結合 HMM 頻譜模型與 ANN 韻律模型之國語語音合成系統
A Mandarin Speech Synthesis System Combining HMM Spectrum Model and ANN Prosody Model

古鴻炎 Hung-Yan Gu　賴名彥 Ming-Yen Lai　蔡松峰 Sung-Fung Tsai

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

本論文研究了一種結合 HMM (hidden Markov model)頻譜模型與 ANN (articifical neural network)韻律模型的國語語音合成系統。在訓練階段，對各個訓練語料音框算出 DCC 係數(discrete cepstrum coefficients)，以作為頻譜特徵參數，接著對於一種音節的多個發音，依 DTW (dynamic time warping)匹配出的頻譜演進路徑作分群，各群建立一個 HMM，並記錄各音節發音的文依性資訊。在合成階段，首先依據文依性資訊挑選出輸入文句各音節的 HMM 模型，接著判定音節 HMM 的各個狀態為無聲、或有聲，然後使用音長 ANN 模型及狀態平均音長來決定 HMM 各狀態應該產生的音框數。除了前人提出的 MLE(maximum likelihood estimate)法，我們另外研究二種內插方法來產生各音框的 DCC 係數，以讓語音合成的速度達到即時處理。接著依據 DCC 係數轉出的頻譜包絡，及 ANN 產生出的基週軌跡與音長，去控制 HNM (harmonic-plus-noise model)作語音信號的合成。聽測實驗的結果顯示，使用所提出的加權式線性內插法來產生 DCC 係數，合成出的語音信號比起使用 MLE 法的，可以得到一些自然度的改進；另外，使用 ANN 音長參數，也比使用 HMM 狀態本身的平均音長，會獲得明顯較高的自然度。

一、前言

早期的語音合成系統，信號的合成方式是，先產生出所用的聲學模型的參數，再拿去控制聲學模型來合成出語音信號，聲學模型如 LPC (linear prediction coding)全極(all pole)模型[1]或幅峰(formant)合成模型[2]，這樣的合成方式也稱爲參數式合成法。參數式合成法在產生出一序列音框的模型參數時，一者由於各音框的模型參數所對應的頻譜曲線不夠精細(參數個數不夠)，或者由於未考慮音框之間頻譜形狀隨著時間作演進的擬真改變方式，而造成所合成出的語音信號令人覺得不夠自然及不夠流暢，即便是已經使用了韻律模型來決定基週軌跡、音長、與音量的情況下，仍是如此。

爲了解決合成語音不夠自然的問題，過去不少研究者改採另一種方向(approach)，亦即錄製大型語料庫、並且研究從語料庫挑選出語音單元(如音節)的方法[3, 4, 5, 6]，希望拿真人所錄的語音單元來直接作串接，如此至少可保證各個語音單元內的信號是自然的，這樣的語音合成方向，稱爲單元選擇式合成法。過去我們曾研究了一些和國語語音合成相關的問題[7, 8, 9, 10]，基本上我們並不願意採取單元選擇式之合成作法，其中一

Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010), Pages 281-295, Puli, Nantou,Taiwan, September 2010.

個原因是，當欲把技術移植到另一種語言(如閩南語、客語)時，仍需花費大量的人力與時間來重新錄製該語言的語料庫；另外一個原因是，我們希望所製作的語音合成系統是具有多重功能的，例如提供說話速度快慢的調整功能、與改變音色(如依女聲原音來合成出男聲語音)的功能。

因此，我們主要採取的是參數式之合成作法，在參數式合成作法的前提下，我們曾研究過一種頻譜演進模型的建立方法[10]，雖然它可以讓音節內的頻譜演進獲得顯著的自然度改進，不過合成出的國語語音，在一些相鄰音節之邊界，仍然會聽出銜接得不夠順暢的情況。對於這樣的問題，我們覺得近幾年已被不少人研究過的 HMM 爲基礎的頻譜演進模型[11, 12, 13, 14]，應是一個不錯的解決方法，所以我們便開始研究以 HMM 來建立國語音節的頻譜演進模型，希望藉以提升合成語音的聲學流暢性(acoustic fluency)，也就是讓合成的語音，不論在音節內部或相鄰音節之間，都能具有流暢的頻譜演進與頻譜銜接。

