network, ANN)預測[1, 2]、全域變異數匹配(global-variance matching, GVM)與真實軌跡挑選(real contour selection, RCS)三者作結合的方法，希望用以提升合成語音的自然度。

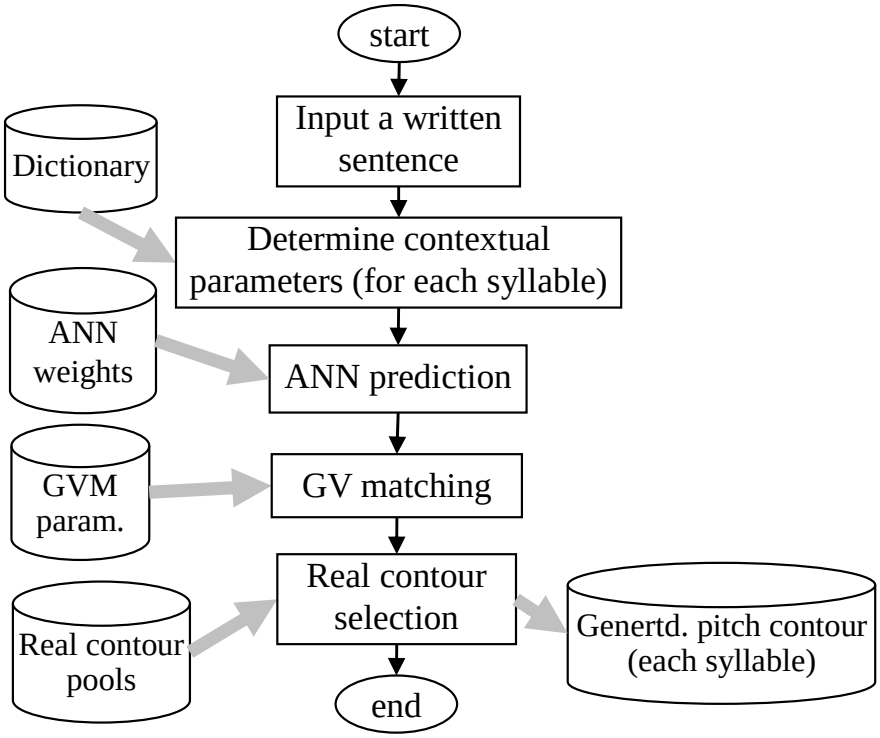
過去 Toda 與 Tokuda 提出 GVM 之作法[9]，來對 HMM 產生的頻譜係數作調整，以減緩發生頻譜過度平滑(spectral over-smoothing)的現象，而藉以提升合成語的音質。在此，我們發現到 ANN 產生的表示基週軌跡的 DCT (discrete cosine transform)係數，也同樣會發生過度平滑(over smoothing)的現象，因此我們覺得對 ANN 產生的基週軌跡 DCT 係數，作 GVM 匹配將有助於提升 ANN 基週軌跡的自然度。此外，我們也受到另一個觀念的啟發，就是語音轉換(voice conversion)領域中前人提出的，以挑選目標語者音框的真實頻譜係數來取代轉換出的頻譜係數，如此用以改進轉換出語音的音質[10]。因此，我們認為把 ANN 產生並且經過 GVM 匹配的 DCT 係數向量 X 作為參考，而據以選出一個最靠近 X 的真實基週軌跡 DCT 係數向量 Y，然後把 Y 拿去取代 X，如此將可更進一步提升所產生的基週軌跡的自然度。關於 RCS 的實作，我們可依據各個音節的語境資料來作語境的分類，然後把屬於不同語境分類的各個真實基週軌跡 DCT 向量，分別放入不同的收集區(pool)裡。

整體來說，我們系統在訓練階段的處理流程如圖一所示。首先對每個錄音語句的的各個音節作基週軌跡分析；接著，把各個音節的基週軌跡轉換成固定維度的 DCT 係數向量；然後拿各個訓練語句的 DCT 向量序列及各音節對應的語境資料，去訓練 ANN 為基礎的基週軌跡產生模型。除了訓練 ANN 模型之外，我們也對各個訓練語句的 DCT 向量序列作分析，以求得 GVM 匹配所需的參數。此外，我們依據各個音節的語境分類，把它的基週軌跡 DCT 向量放入對應的收集區裡。

