用於英語語音合成之聲調預測與基週軌跡產生方法

Tone Prediction and Pitch Contour Generation Methods for English Speech Synthesis

古鴻炎 (Hung-Yan Gu)
國立台灣科技大學資訊工程系
guhy@mail.ntust.edu.tw

陳忠緯 (Chung-Wei Chen)
國立台灣科技大學資訊工程研究所
M9515053@mail.ntust.edu.tw

摘要

本論文研究了英語語音合成相關之聲調預測、基週軌跡產生的問題。關於英語音節的聲調預測，我們提出一個基於最大可能性之預測方法，而在實施的作法上，研究以動態規劃來加快尋找最佳的聲調序列，並且基於 PPMC 逃脫機率估計法來估計局部機率，再對 PPMC 法作了改良。關於音節基週軌跡之產生，我們研究以半音節之分類作為語境資料，用以建立 ANN 基週軌跡產生模型。此外，我們採用規則式作法來設定音節音量與音長，語音信號合成則是採用 HNM 合成法，目前已建造出一個英語的語音合成系統，並且用它合成出的語音來進行聽測實驗，聽測的評分顯示，當聲調預測的正確率愈高時，所合成語音的自然度就會愈好。

關鍵詞：語音合成、聲調預測、基週軌跡。

1. 前言

目前台灣有許多單位正在進行智慧型機器人的研究，我們考慮到智慧型機器人未來在國際上，需具有"與人對話"、"表演脫口秀"等功能，因此我們以先前研究英語歌聲合成的經驗為基礎[2]，來進行英語語音合成的研究。

英語與國語之間存在多項不同的特性，例如音節的組成、音節的數量、聲調的表示方式等，都是有差異的，其中我們認為聲調的表示及預測會是影響英語語音合成的最重要的因素。國外研究英語合成的學者，如 R. A. J. Clark [13]，也是認為聲調預測及音高軌跡產生是相當重要的，所以可查到許多語調模型(intonation model)方面的研究文獻，比較有名的如 ToBI [9]、Tilt [12]和 IPO [7]。

一個知名的英語語音合成系統是，愛丁堡大學的 Festival 系統[14]，Festival 在產生語調(intonation)時會先預測出音高事件(pitch event)，而音高事件可分為二個種類，一類是 pitch accent (如 H*, L*, L*+H, H*+L, …)，另一類為 boundary tone (如 H%, L%)。因此，Festival 使用了二個分類器來作音高事件的預測，而進行預測實驗後得到的結果是，對於出現最多的 pitch accent 的預測正確率可達 70%，但是出現少的 pitch accent 的預測正確率就很差了，例如訓練語料中共出現了 176 次的 L* accent，但是它的預測正確率卻為 0%；另外，出現最多次的 boundary tone，其預測正確率可達 90%，但是對於另一個 boundary tone 的預測正確率就相當低。

如圖 1 所示，一個英語句子在作語音合成之前，必需先作文字至音節的轉換，再作聲調(或音高事件)的預測，然後再依預測出的聲調去作音高軌跡(pitch contour)的產生，接著才能作信號波形的合成。在本論文裡，我們的研究焦點是英語音節的*聲調預測*，及*音高軌跡產生*；而對於*音節轉換*和*信號波形合成*，則採用前人發展出的程式模組[2]，再做必要的修改。

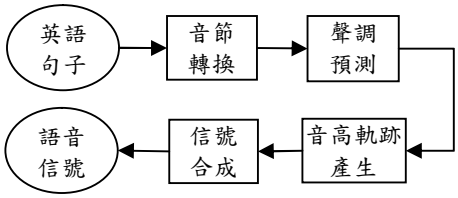

圖1 英語語音合成之主要流程

關於聲調或音高事件的預測，首先我們嘗試以漢語聲調的定義方式，來把英語音節的音高事件定義成數種聲調型式，而不沿用國外的音高事件之定義方式。至於聲調的預測，我們提出了一種基於 PPM (prediction by partial matching)機率估測模型的最大可能性(maximum likelihood)預測法(MLP)，此外，我們也使用了知名的分類樹軟體 Weka [11]，來對聲調預測的準確率進行實驗，以檢驗所提出的方法。關於音高軌跡(pitch contour)的產生，我們依據先前研究國語音高軌跡產生的經驗[6]，採用類神經網路模型來作音高軌跡的產生。