另外，關於韻律上的流暢性，也就是一個合成語句的前後音節之間，不要發生音調忽高忽低、說話速度忽快忽慢的情形，我們覺得要獲得高水準的韻律流暢性，應該要採用專門作韻律參數產生的模型與方法，而不需要像別人一樣，把韻律參數的產生和頻譜演進的控制綁在一起，亦即使用 MSD (multi space probability distribution) HMM 模型來含蓋韻律參數和頻譜演進。我們的理由之一是，過去已有不少優秀的韻律參數產生方法被提出[15, 16, 17, 18]，就算是我們自己的方法不好，也還可以仰賴別人的方法及其所發展的程式模組；另外一個理由是，我們查閱了以 MSD-HMM 來建立基週軌跡模型的文獻[11, 12]，發現他們只是使用 context-dependent HMM 來建立各語音單元的基週軌跡模型，而並沒有考慮到一個語句的句首與句末音節，會處於不同的韻律狀態的觀念[8, 18]，因此我們懷疑 MSD-HMM 是否能夠產生出像一個真實語句那樣具有下傾(declining)及抑揚變化的句子基週軌跡。

由前述說明可知，本論文建造的國語語音合成系統，雖然採取了 HMM 來建立各國語音節的頻譜演進模型，不過產生音節的基週軌跡、音長之韻律模型則是另外採取 ANN 來建立的。前面提到的 HMM 頻譜演進模型及 ANN 韻律模型，它們的構造與訓練方式將在第 2 節中介紹；而合成階段的處理流程與方法，則將在第 3 節中介紹；第 4 節則呈現合成語音的頻譜例子，及說明聽測實驗的結果。

二、模型訓練

在訓練階段我們首先進行語料的錄音，然後對語料作標音、切音的處理；之後使用基於離散倒頻譜(discrete cepstrum)之頻譜包絡(spectral envelope)估計方法[19, 20]，去估計出一個音節各音框的頻譜參數；接著對各個語句的組成音節進行 DTW 匹配來求得頻演(頻譜演進)路徑，以便依此路徑來對相同拼音的音節發音作分群；分群後，再對各群的音節發音作 HMM 模型訓練。此外，我們也對各個切音得到的音節，量測它的音長(duration)，再配合文句分析所得到的資料，用以訓練 ANN 音長模型。訓練階段的主要處理流程如圖 1 所示，以下就對圖 1 各方塊加以說明。

2.1 錄音、標音

我們使用 Acoustic Systems 公司之 RE-242 隔音室，請一位男性錄音者來錄製訓練語料，錄音者發音的速度稍微比正常說話速度慢，目的是讓每個字的發音較完整也比較方便標記出音節的正確邊界。首先我們錄各個國語基本音節的單獨發音，以便用於 DTW