另一方面，產生基週軌跡的整體流程如圖二所示。首先輸入一個文句，接著經由搜尋詞典來確認各個中文字的音節發音與音調；依據查詢出的一序列音節發音與音調，就可為各音節準備它對應的語境參數，然後將各音節的語境參數輸入 ANN 模型，去預測該音節的基週軌跡(即 DCT 係數)；對於 ANN 預測出的基週軌跡，接著使用訓練階段儲存的全域變異數(GV)參數去對 DCT 係數進行 GVM 匹配；之後，依據 GVM 匹配調整過的基週軌跡，我們從訓練階段建立的、且和目前音節具有相同語境類型之真實基週軌跡收集區中，去找出最接近 GVM 匹配過之基週軌跡的一個真實基週軌跡。



圖一、基週軌跡模型之參數訓練的主流程



圖二、基週軌跡產生階段之主流程

二、模型參數訓練

如圖一所示，我們需要訓練 ANN 模型的權重，分析出 GVM 匹配所需的參數，及分別儲存不同語境類型的真實基週軌跡(即 DCT 係數向量)。

（一）、語句錄音與基週軌跡偵測

在此研究中，我們邀請了一位男性語者於錄音室中錄製了 810 句語句，而總音節數為 7,161 個音節。在錄音之後，先以 HTK (HMM toolkit) 軟體進行自動標音，再使用 WaveSurfer 軟體來對各音節的時間邊界作人工微調。

關於音節基週軌跡之偵測，我們使用 HTS (HMM-based speech synthesis system)軟體內含的 SPTK (Speech Signal Processing Toolkit)模組[8]來進行，並且設定音檔的取樣率設為 22,050 Hz，而音框位移則設為 110 個樣本點。在自動偵測基週軌跡之後，我們發現有許多音框所偵測出的基頻值是錯誤的，例如一個有聲(voiced)音框的基頻值可能被偵測為 0，即誤判為無聲(unvoiced)，或是被偵測成真實頻率的一半或兩倍的情形。因此我們撰寫了一個工具程式，來對偵測錯誤的基週軌跡作半自動或是手動的更正處理。

（二）、離散餘弦轉換

由於一個語句中各音節的基週軌跡長度不一，長度可能介於 30 至 80 個音框之間，為了把基週軌跡表示成固定維度數的資料，我們選擇以離散餘弦轉換(DCT)之係數來表示各音節的基週軌跡。至於維度數量之選擇，在比較過多種維度數之 DCT 轉換與反轉換回來的基週軌跡曲線後，我們決定將維度數設為 24。一個原始的基週軌跡、和 DCT 反轉換回來之曲線例子如圖三所示，我們覺得 16 階 DCT 反轉換所得之曲線，仍不夠忠實於原始曲線。

詳細來說，本研究裡使用的是 DCT-I 之離散餘弦轉換[11]，其正向轉換之公式為：

$$c(m) = x(0) + (-1)^m \cdot x(N-1) + 2 \cdot \sum_{k=1}^{N-2} x(k) \cdot \cos\left(\frac{m \cdot k \cdot \pi}{N-1}\right), \quad m = 0,1,...,23 \tag{1}$$

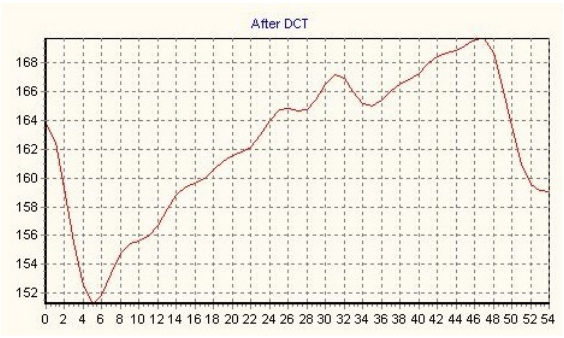
其中 $x(k)$表示一個音節基週軌跡的第 k 個音框的基頻值(以 Hz 為單位)，$c(m)$表示 DCT 轉換後的第 m 階係數，而 N 則是該音節的音框數。