此外，我們也依據圖 1，來進行系統的實作，希望發展出一個可作即時英語語音合成之系統，此系統的發展，可分成訓練和合成兩個階段，這兩階段的處理流程分別如圖 2 與 3 所示。在訓練階段，我們先將英語語料庫裡的語句作標音(標記音標符號)、切音處理，之後再對各音節標記聲調符號、及分析音高軌跡；音高軌跡分析時，先對各音節進行基週偵測(pitch detection)，然後取 16 個正規化時間點上的音高頻率值來代表一個音節的音高軌跡；接著，拿標記好的資料以及音高軌跡資料去作相關模型的訓練。在合成階段，首先讀入一個文句，接著查詢詞典以轉換出音節序列；接著使用最大可能性預測法，預測出各音節的聲調；之後，再

依所預測出的聲調、搭配前後文句資料來產生各音節的音高軌跡；另外，採規則式作法來決定各音節的音長、音量；最後呼叫 HNM (harmonic plus noise model)信號產生模組，以合成出英語語音的信號波形。

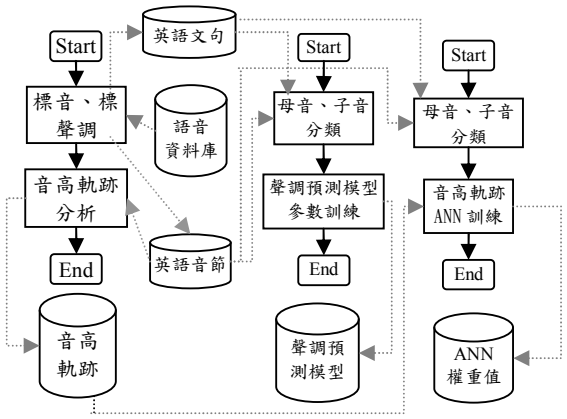

圖 2 訓練階段之處理流程

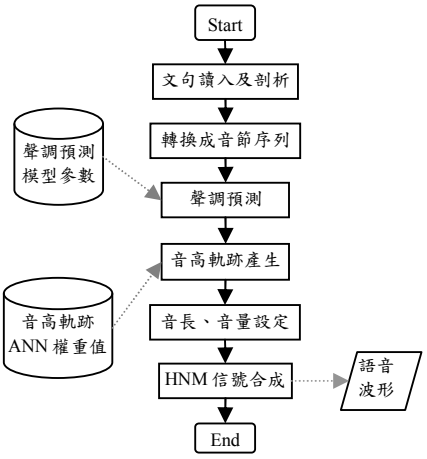

圖 3 合成階段之處理流程

2. 語料收集與分析

我們從數期的英語會話教材『空中英語教室』中，取出同一位女士發音的 35 篇文章的 MP3 音檔，然後把這些音檔的規格轉換為取樣率 22,050Hz，16bits/sample。這些音檔共有 8,106 個音節，我們把其中 7,350 個音節作為模型訓練的語料，剩餘的 756 個音節則作為測試語料。

關於音檔的標音，我們使用 WaveSurfer 軟體來對每一個音節作逐一的標記，標記了音節的邊界、拼音符號、及聲調編號。拼音符號使用的是 CMU 音標代碼[4]，而我們訂定的聲調編號則如表 1 所示。依據表 1 的聲調定義，我們對語料中各音節以手動方式標記聲調編號，聲調的判定是由人耳聆聽及觀察所分析出的音高軌跡。

在音節的組成結構上，國語的主要結構是聲母加上韻母(即 C + V)，但是英語音節的組成有許多是 C + V + C 的情況，其中 C 可以為多個子音或無子音，如此可能組合出的音節數量會是很多的，依據 CMU 詞典可知音節數量大於 15,000。當考慮音高軌跡模型所需使用的語境參數時，如果以音節作為語境參數之單位，則可能組合出的不同語境的數量將會相當龐大。

表 1 聲調定義表

英語聲調	調形	相似調形之國語聲調
1	高平調	1
2	上升調	2
3	低平調	3
4	下降調	4
5	中平調	
6	高短調	5
7	低短調	5

因此我們採取的一種簡化的語境單位是音節核心母音，它只有 16 類而已。不過在基週軌跡模型裡，考慮到子音的影響，所以又採取另一種簡化方式，就是將音節切割成前、後半音節(demi-syllable)之語音單位，並且對可能的半音節組合作分類，以兼顧語料不足問題、和語境參數的精細度。一個語境的例子如: “my leg and” 這 3 個 word 的音標為 / m ay /、/ l eh g /、/ ae n d /，當考慮 / l eh g / 這個音節時，它的語境參數為/-ay/ (前一個音節的後半)、/leh-/ (本音節的前半)、/-ehg/ (本音節的後半)、/ae-/ (後一個音節的前半)。在此我們把前半音節分成 26 類，另外把後半音節分成 31 類，詳細的分類方式請參考第二作者的碩士論文[3]。