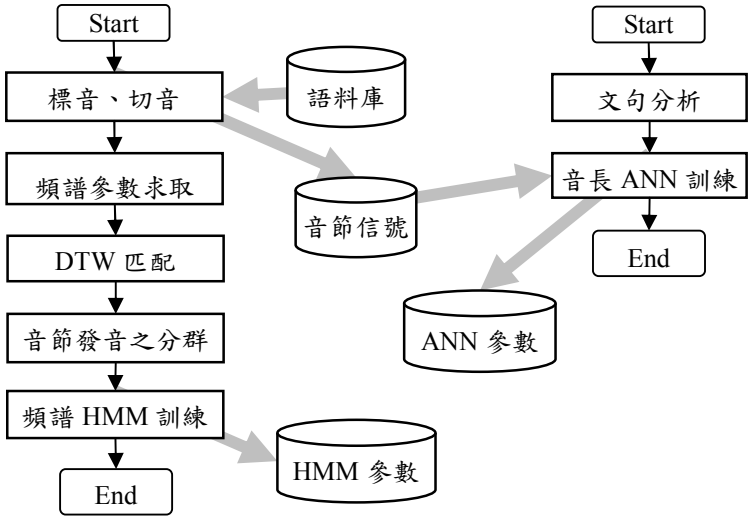


圖 1 訓練階段之主要處理流程

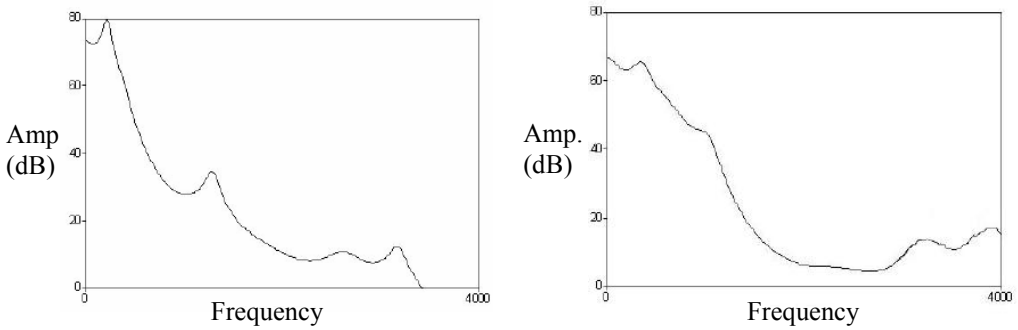
匹配方塊，然後再以整句發音方式來錄訓練語料，總共錄製了 1,208 個語句，音節總數爲 10,173 個，語句的內容則參考自 TCC300 語料庫，由於 TCC300 採用的語句內容已經過篩選，所以可含蓋國語的 407 個基本音節。錄音之後，將語音音檔轉換成取樣率 22,050 Hz、解析度 16bits/sample 之音檔格式。

爲了擷取出各個語句裡的各個音節的語音信號，我們必須先作標音的動作，亦即在時間軸上標示出各個音節的左右邊界、音標與聲調。由於需作標音的音節數量不少，所以我們先使用 HTK (HMM tool kit)來作音節邊界之 forced alignment 處理，即自動標音，不過 forced alignment 所標示的音節邊界不夠準確，這將會影響到後面的處理步驟，因此我們再使用 Wavesurfer 軟體來對音節邊界作人工方式之微調、更正，然後依此標音結果來進行切音，切出來的各個音檔，再依據語句編號、句內音節序號、及音節拼音來對它們命名。

2.2 頻譜參數求取

前人的研究成果裡提到[14]，合成語音的信號品質跟頻譜包絡的形狀有很大的關係，如果頻譜參數還原出的頻譜包絡有 over smoothing 的現象，如圖 2(b)所示，就會導致合成的語音信號變得模糊、不清楚，相對地較爲精確的頻譜包絡，如圖 2(a)所示，會顯現出較窄的共振峰頻寬，而可用以合成出較清晰的語音信號。過去我們曾以離散倒頻譜爲基礎，研究提出一種頻譜包絡之估計架構[20]，來求取各音框的離散倒頻譜係數(DCC, discrete cepstrum coefficients)，先前的實驗顯示，DCC 係數可還原出頗爲精確的頻譜包絡，所以在此我們決定採用 DCC 係數作爲頻譜參數。

圖 3 爲 DCC 係數求取的流程圖。當輸入一個語音音框之後，首先進行基頻偵測，這裡使用了 ACF (autocorrelation function)搭配 AMDF (absolute magnitude difference function)的作法[21, 22]，以判斷音框是否爲有聲及求出基頻值，作爲之後頻譜峰點挑選的根據；接著將音框信號乘上漢寧(Hanning)窗，並且音框後面補上零，使其長度成爲 1024 點後，再對音框作 FFT (fast Fourier transform) 計算而得到 FFT 頻譜；然後依據 FFT 之振幅頻譜(magnitude spectrum)挑選頻譜曲線上的峰點，再將選到的峰點的頻率值作尺度轉換；之後使用挑選出的頻譜峰值和轉換後的頻率值，去估計出 DCC 係數 $c_0, c_1, c_2, \ldots, c_p$，其中 p 表示階數，在此我們設定 p 爲 38。有了 DCC 係數，就可據以還原算出頻譜