對於公式(1)，其對應的 DCT 反轉換公式為：

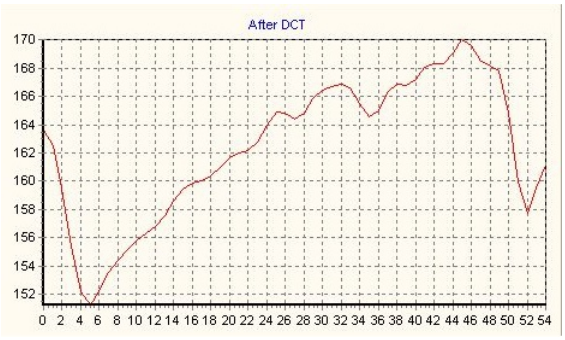
$$x(k) = \frac{1}{2(N-1)}\left[c(0) + (-1)^k \cdot c(M-1) + 2 \cdot \sum_{m=1}^{M-2} c(m) \cdot \cos\left(\frac{k \cdot m \cdot \pi}{M-1}\right) \right], \quad k = 0,1,...,N-1 \tag{2}$$

其中 M 表示 DCT 轉換的維度數，在此 M 設為 24。

(a) 原始之基週軌跡曲線



(b) 16 階 DCT 反轉換之曲線

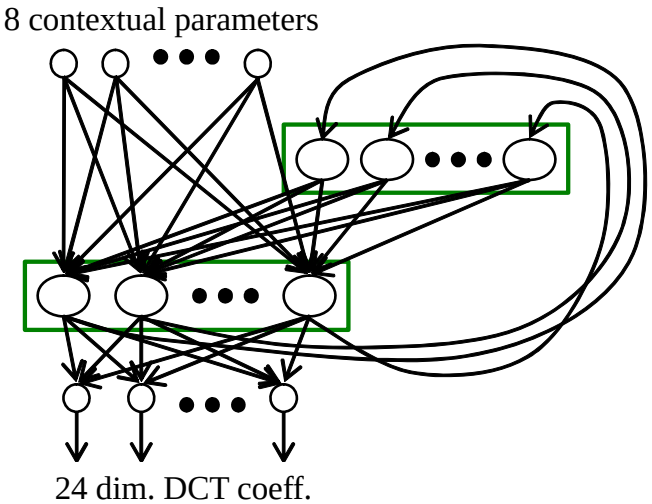


(c) 24 階 DCT 反轉換之曲線

圖三、原始與 DCT 反轉換之基週軌跡曲線

（三）、ANN 模型訓練

在此我們設計使用的 ANN 結構如圖四所示，就如同前人所設計的 ANN 結構[1, 2]，它是一種遞迴式的類神經網絡，輸入層有 28 個節點來輸入 8 種語境參數，輸出層則有 24



圖四、本研究設計之 ANN 結構

個節點來輸出 24 維的 DCT 係數。前述的 8 種語境參數包括: (a)前一個音節的聲調和韻

母類別資料；(b)目前音節的聲調、聲母和韻母資料；(c)後一個音節的聲調和聲母類別資料；(d)目前音節在句子內的位置序數。關於聲母與韻母的分類方式、及 8 種語境參數如何編碼成 28 個輸入節點，其詳細的作法可參考我們先前發表的論文[12]。此外，關於隱藏層中應放置的節點數，在此我們依實驗量測出的平均預測誤差值來決定，所測試的節點數從 12 變化到 20，而使用的語料則是前 750 個語句，結果發現把節點數設為 16 是最好的選擇。