3. 音節聲調之預測

因為英語詞典中並沒有聲調資料可供查詢，所以我們必需先預測出一個句子裡各音節的聲調調號，然後才能把聲調資料帶入基週軌跡 ANN 模型，以產生出各音節的基週軌跡。

當我們把訓練語料中手動標記的聲調資料，依各音節的母音類別作統計，結果得到如表 2 之聲調分佈情形。以母音 ax 來看，7 種聲調中除了聲調 4 之外，其它聲調出現的機率都非常高，因此要正確預測 ax 是發那一種聲調是有難度的。另外，觀察母音 oy 的聲調分佈，則只有 3 種聲調可能出現，其中聲調 2 的出現機率又比其它聲調的機率高許多，因此較容易預測出正確的聲調。

3.1 最大可能性預測法

當一個輸入的英語句子經由查詞典後，可得到一序列的音節, $S = S_1, S_2, \cdots, S_N$，依據 S，我們希望找出機率上最可能的一個對應的聲調序列，$T^* = T_1^*, T_2^*, \cdots, T_N^*$，也就是要從各音節聲調 T_k, k=1, 2, …, N 的所有可能的組合中，找出最可能的一個聲調序

列，其公式為:

$$T^* = \underset{T=T_1, T_2, .., T_N}{arg\,max}\ \mathrm{P}(T \mid S) = \underset{T_1, T_2, .., T_N}{arg\,max}\ \mathrm{P}(T, S)\ . \quad (1)$$

表 2 訓練語料之聲調分佈

聲調 / 母音	1	2	3	4	5	6	7
ae	66	217	145	48	82	4	7
eh	26	301	127	58	60	0	1
ih	225	320	309	29	108	67	58
aa	30	204	98	40	47	1	0
ax	325	200	433	6	134	182	210
ah	29	99	45	13	15	2	1
ao	12	157	61	33	31	0	0
uh	11	27	17	9	9	0	4
iy	131	201	157	33	64	17	18
ey	15	187	77	39	37	1	1
ow	26	130	42	16	16	0	0
uw	28	125	36	27	39	57	66
er	105	156	142	27	35	7	6
ay	36	178	78	45	52	0	0
oy	0	13	2	1	0	0	0
aw	10	71	27	19	11	0	0

實作上需要再考慮的是，機率項 P(T, S)應如何定義，以便其可符合實作的需求；另外，需考慮的是，如何從所有可能的聲調組合中作搜尋，以便降低計算量。關於 P(T, S)的定義，在此我們仿效馬可夫鍊(Markov chain)語言模型的觀念[10]，把全域(global)機率的定義，化簡成局部(local)機率之連乘積，也就是令

$$\begin{aligned}\mathrm{P}(T,S) &= \prod_{k=1,2,\ldots,N} \mathrm{P}_{\mathrm{loc}}(T_k) \\ &= \prod_{k=1,2,\ldots,N} \mathrm{P}(T_k \mid T_{k-1}, T_{k+1}, S_{k-1}, S_k, S_{k+1}).\end{aligned} \quad (2)$$

不過，由第 2 節的說明可知，英語的可能音節的數量很大，大於 15,000 個，如此連續三個音節的可能組合數量就大於 $3.375*10^{12}$，所以實作上仍然很難去估計局部機率項 $\mathrm{P}_{\mathrm{loc}}(T_k)$的參數值。因此，我們對局部機率的定義再作進一步的簡化，也就是令

$$\mathrm{P}_{\mathrm{loc}}(T_k) = \mathrm{P}(T_k \mid T_{k-1}, T_{k+1}, V_{k-1}, V_k, V_{k+1})\ . \quad (3)$$

其中 V_k 表示第 k 個音節的母音類別，而如第 2 節所說，CMU 音標裡共有 16 個母音類別。

3.2 基於 PPM 之局部機率估計

雖然公式(3) 裡的局部機率 $\mathrm{P}_{\mathrm{loc}}(T_k)$的定義已經經過簡化，但是我們現有的語料數量(7,350 個音節)，仍然顯得太少，而未能含蓋所有的聲調與母音的組合。因此，我們採取 PPM 的作法[8]，來估計局部機率 $\mathrm{P}_{\mathrm{loc}}(T_k)$。為了書寫方便，在此令 $a=V_{k-1}$、$b=V_k$、$c=V_{k+1}$，此外令 $d=T_{k-1}$、$e=T_k$、$f=T_{k+1}$，如此，$\mathrm{P}_{\mathrm{loc}}(T_k)$ 機率值的一個基本的估計方式是:

$$\mathrm{P}(e \mid a,b,c,d,f) = \frac{count(a,b,c,d,e,f)}{count(a,b,c,d,f) + count(esc_5 \mid a,b,c,d,f)} \quad (4)$$