(a)較精確之頻譜包絡[14]　　(b) 較粗糙之頻譜包絡[14]

圖 2 頻譜包絡曲線之例子

包絡，公式爲:

$$\log S(f) = c_0 + 2\sum_{k=1}^{p} c_k \cdot \cos(2\pi \cdot f \cdot k), \quad f = \frac{0}{N}, \frac{1}{N}, \frac{2}{N}, ..., \frac{N-1}{N} \tag{1}$$

其中，N 表示 FFT 的點數(即 1,024)，f 表示尺度正規化後的頻率變數(數值範圍在 0 至 1 之間)，$S(f)$表示逼近出的頻譜包絡曲線。關於圖 3 流程的較爲詳細的說明，請參考我們先前的論文[20]。

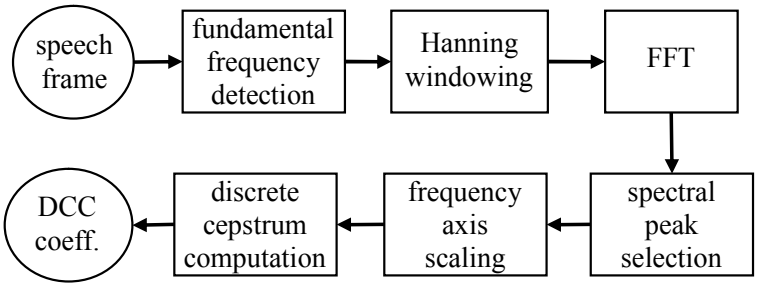


圖 3 DCC 係數求取之流程

2.3 DTW 匹配 與 音節發音分群

在不同的前後文情況下發出同一個音節的語音，雖然音節的拼音符號一樣，但是受到前後文的影響，它們的頻演(頻譜演進)方式是會有差異的。如果簡單地把聲母與韻母視爲音節內的兩個狀態時，則可發現在不同前後文情況下，兩個狀態所佔音節音長的比例是會有差異的。考慮到前後文對頻演方式的影響，過去 HMM 爲基礎的語音合成作法，對於一個語音單元在不同前後文音素(phoneme)組合下的多個發音，採取以決策樹(decision tree)來進行**文依性**(前後文相依性)分群，之後才對各群發音分別建立各自的 HMM 模型 [12]。

在本論文裡，我們先應用先前研究的頻譜演進觀念[10]，來得到一個音節在不同前後文情況下的多個發音的頻演路徑，亦即把該音節的單獨發音放在縱軸，而把該音節在語句裡的各個發音放在橫軸，來作 DTW 匹配以求得一條最小頻譜距離之路徑，稱爲頻演路徑。之後，依據頻演路徑來爲一個國語音節的多個發音作文依性分群，如此具有相似頻演路徑的發音就會分到一群，再據以訓練出一個對應的 HMM 模型，如此在訓練語料中具有多次發音的一種國語音節，就會訓練出數個 HMM 模型。本論文採取以頻演路徑作爲音節發音的文依性分群的依據，不同於一般以決策樹作文依性分群依據的作法，

我們的作法比較具有聲學上的實證特性。

由於國語各種音節的使用頻率並不一致，而在訓練語料中各種音節的發音數量也是有多有少，所以先根據一個音節的發音數量來決定分群數是比較恰當的，以避免 HMM 模型訓練時，發生語料不足的情形。在此我們嘗試了幾種分群數之決定規則，最後決定採取的規則是: 群數 = 發音數 / 10。至於分群的方法，我們先對各條頻演路徑作橫軸時間的正規化，以便把一條頻演路徑表示成 32 維(dimension)的向量，然後就可使用 K-means 演算法來作分群。分群結果之好壞的評定，我們使用 CH (Calinshi and Haravasz) 評估法[23]來量測，由於 K-means 分群的初始中心點爲隨機選取，所以各次分群的結果可能都不一樣，因此每一次作完 K-means 分群，就依分群結果去計算它的 CH 值。在此我們對各個國語音節的發音都作 20 次的分群，然後拿 20 次分群中具有最大 CH 值的那一次分群結果當作最後的結果。此外，各個國語音節之分群完成時，我們也要記錄各群發音的來源資訊及前後文資訊，以便在合成階段能夠據以作搜尋，來決定一個待合成的音節，應該挑選那一群所訓練出的 HMM 模型來使用。