（四）、GVM 參數分析

GVM 匹配原本被提出來對一序列語音音框的頻譜係數作調整[9]，然而在這裡，我們將改成以音節為單位，這是因為一個音節的基週軌跡在此僅以一個 24 維的 DCT 向量作表示。假設一個語句的長度在 4 到 20 個音節之間，則將只有 4 到 20 個 DCT 向量可用來計算該語句基週軌跡 DCT 向量之各維度的變異數，若要估計第 k 語句之 DCT 向量第 j 維的變異數，我們使用的公式為：

$$v_i^k = \left[\sum_{j=1}^{n(k)} (c_i^k(j) - m_i^k)^2 \right] \Big/ n(k), \tag{3}$$

其中 $n(k)$表示第 k 語句裡的音節個數，$c_i^k(j)$ 表示第 j 個音節基週軌跡 DCT 向量之第 i 維的係數，而 m_i^k 表示 $c_i^k(j)$, j=1, …, $n(k)$的平均值。

如此，橫跨 750 句訓練語句的 DCT 向量第 i 維之全域變異數，就可以公式(4)來作計算：

$$g_i = \frac{1}{N} \sum_{k=1}^{N} v_i^k, \tag{4}$$

其中 N 表示訓練語句的個數(在此是 750 句)，g_i 表示估計出之第 i 維的全域變異數。

（五）、真實基週軌跡之收集

若要實現真實軌跡之挑選，則在訓練階段裡，我們必須為每一種類型的語境(context)組合去收集屬於該類型語境組合的真實基週軌跡。那麼，如何去定義語境類型呢? 首先我們考慮的是一個熟知的現象，就是一個語句的音調(intonation)會隨著音節在句子裡的位置而發生音高下傾(pitch declining)之現象，因此我們就粗略地把每一語句的組成音節分割成三個片段，如此可推得落在最前片段的音節將會有較高的音高，而落在尾部片段的音節，其音高將會較低。

其次，我們認為影響音節基週軌跡之形狀與高度的一個主要因素是，本音節和它鄰接的前後兩音節的聲調組合。假設一個語句裡第(j-1)個音節的發音聲調編號為 P_{j-1}，第 j 個音節的聲調編號為 P_j，且第(j+1)個音節的聲調編號為 P_{j+1}，再者國語共有五種聲調，所以

第 j 個音節的聲調組合索引值，在此的計算方式訂為 $25\times P_{j-1} + 5\times P_j + P_{j+1}$，如此可被組合出的聲調組合索引值會有 125 種，如果一個音節之前面或後面沒有連接其它音節，則其前接或後接音節的聲調，在此就直接定義為輕聲。

考慮前述的兩個因素，在此便訂定出 3 (片段) × 125 (聲調組合) = 375 種語境組合的類型，因此我們設置了 375 個收集區來分別收集所屬的真實基週軌跡 DCT 向量。對於 750 句的訓練語句，各語句裡各音節的基週軌跡 DCT 向量，便可依該音節的語境組合編號，將它的基週軌跡 DCT 向量放入對應的收集區中。

三、基週軌跡產生與實驗評估

(一)、基週軌跡產生

依據圖二之流程，對於一個輸入的國語文句，我們會先查詢出它的一序列音節發音和聲調，然後就可依序把各音節的語境參數餵入 ANN 模組，以預測出 24 維的 DCT 係數所代表的基週軌跡。當所有音節的基週軌跡 DCT 向量都預測出來後，接著就對各個音節進行 GVM 匹配之處理，藉以求得起伏較明顯的基週軌跡曲線。在此，作 GVM 匹配所用的公式為：

$$\hat{c}_i = (c_i - m_i)\left[(w\cdot\sqrt{g_i / v_i}) + 1\right] + m_i \ , i = 1, ..., 23, \qquad (5)$$

其中 c_i 表示 ANN 預測出的 DCT 向量的第 i 維係數；m_i 與 v_i 分別表示第 i 維係數的平均值與變異數，它們是依據各語句中全部音節的 c_i 去計算出來的；g_i 是依公式(4)所算出的第 i 維全域變異數；w 表示匹配強度的權重值，其值介於 0 到 1 之間。在此，我們僅對 1 至 23 維之 DCT 係數作 GVM 匹配，因為 ANN 所產生的第 0 維 DCT 係數 c_0，作 GVM 匹配並不會影響到基週軌跡的形狀，反而是會影響該音節之音高水平高度。