其中 $count(a,b,c,d,e,f)$ 表示這 6 個因素組合在訓練語料中的出現次數，同理 $count(\cdots)$ 表示括弧內因素組合在訓練語料中的出現次數，而 $count(esc_5)$ 代表虛擬的逃脫符號 esc_5 的出現次數，關於 $count(esc_5)$ 值的取得方法在後面說明。當公式(4)發生分子的值為 0，而使估計出的機率值為 0 時，我們就採取 PPM 的降階機制，以降階後的公式來估計機率值。

PPM 依據一個虛擬的逃脫符號 esc 所對應的脫逃機率，來處理訓練語料裡沒出現過的因素組合，當遇到最高階(a,b,c,d,f 等 5 個因素組合為條件)的局部機率估計值為 0 時，就去掉其中一個因素，下降一階(只用 4 個因素組合之條件)去估計局部機率值，但是估計得到的機率值必需再乘以上一階的脫逃機率值，來作為新的機率估計值，其公式為:

$$\mathrm{P}(e \mid a,b,d,f) = \mathrm{P}(esc_5) \cdot \frac{count(a,b,d,e,f)}{count(a,b,d,f) + count(esc_4 \mid a,b,d,f)}\ . \quad (5)$$

如果降一階後所估計出的局部機率值仍然為 0，那就再次作降階，如此繼續，直到局部機率的估計值不為 0 為止。本論文使用的 PPM 降階機制，依序捨棄的因素是：先捨棄後一音節母音，再捨棄前一音節母音，接著捨棄本音節母音，然後捨棄後一音節聲調，再捨棄前一音節聲調，最後即是計算本音節聲調在訓練語料中的出現機率。

關於脫逃機率的估計，我們採用了 PPMC 之估計方法[8]，其公式為：

$$P(esc) = \frac{N_g}{N_t + N_g}, \quad (6)$$

其中 N_g 表示本階因素的組合之中，如公式(4)裡的 a, b, c, d, f 等 5 個條件因素再和不同的 e 值作組合時，在訓練語句裡的出現次數不為 0 的不同 e 值的個數，這樣的公式相當於令 $count(esc_5 | a,b,c,d,f) = N_g$。由於我們觀察到，逃脫機率的值有時候會相當大(如為 1)，因此我們嘗試把 PPMC 局部機率估計法作改良，就是每次下降一階時，將所求得的逃脫機率值再多乘以 0.05。

3.3 最佳聲調序列之搜尋

我們研究以兩層式的架構來作最佳(最大可能性)聲調序列的搜尋，如圖 4 所示，第一層經由動態規劃來尋找最佳的聲調序列，第二層則用以估計個別音節被設為某一聲調時的的局部機率。

詳細作法為，將一個句子的各個音節看成是動態規劃的階段，並且在每個階段裡，以前一音節、本音節、後一音節的可能聲調來組合出 8 * 8 * 8 = 512 個狀態，這裡各個 8 分別代表 7 種英語聲調加上一個空聲調(null)，空聲調用以表示第一音節之前及最後音節之後的聲調。

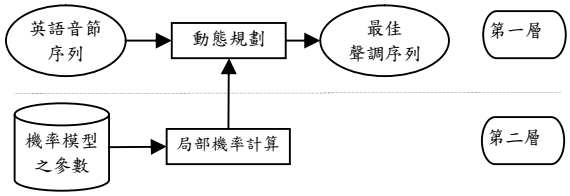

圖 4 最佳聲調序列之搜尋架構

由於一條路徑是由各個階段分別提供的一個狀態連結而成，而可能連結出的不同路徑的數量非常龐大，因此必需以動態規劃方式來尋找最佳路徑。在每一個階段(音節)裡我們分別計算該階段各狀態的累加機率值，累加機率值是由局部的狀態機率值取對數後相加而得到，所以也可視為是局部的狀態機率值連乘的結果。另外，也可把本階段某一狀態的累加機率值作分解，看成是前一個階段的累加機率值最佳者加上本階段某一狀態的局部機率值所得到，如公式 (7)、(8) 所表示的：

$$\delta_{k+1}(i) = \left(\underset{\substack{n \\ s.t.\ n \text{ and } i \\ \text{can be connected}}}{MAX} \delta_k(n) \right) + \log\left[\mathrm{P}_{\mathrm{loc}}(T_{k+1}(i))\right] \quad (7)$$

$$\delta_1(i) = \log\left[\mathrm{P}_{\mathrm{loc}}(T_1(i))\right] \quad (8)$$

其中 $\delta_{k+1}(i)$ 表示第 k+1 階段裡第 i 狀態上的累加機率值；$\mathrm{P}_{\mathrm{loc}}\left(T_{k+1}(i)\right)$ 表示第 k+1 階段裡第 i 狀態的局部機率值；$T_{k+1}(i)$ 表示第 i 狀態所對應的音節 S_{k+1} 的聲調 T_{k+1}，也就是 $T_{k+1} = (i / 8) \% 8$；$\delta_1(i)$ 表示第 1 階段裡第 i 狀態的累加機率值，其值等於第 1 階段裡第 i 狀態上的局部機率值取對數。依公式(7)遞迴計算到最後階段，即可求出整體上最大的累加機率值，然後再追溯(backtrack)出最佳的路徑，而依此路徑就可得到對應的聲調序列。