2.4 頻譜 HMM 訓練

雖然在 2.2 節提到，一個音框會分析出 $c_0, c_1, c_2, \ldots, c_{38}$ 等 39 個 DCC 係數，但是由 HNM 信號合成的實作上，我們發現參數 c_0 並不能用以還原出原始語音波形的振幅包絡，因此後來我們改成以音框能量之對數值來作爲 c_0 參數；此外，在合成階段我們需要對一個被挑選出的 HMM 的前幾個狀態，作無、有聲之判斷，因此我們必須增加一個參數來記錄一個音框是否有週期性的資訊，也就是當圖 3 的基頻偵測方塊偵測一個音框爲有聲時，就令 $c_{39} = 1$，不然則令 $c_{39} = 0$。如此在本論文裡，一個音框會分析出 40 個頻譜參數 c_i，及 40 個頻譜差分(delta)參數Δc_i, $i = 0, 1, 2, \ldots, 39$。

考慮到一個音節的發音經過分群後，一群內的發音個數已經不多，再者參考前人的研究成果[11, 12]，我們決定一個狀態上只設定一個高斯混合(mixture)，並且對於高斯混合的機率計算，只採用對角化的共變異矩陣(covariance matrix)；此外，由於 HMM 訓練的語音單位爲音節，所以在此設定 HMM 的狀態數爲 8 個，以便能保有足夠細緻的聲、韻母頻譜演進的資訊，而 HMM 的結構，則設定成由左至右(left-to-right)。

當一個音節的多個發音經過分群後，我們就逐一拿各群的發音來訓練出該群所對應的 HMM 模型，訓練的程序是，先求取 HMM 初始模型，再由初始模型開始分段 K 中心法 (segmental K-means) [24] 之反覆式(iterative)訓練步驟，反覆的終止條件是，當本次反覆之維特比(Viterbi)搜尋後，各 HMM 狀態上所收集到的音框內容與上一次反覆所收集的音框內容完全相同時，即結束訓練。此外，考慮一個音節作合成時，該音節的音長數值是由韻律模型所產生，但是我們必須依此音長值來調整音節內部各個狀態所佔的音長比例，因此在 HMM 模型訓練時，也需要加入狀態音長參數的訓練，也就是估計 HMM 第 i 個狀態被停留的平均音框數 D_i 及其變異數 V_i。

2.5 音長 ANN 訓練

韻律模型負責產生出韻律參數，其中一項是音節音長，在此我們使用先前開發的 ANN 模型的訓練程式[10]，來對新錄的語料(共 10,173 個音節)作音長 ANN 模型的訓練；另外一項重要的韻律參數是音節的基週軌跡，在此直接使用先前開發的基週軌跡產生程式[25]；至於音節的音量，我們並未採用獨立的音量模型，而是直接依據 HMM 各狀態上的高斯混合平均向量裡的 c_0 參數。

所採用的音長 ANN 模型，其結構包含: 輸入層、隱藏層、遞迴隱藏層、輸出層。輸入層共有 28 個節點(nodes)，用以輸入 8 種語境參數，詳細的配置方式如表 1 所列，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 與表 3 所列，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；至於句中位置參數，它是一個時間比例之數值，用以代表本音節在整句話中的時間位置。

表 1 語境參數與所佔 bit 數

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	句中位置參數	後音節聲調	後音節聲母類別
bit 數	3	4	3	5	6	浮點數	3	3