經過 GVM 處理之後，接著進行真實軌跡之挑選。令 X_j 表示一語句經過 GVM 匹配後的第 j 個音節的 DCT 向量，在此先依據第 2.5 節所說明的方式，去決定 X_j 為屬於本語句的前、中、後片段的那一段，及計算第 j 個音節與前後鄰接音節的聲調組合，然後計算出 X_j 所對應的語境類型編號 T_j。接著，搜尋編號 T_j 之收集區中的真實基週軌跡 DCT 向量，以找出幾何距離上與 X_j 最接近的真實軌跡 DCT 向量 Y_j，然後拿 Y_j 來取代 X_j。

圖二中 ”GV matching” 與 ”Real contour selection” 這兩個區塊，我們欲研究它們所能發揮的效用，因此我們接著實驗了六種不同的基週軌跡產生方法，這些產生方法的差別為：前述的兩個區塊有否被包含進去，以及在於公式(5)中設定使用不同的權重值 w。以下我們以符號 MA、MB、MC、MD、ME 與 MF 來代表這 6 種基週軌跡產生方法，它們的細節設定是：

MA：只使用 ANN 而不使用 GVM 與 RCS；
MB：使用 ANN 和 GVM，且設定 w=0.33，但不使用 RCS；
MC：使用 ANN 和 GVM，且設定 w=0.5，但不使用 RCS；
MD：使用 ANN 和 RCS，但不使用 GVM；

ME：使用 ANN、GVM 和 RCS，且設定 w=0.33；
MF：使用 ANN、GVM 和 RCS，且設定 w=0.5；

(二)、客觀評估

在內部測試時，我們仍然使用訓練 ANN 模型與 GVM 參數的前 750 個語句，來量測原始語音分析出的基週軌跡 DCT 向量和程式產生的基週軌跡 DCT 向量之間的幾何距離、及兩者之間的變異數比值(variance ratio, VR)；而在外部測試時，則僅拿未在訓練階段使用的剩餘之 60 個語句來作量測。在量測內部語句的幾何距離平均誤差之後，我們發現前述的六種基週軌跡產生方法之間並沒有明顯的數值差異，詳細的幾何距離平均誤差數值如表一所示；因此我們就改成採取前人提出的原用於比較轉換出語音(voice conversion)

方法	MA	MB	MC	MD	ME	MF
平均誤差	2.066	2.072	2.081	2.072	2.075	2.080

表一、基週軌跡 DCT 向量之幾何距離平均誤差

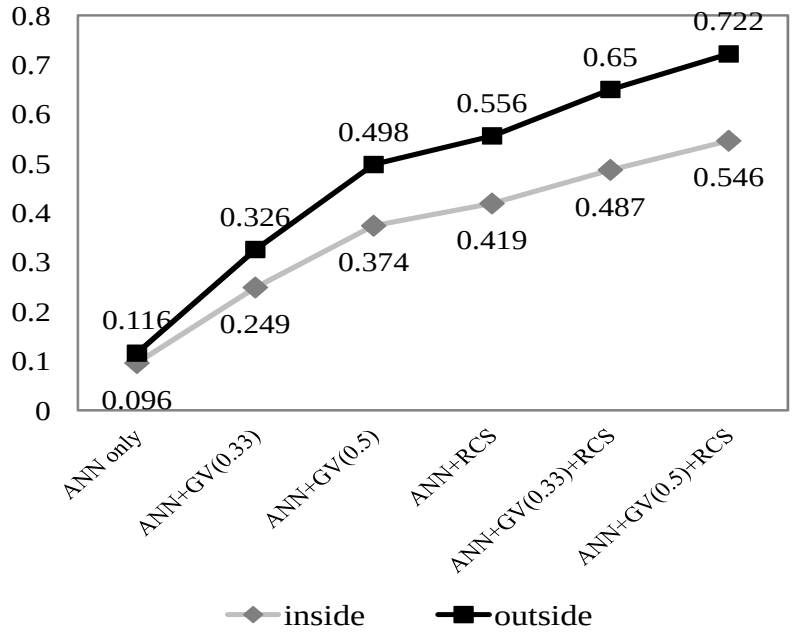
品質之變異數比值(VR)量測[13]，來比較這六種基週軌跡產生方法，變異數比值的計算公式為：