由於各階段裡的狀態是由三連音節(前一音節、本音節、後一音節)的聲調作組合而形成，因此只能從前一音節的相容狀態連結到本音節目前所考慮的狀態，所謂的”相容狀態”，表示相鄰音節之間的狀態連結是有限制的，如圖 5 所示，以階段 k 的狀態(2,5,3)來看，其來源路徑只有從階段 k-1 裡本音節及後一音節聲調分別為 2 和 5 的狀態連結過來，如圖中階段 k-1 的(7,2,5)和(6,2,5)狀態可以連結到(2,5,3)狀態，所以它們之間是相容的。但是階段 k-1 的(7,3,5)及(4,2,1)狀態就不能夠連結到(2,5,3)狀態，因為它們之間是不相容的。

3.4 聲調預測實驗

使用本節的最大可能性預測法(MLP)來作聲調序列的預測，然後將預測出的音節聲調與手動標記的音節聲調作比對以計算正確率，結果得到如表 3 所示的正確率數值。在此 inside(參加機率模型訓練)語句有 744 句，共 7,350 個音節；而 outside(未參加機率模型訓練) 語句只有 69 句，共 756 個音節。

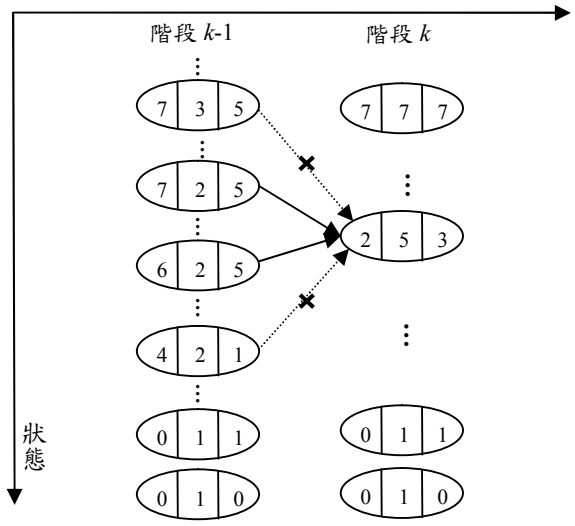

圖 5 狀態限制連接之例子

表 3 MLP 法聲調預測之正確率

實驗 \ 方法	MLP + 原始 PPMC	MLP + 改良 PPMC
Inside 測試	58.0%	95.2%
Outside 測試	31.8%	35.2%

由表 3 可知，當使用改良過的 PPMC 逃脫機率估計法，來取代原始的 PPMC 估計法，不僅可讓內部測試的聲調預測正確率，從 58.0%大幅提升到 95.2%，還可讓外部測試的正確率，小幅提升 3.4% (35.2% － 31.8%)。不過，外部測試的正確率 35.2%，比起內部測試的正確率 95.2%，可說是差了很多，這是否意味 MLP 法有問題? 為了作檢驗與比較，我們拿相同的語料，並使用相同的 5 個特徵項(如公式(3) 裡的條件項)，去給 Weka 軟體作分類實驗，結果得到如表 4 所示的分類正確率數值，由表 4 可知，Weka 的分類正確率，在外部測試時，會比我們的 MLP 法好約 6%，不過在內部測試時，就比 MLP 法的差很多(95.2% - 51.8% = 43%)。所以 MLP 法的特色是，對於訓練語料音節的聲調，具有很好的預測能力，然而對於測試語料的音節聲調預測，效能會稍差一些，這應是與逃脫機率估計方法的好壞有關。

表 4 Weka 聲調分類之正確率

實驗 \ 方法	Weka + J48 分類	Weka + SimpleCart
Inside 測試	51.8%	45.7%
Outside 測試	38.2%	41.0%

4. 音高軌跡產生

關於音節之音高軌跡的產生，我們採用類神

經網路(ANN)來建立音高軌跡模型，再用以產生音高軌跡。所用之 ANN 其結構如圖 6 所示，包括輸入層、隱藏層、遞迴隱藏層、和輸出層，輸入層具有 45 個節點，用以輸入 11 項語境參數，輸出層具有 16 個節點，用以輸出一個音節的 16 個正規化時間點上之音高頻率值。