表 2 聲母分類表

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 韻母分類表

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		

ANN 模型的輸出層，只有一個節點，該節點輸出的是正規化的音長值，若乘以平均音長之秒數，就可以還原成實際的音長秒數值。此外，隱藏層與遞迴隱藏層的節點個數是一樣的，至於應使用多少個節點，必需依據 ANN 模型訓練的結果來決定。在此我們採用最陡坡降學習法，以 9,156 個錄音音節來訓練 ANN 的權重值，而另外 1,017 個錄音音節則作爲外部測試之用，此外分別設定隱藏層的節點數爲 15、16、17、18、19、20 等，如此在 5,000 次反覆學習之後，分別量測內部訓練音節、與外部測試音節的音長預測誤差，計算平均值後得到如表 4 所示的數值，依據表 4 外部測試音節的平均誤差值，我們決定把隱藏層的節點數設爲 17。

表 4 不同之隱藏層節點數的誤差值

節點數	15	16	17	18	19	20
內部訓練 平均誤差	0.03600	0.03625	0.03521	0.03478	0.03514	0.03491
外部測試 平均誤差	0.04336	0.04371	0.04285	0.04387	0.04304	0.04362

三、語音合成處理

當一個輸入的文句作完文句分析與斷詞之後，接著使用 ANN 模型來產生該文句各

音節的音長與基週軌跡之韻律參數；然後，依據各個音節在句子中的前後文去挑選出適當的音節 HMM，再依據所挑出的 HMM 的參數來產生出該音節的各個音框的頻譜參數，也就是 DCC 係數；此外需要對一個音節所對應的 HMM 的狀態，作無、有聲之判斷，再依基週軌跡參數來計算出各音框的音高頻率值；之後，採取以 HNM 信號模型來合成出語音信號，HNM 信號模型可依序接受各音框的頻譜參數、音高參數來產生出信號樣本。合成階段的主要處理流程如圖 4 所示，以下就對圖 4 各方塊加以說明。

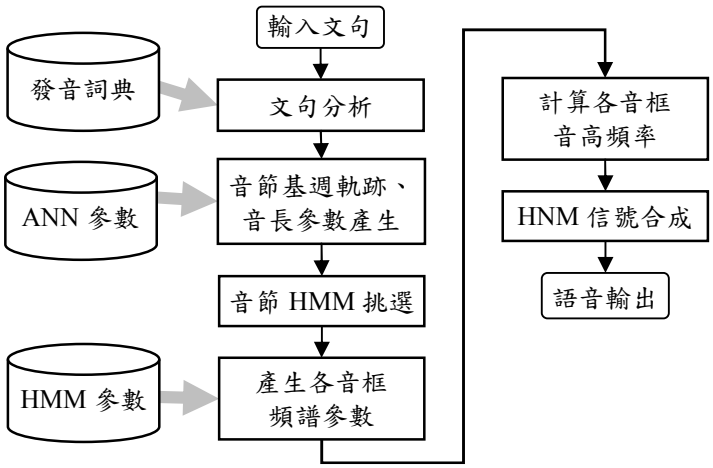


圖 4 語音合成之處理流程

3.1 音節音長、基週軌跡參數產生

我們使用獨立於頻譜 HMM 之外的音長 ANN 模型，來產生一個語句各音節的音長，希望合成出的語音能夠有更佳的自然度表現。由 2.5 節可知所用的音長 ANN 的結構，因此我們必須先把各音節的 8 個語境參數數值準備好，然後逐一帶入音長 ANN，去計算出正規化的音長值，再將此正規化的音長值乘以訓練階段保存的平均音長之秒數，如此就可得到一個音節的音長秒數值。之後，一個音節的音長秒數就可用以換算出音框數量，再據以決定該音節的 HMM 八個狀態各別應分配的音框個數。

關於音節的基週軌跡，我們也是使用獨立於頻譜 HMM 之外的基週軌跡 ANN 模型[25]，來產生一個語句各音節的基週軌跡參數。由於一個音節有 16 個基週軌跡參數，分別代表 16 個正規化時間點上的音高頻率，所以在此所用的基週軌跡 ANN 之結構，輸出層具有 16 個節點用以輸出 16 個參數，隱藏層與遞迴隱藏層都使用 30 個節點，而輸入層則和音長 ANN 的輸入層一樣，使用相同的語境參數、且具有 28 個節點。因此，我們同樣地先把各音節的 8 個語境參數數值準備好，然後逐一帶入基週軌跡 ANN，去計算出正規化的基週軌跡參數值，再將這些正規化的數值乘以一個常數倍率，就可以得到各音節的基週軌跡參數。