$$VR = \frac{1}{L}\sum_{k=1}^{L}\frac{1}{D}\cdot\sum_{d=1}^{D}\frac{\hat{\sigma}_k^d}{\sigma_k^d} \quad , \tag{6}$$

其中 L 表示國語韻母的類別數(在此 L=36)； D 表示基週軌跡 DCT 向量的維度數；$\hat{\sigma}_k^d$ 表示程式產生出的基週軌跡 DCT 向量之中，把屬於第 k 類韻母之 DCT 向量第 d 維的係數拿去計算出的變異數；σ_k^d 則表示原始音節語料分析出的基週軌跡 DCT 向量之中，把屬於第 k 類韻母之 DCT 向量第 d 維的係數拿去計算出的變異數。需注意的是，在此 D 的值為 23，因為 ANN 產生的 DCT 係數 c_0，我們並未對它作 GVM 調整，也未把它取代成 RCS 選出的 DCT 向量之 c_0。

我們把前述六種基週軌跡產生方法輸出的 DCT 向量，分別帶入公式(6)作 VR 值的計算，然後把 VR 值畫成圖五。根據量測出的 VR 值可發現，若只使用 ANN 來產生基週軌跡 DCT 向量(即方法 MA)，則量得的 VR 值會很低，約在 0.1 附近。但是，如果在 ANN 產生出基週軌跡 DCT 向量之後，再拿 DCT 向量去作 GVM 匹配(即方法 MB 或 MC)、或 RCS 挑選(即方法 MD)，則量得的 VR 值都會有顯著的提升，這表示基週軌跡 DCT 係數之過度平滑現象顯著減少，理論上可以使基週軌跡得到更高的自然度。更進一步，如果在 ANN 產生出的基週軌跡 DCT 向量之後，接續作 GVM 調整和 RCS 挑選(即方法 ME 或 MF)，則量得的 VR 值會更為提升。圖五中的兩條曲線分別代表拿內部或外部語料去作 VR 值量測所得到的結果，由這兩條曲線可看出，兩曲線的變化趨勢都與前面說明的現象具有一致性，因此， GVM 調整和 RCS 挑選可以有效地改進基週軌跡 DCT 係數過