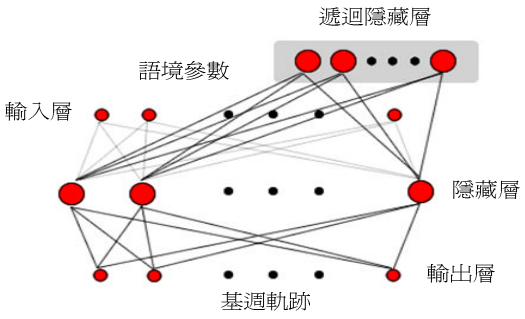

圖 6 ANN 結構

ANN 的輸入稱為語境參數，語境參數可設定使用不同層級的語言資訊，過去我們研究國語之語音合成，主要是以音節作為語境的組成單位，但是從第 2 節的說明可知，英語由於音節數量很大，並不適合以音節作為語境的組成單位，因此本研究裡使用半音節的分類作為語境的單位，詳細的配置情形如表 5 所示。

表 5 ANN 輸入層之語境參數

項目	前音節聲調	前音節之後半音節分類	前音節之母音類別	本音節聲調	本音節之前半音節分類	本音節之後半音節分類	本音節母音類別	句中位置	後音節聲調	後音節之前半音節分類	後音節之母音類別
bits 數目	3	5	5	3	5	5	5	浮點數	3	5	5

我們使用 744 句共 7,350 個音節分析出的基週軌跡，來訓練前述之 ANN 模型，訓練循環的次數為 5,000 次。在此採取均方根誤差(RMS error)之量測方式，來量測音高軌跡之預測誤差，公式如下：

$$d(X,Y)=\sqrt{\frac{1}{16}\sum_{i=0}^{15}(x_i-y_i)^2} \qquad (9)$$

其中 $X=<x_0,x_1,\cdots,x_{15}>$ 表示一個音節實際唸的 16 點音高軌跡值，$Y=<y_0,y_1,\cdots,y_{15}>$ 表示 ANN 所預測出的音高軌跡。我們求取所有音節的預測誤差的平均值(AVG error) 、及誤差之標準差(STD error)，來衡量 ANN 模型的好壞。

對於隱藏層節點個數的設定，我們使用外部語料 69 句(共 756 個音節)來作測試，嘗試從 8 個節點變化到 18 個節點，結果得到如表 6 所示之誤差數值。當把隱藏節點個數設為 9 時，測試語句的音高軌跡預測之 AVG 誤差及 STD 誤差最低，因此我們選擇設定隱藏層的節點數為 9。

5. 英語合成系統與聽測實驗

我們建造的語音合成系統，其處理流程如圖 3，而圖 3 各方塊的處理動作為: (1)讀入欲合成的英文文字檔案，經由剖析切割出各個句子；(2)藉由查詢拼音詞典，將各個英語詞(word)轉成音節，並且產生各音節的語境參數；(3) 預測各音節的聲調，使用第 3 節說明的方法；(4)依據預測出的聲調序列及其它語境參數，帶入 ANN 模型去產生出基週軌跡，如第 4 節的說明；(5)以規則式作法設定各音節的音長與音量；(6)使用 HNM 合成模組去合成出英語語音信號。

表 6 音高軌跡 ANN 模型之測試誤差

隱藏層單元數	AVG error	STD error
8	0.02924	0.01949
9	0.02852	0.01922
10	0.02891	0.02003
11	0.02890	0.01965
12	0.02922	0.02013
14	0.02942	0.02042
16	0.02996	0.02127
18	0.03104	0.02301

5.1 音量與音長設定

先將音節的振幅調整到一個定值，再依音節的核心母音(代表主要之嘴型)來調整音量，規則如下：(1)當核心母音為/aa/、/ah/時，該音節音量不變；(2)當核心母音為/ae/、/eh/、/ey/時，音節音量下降 1dB；(3)當核心母音為/ow/、/oy/、/aw/、/ao/時，音節音量下降 2dB；(4)當核心母音為/ih/、/iy/時，音節音量下降 4dB；(5)當核心母音為/er/、/ax/時，音節音量下降 5dB；其餘情況音量都降低 3dB。

在音節的音長方面，我們也採用規則式的作法來作設定，首先將音長的長度調整到一個定值，再依音節的母音來調整音長，規則設定如下：(1)當核心母音為長母音時，音節長度不變；(2) 當核心母音為短母音時，音節長度縮短為 0.8 倍；(3)當一個音節的開始與結束都無子音時，此音節的長度不變；(4)當音節開始或結束的子音為有聲(如/b,d,g/)或無聲爆破音(如/p,t,k/)及無聲摩擦音(如/f, s, ʃ /)時，音節長度增長 0.15 倍；(5)當音節開始或結束的子音為其它有聲子音(m,v,z,l)時，音節長度增長 0.1 倍；(6)當 word 為單音節時，音節長度不變；(7)當一個英文 word 有 2 個音節時，則各組成音節之長度縮短為 0.9 倍；(8)當一個英文 word 有 3 個音節時，各音節長度縮短為 0.8 倍；當一個英文 word 有 4 個或 4 個以上的音節時，各音節長度縮短為 0.75 倍。

5.2 HNM 信號合成

HNM 翻譯為諧波加雜音模型，從名稱來看就

是要把聲音信號 $s(t)$ 分解成諧波 $h(t)$ 及噪音 $n(t)$ 兩部分。HNM 提供了一個最大有聲頻率(maximum voiced frequency, MVF)的訂定方法[15]，且以 MVF 作為分界點，對於頻率低於 MVF 的頻帶，就產生出諧波信號，而對於頻率高於 MVF 的頻帶，就產生出雜音音信號，然後再將這兩種信號加起來作為 HNM 的合成信號，也就是令 $s(t)=h(t)+n(t)$ 。

產生諧波信號時，以頻率值有倍數關係的多個弦波來作合成，如公式(10)所示：

$$h(t)=\sum_{k=1}^{K(t)} a_k(t)\cdot\cos\left(\phi_k(t)\right) \qquad (10)$$

其中 $a_k(t)$ 及 $\phi_k(t)$ 表示時間 t 時，第 k 個弦波的振幅及相位，$K(t)$ 則表示時間 t 時諧波的數目。當產生雜音信號時，則是以間隔固定為 100Hz 的弦波信號來作為信號成分，至於弦波的振幅，則依倒頻譜(cepstrum)係數轉換出的頻譜包絡來決定。對於無聲音素(如/p/)的信號合成，可把 MVF 設為 0，即整個頻譜都視為雜音部分。較詳細的作法可參考原始文獻[15]或我們先前的論文[1]。

5.3 聽測實驗

系統製作完成之後，接著我們對合成出的英語語音作聽測評估，聽測實驗分成二次實施，第一次實驗稱為系統內聽測，目的是比較我們系統所合成出兩類型語句的語音在自然度上的差異，第一類型語句的例子如圖 7 所示，是作模型訓練用的語句，第二類型語句如圖 8 所示，是作外部測試用的語句(即未參加模型訓練)，因此語句內容是不同的。第二次聽測稱為系統間聽測，目的是比較我們系統和 Festival 的 online demo – HTS 程式[14, 5]在合成語音的自然度上的差異，使用的語句是外部測試用的語句。我們系統合成出的語音音檔，可從網頁 http://guhy.csie.ntust.edu.tw/~chunwei/experiment.htm 去下載。

```
not many teenagers run their own company.
but the Olsens are used to being unique.
they grew up in the public eye.
Mary Kate and Ashley Olsen.
Mary Kate and Ashley were born on June thirteenth.
they grew up with the whole world watching.
we're so short and tiny.
```

圖 7 模型訓練用的語句內容

```
Listen attentively when your friend share some thing with you.
But that responsibility proved too much for her.
Building a successful friend ship is never easy.
he may not be very much fun to be around.
to have a friend you must first be a friend.
So she ran away from home when she was thirteen.
```

圖 8 外部測試用的語句內容

進行系統內聽測實驗時，將外部測試語句所合成之音檔作為音檔 A，而模型訓練用的語句所合成之音檔作為音檔 B。另外，進行系統間聽測實驗時，把我們系統合成出的音檔作為音檔 A，而把 Festival HTS 合成出的音檔作為音檔 B。評估的方式是，當受測者聽完 A、B 兩音檔之後，詢問他音檔 B 比音檔 A 的語音自然度好或差，並請他給一個評分，評分的範圍由最高 9 分到最低 1 分，9 分代表 B 比 A 好很多，7 分代表 B 比 A 好一些，5 分代表沒差別，3 分代表 B 比 A 差一些，1 分代表 B 比 A 差很多，只可以打整數分數。