於平滑的現象，而可讓基週軌跡的自然度獲得提升。



圖五、不同基週軌跡產生方法之 VR 量測值折線

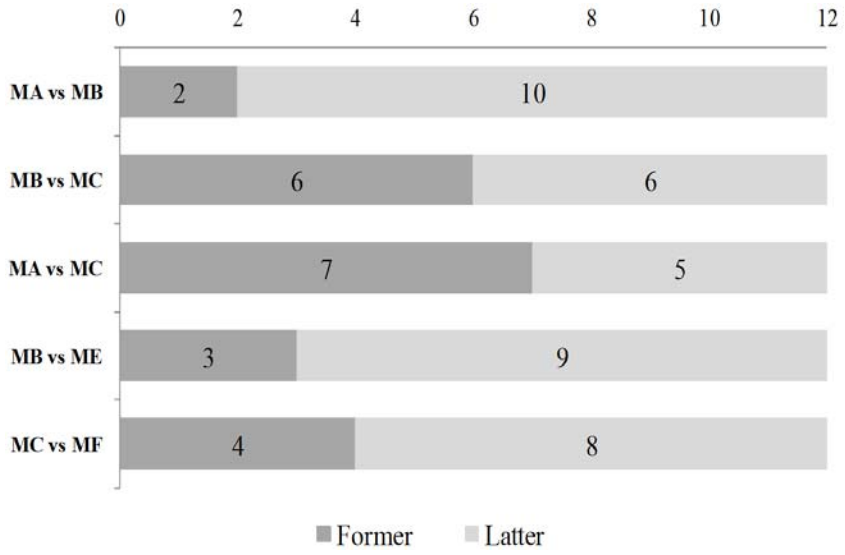
(三)、主觀評估

一個基週軌跡產生方法產生出的軌跡，各音節的聲調必須聽起來正確無誤，並且軌跡曲線本身也要有明顯的抑揚變化，如此才能讓人聽起來具有較高的自然度。在此我們針對前述的六種產生方法進行主觀聽覺之測試，共邀請了兩組人士來參加聽測，第一組的 11 人具有語音處理的研究背景，第二組的 11 人則無語音處理之經驗。

我們隨機選取了三篇短文文句，然後使用前述六種產生方法(MA 至 MF)去對各篇文句產生出基週軌跡，接著再和其它韻律參數(音節音長和音量)作組合，以帶入語音信號合成模組[14]，來對各篇文句合成出 6 個語音信號檔，各對應於六種基週產生方法之一。此外，我們也把產生出的基週軌跡從男聲的音高轉換成女聲的音高，這可透過語音轉換領域常用的音高轉換方法[10，13]，然後把轉換過的基週軌跡送給先前以女聲錄音所訓練出的 HMM 頻譜模型，去合成出女聲基週軌跡的音檔，以便對不同性別的基週軌跡作聽測，讓聽測實驗能夠兼顧性別而更具有一般性。如此，對於每一種基週軌跡產生方法來說，都會有 6 個合成出的語音音檔(3 篇短文 × 2 種音高)。

在此聽測實驗的進行方式是，每次播放兩個合成語音的音檔給受測者聽，以比較兩音檔的自然度，然後請受測者打一個分數，來顯示那一個音檔比較自然。打分數的規則是，如果前者(後者)的自然度明顯高於後者(前者)，則給予 1 分(5 分)，如果前者(後者)僅比後者(前者)稍好一點，則給予 2 分(4 分)，如果無法區分出兩者的自然度優劣則給予 3 分。

如果要把六種產生方法兩兩作組合去作聽測實驗，則需要進行 15 組的聽測實驗，將會非常花費人力，因此我們在此只選擇其中五組來進行聽測實驗，也就是(a)MA 比 MB、(b)MB 比 MC、(c)MA 比 MC、(d)MB 比 ME、和(e)MC 比 MF。對於每一組方法的比較，每一個受測者須依序聽取兩個產生方法所合出的 6 對音檔，並且給 6 對音檔分別打分數。在聽測實驗結束之後，我們區分受測者所隸屬的組別、並且對 6 對音檔分別收集評分，然後分別計算出平均評分。在此，我們將平均評分視為一種投票，當平均評分小於 3 時，就給聽測時先播放之音檔對應的產生方法增加一票，而當平均評分大於 3 時，就給聽測時後播放之音檔對應的產生方法增加一票。由於每個產生方法都有 6 個合成音檔，並且受測者分成兩組(各 11 人)，所以每一組產生方法之比較總共有 12 張票，統計投票結果後，5 組作自然度聽測比較的基週軌跡產生方法，各自所得到的票數就如圖六所示。



圖六、5 組產生方法作聽測比較之投票結果

根據圖六所顯示的投票結果，我們可看出 MA 比 MB 的票數為 2 比 10、MB 比 ME 的票數為 3 比 9、並且 MC 比 MF 的票數為 4 比 8 。所以我們可說，方法 MB (ANN 加 GVM)產生出的基週軌跡要比方法 MA(只使用 ANN)的更為自然，此外從 MB 比 ME 和 MC 比 MF 這兩組方法的得票數結果，我們可說使用 RCS (真實軌跡挑選)，可更為提升自然度。另一方面， MB 比 MC 的票數為 6 比 6，而 MA 比 MC 的票數為 7 比 5，所以在自然度上，方法 MB 和 MC 之間並沒有顯著的差別，也就是 GVM 處理的權重值並不會造成明顯的差別。