我們請了 15 位聽測者來進行聽測評估，15 位當中有 4 名為本實驗室與校內的學生，另外 11 名則為校外的人士。在聽測的過程中，受測者可以自由選取合成音檔的任何一段反覆試聽。兩次聽測實驗之後，分別算出的平均分數如表 7 所示，由系統內聽測的平均評分 5.67 可知，模型訓練用語句的自然度稍微優於外部測試用的語句，其原因應是聲調預測的正確率差異，即模型訓練用語句的正確率比外部測試用語句的正確率高很多。另外，由系統間聽測的平均評分 8.20 可知，我們系統所合成語音的自然度比 Festival HTS 的合成語音差很多，這是因為大部份的受測者認為我們系統在音長及停頓的設定方面感覺不是很自然，並且相鄰音節之間感覺過於獨立而沒有連續性。

表 7 二次聽測實驗之評分結果

	系統內聽測	系統間聽測
平均值	5.67	8.20
標準差	0.943	0.909

6. 結論

英語語音合成的一個重要問題是，英語詞典並沒有提供音節聲調的資訊，因此我們的研究焦點就放在英語音節的聲調預測，及基週軌跡的產生。

在聲調預測方面，我們研究提出基於最大可能性之預測方法，而在實施的方法上，提出以兩層式的架構來尋找最佳的聲調序列，第一層藉由動態規劃來尋找最佳的聲調組合之狀態序列，第二層則是估計各個音節的局部機率，我們基於 PPMC 法來估計局部機率，並且改良了 PPMC 法，當以模型訓練用的語句作測試時，聲調預測的正確率可達到 95.22%，而當以外部測試用的語句測試時，聲調預測的正確率則僅達到 35.19%。

在基週軌跡產生方面，我們研究以 ANN 來建立基週軌跡模型。過去使用 ANN 作國語基週軌跡產生時，ANN 的語境參數是以音節為單位，但是英語由於音節的數量太大，我們便研究以半音節的分類來作為語境的單位。在 ANN 的隱藏層節點數的實驗中，使用外部測試用的語句，當把隱藏層的節點數設為 9 時，才會得到最小的 AVG 和 STD 預測誤差。

另外，我們將所研究的聲調預測及基週軌跡產生方法製作成程式模組，用以建置一個初步的英語語音的合成系統。其它的程式模組，我們採用規則

式作法來設定音量、音長，而信號合成部份則是採用先前發展的 HNM 模組來作信號合成。對於所建造的英語語音合成系統，我們也作了聽測評估，聽測實驗的評分顯示，如果音節聲調的預測正確率越高，則合成的語音聽起來就會更為自然。此外，與 Festival 的 online demo – HTS 的合成語音作比較，發現在語音的自然度上，我們系統比 Festival HTS 明顯差了很多，其原因除了聲調預測正確率不夠高之外，我們以規則式作法來設定音長、音量，應也是造成自然度下降的重要因素。所以，未來可再研究提高聲調預測的正確率，此外也要研究較佳的音長、音量的設定方法，及解決相鄰音節顯得過於獨立的問題，以提高合成語音的自然度。

參考文獻

[1] 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 2006 國際電腦音樂與音訊技術研討會*，台北，*session 2* (音訊處理 I)， 2006。

[2] 梁弘學，*英語歌聲合成之研究*，碩士論文，國立台灣科技大學 資訊工程研究所，台北，2009。

[3] 陳忠緯，*用於英語語音合成之基週軌跡產生方法*，碩士論文，國立台灣科技大學 資訊工程研究所，台北，2010。

[4] Carnegie Mellon University, *The CMU Pronouncing Dictionary*, http://www.speech.cs.cmu.edu/speech/ .

[5] HTS working group, *HMM-based Speech Synthesis System* (HTS), http://hts.sp.nitech.ac.jp/ .

[6] 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 (4), pp. 371-390, 2007.

[7] J. E. Cahn, *Generating Expression in Synthesized Speech*, Technical Report, Media Lab, MIT, Cambridge, Boston, 1990.

[8] K. Sayood, *Introduction to Data Compression*, 3’rd ed., Morgan Kaufmann, San Francisco, USA, 2005.

[9] K. Silverman, *et al.*, “ToBI: A Standard for Labeling English Prosody”, *Int. Conf. on Spoken Language Processing*, pp. 867-870, Banff, 1992.

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

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update", *SIGKDD Explorations*, Vol. 11 (1), 2009.

[12] P. Taylor, “The Tilt Intonation Model”, *Int. Conf. on Spoken Language Processing*, Sydney, 1998.

[13] R. A. J. Clark, *Generating Synthetic Pitch Contours Using Prosodic Structure*, Ph.D. thesis, University of Edinburgh, Edinburgh, UK, 2003.

[14] The Centre for Speech Technology Research, *The Festival Speech Synthesis System*, http://www.cstr.ed.ac.uk/projects/festival/

[15] Y. Stylianou, *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification*, Ph.D. thesis, Ecole Nationale Superieure des Telecommunications, Paris, France, 1996.