四、結論

我們發現 ANN 產生的基週軌跡 DCT 係數存在有過平滑(over smoothing)的現象，因此在本論文中，我們嘗試於 ANN 預測模組之後再串接兩種處理模組，即 GVM 和 RCS，以設法提升 ANN 產生之基週軌跡的自然度。對不同產生方法作客觀評估時，我們採取以計算 VR 值來反映過平滑的程度，依據量測出的 VR 值結果，我們發現 GVM 和 RCS 模組兩者都能明顯地提升 VR 值，因此 GVM 和 RCS 兩種處理動作確實都有助於緩和基週軌跡 DCT 係數之過平滑問題，並且當把 GVM 與 RCS 串接起來作處理時，更能夠進

一步提升 VR 值。

另外在主觀評估方面，我們進行了聽測實驗，來比較五組基週軌跡產生方法的自然度。在聽測實驗之後，把受測者所給的評分依據受測者的組別和所聽的音檔，分別作收集再計算平均評分，然後把各個平均評分值當作對兩聽測音檔之自然度比較的投票。統計票數後，我們發現方法 MB (ANN 加 GVM)的票數明顯高於方法 MA (只使用 ANN)；此外，方法 ME 的票數高於 MB，且方法 MF 的票數高於方法 MC，所以 RCS (用於方法 ME 與 MF)確實可有效地提高所產生之基週軌跡的自然度。

參考文獻

[1] S. H. Chen, 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.

[2] C. T. Lin, 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, 2004.

[3] C. C. Hsia, C. H. Wu, and J. Y. Wu, “Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in HMM-based speech synthesis”, IEEE trans. Audio, Speech, and Language Processing, Vol. 18, No. 8, pp. 1994-2003, 2010.

[4] H. Y. Gu and C. C. Yang, “An HMM based pitch-contour generation method for Mandarin speech synthesis”, Journal of Information Science and Engineering, Vol. 27, No. 5, pp. 1561-1580, 2011.

[5] M. Dong, K. T. Lua, “Pitch contour model for Chinese text-to-speech using CART and statistical model,” Int. Conf. on Spoken Language Processing, Denver, USA, pp. 2405-2408, 2002.

[6] L. Gao, Z. H. Ling, L. H. Chen, and L. R. Dai, “Improving F0 prediction using bidirectional associative memories and syllable-level F0 features for HMM-based Mandarin speech synthesis”, Proceeding of ISCSLP, Singapore, pp. 275-279, 2014.

[7] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM based speech synthesis”, Proceeding of EUROSPEECH, Budapest, Hungary, pp. 2347-2350, 1999.

[8] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, “The HMM-based speech synthesis system (HTS) version 2.0”, Proceeding of 6-th ISCA Workshop on Speech Synthesis, Bonn, Germany, pp. 294-299, 2007.

[9] T. Toda and K. Tokuda, “A speech parameter generation algorithm considering global variance for HMM-based speech synthesis”, IEICE trans. INF. & SYST., Vol. E90–D, No. 5, May 2007.

[10] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, and Y. Stylianou, “Towards a voice conversion system based on frame selection”, Int. Conf. Acoustics, Speech, and signal Processing, Honolulu, Hawaii, pp. 513-516, 2007.

[11] A. V. Oppenheim and R. W. Schafter, Descrete-time Signal Processing, second ed., Prentice-Hall, 1999.

[12] H. Y. Gu, Y. Z. Zhou, and H. L. Liau, “A system framework for integrated synthesis of Mandarin, Min-nan, and Hakka speech”, Int. Journal of Computational Linguistics and Chinese Language Processing, Vol. 12, No. 4, pp. 371-390, 2007.

[13] E. Godoy, O. Rosec, and T. Chonavel, “Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora”, IEEE trans. Audio, Speech, and Language Processing, Vol. 20, No. 4, pp. 1313-1323, 2012.

[14] H. Y. Gu, M. Y. Lai, and W. S. Hong, “Speech synthesis using articulatory-knowledge based HMM structure”, Int. Conf. on Machine Learning and Cybernetics, Lanzhou, China, pp. 371-376, 2014.