3.2 音節 HMM 挑選

由於一個音節內部的頻譜演進是由 HMM 來作控制，而一個音節的訓練發音可能訓練出數個 HMM，因此在合成階段要合成輸入文句裡的一個音節時，考慮到音節之間頻譜銜接的流暢性，我們必需依據前後音節發音音素的組合來挑選出本音節適當的 HMM 模型。作法上我們首先依訓練語料各音節發音的音標建立一張搜尋表，然後當輸入一個欲合成的文句時，就可依此表來搜尋出文句中各個音節對應的 HMM 模型編號。搜尋表

的格式如圖 5(a)，每列中各欄位的意義分別爲 <本音節>, <前音節韻母>, <後音節聲母>, <前音節聲母>, <後音節韻母>, <HMM 模型索引>，以有框線的那一列爲例來作說明，”a_ai_l_z_ian 1” 表示本次音節的發音是/a/，它的前一個音節發音爲/zai/，而後一個音節發音爲/lian/，當聲母爲空聲母或韻母爲空韻母時我們以/*/表示，該列最後的數字 1 表示該發音參加了編號 1 之 HMM 模型的訓練。如果輸入文句中有一個音節剛好被拆解成”a_ai_l_z_ian 1”，則可以搜尋到此有框線的列，而得知應使用/a/音節編號 1 的 HMM 模型。

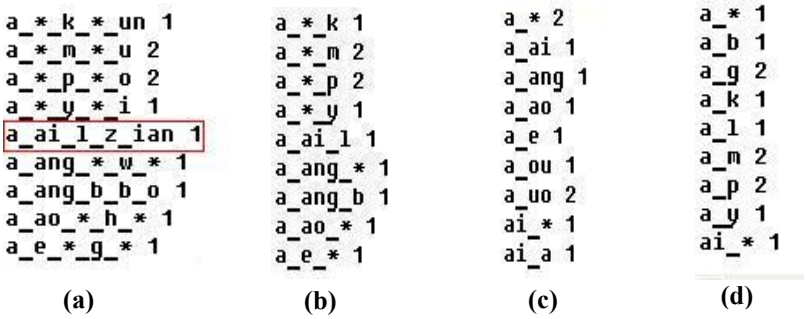


圖 5 HMM 模型搜尋表

然而隨機的文句輸入，可能會找不到完全一致的文依性關係，所以需要有替代方案，以避免找不到 HMM 模型。在此我們先將聲母、韻母作分類，同一分類的成員表示其發音特性較爲相似，當發生找不到完全一致的前後音節聲、韻母時，就可作爲替代尋找的目標，我們採取的分類方式，聲母部分的列於表 2，韻母部分的則列於表 3。

較詳細來說，我們設定的替代尋找之規則如下:

(規則 1): 將待搜尋音節的文依性條件的末二個成分去除，如此作搜尋的表格就變成如圖 5(b)所示。如果此表格仍然找不到，則一一替換前一音節韻母、後一音節聲母在同一類別裡的所有聲母、韻母，再作搜尋。若仍然找不到，則採取下一個替代規則。

(規則 2): 將待搜尋音節的文依性條件的末三個成分去除，如此作搜尋的表格就變成如圖 5(c)所示。如果此表格仍然找不到，則一一替換前一音節韻母在同一類別裡的其它韻母，再作模型的搜尋。若仍然找不到，則採取第三替代規則。

(規則 3): 將文依性條件的四個成分中的第 1 個及第 3、4 個去除，即即只保留後一音節之聲母，如此作搜尋的表格就變成如圖 5(d)所示。如果此表格仍然找不到，則一一替換後一音節聲母在同一類別裡的其它聲母，再作搜尋。若仍然找不到，則表示該音節只有一個 HMM 模型，就直接選取。

3.3 頻譜參數產生

挑選出一個待合成音節的頻譜 HMM 模型之後，接著要決定該 HMM 的 8 個狀態分別應駐留多少個音框。然而依據 3.1 節產生出的音長參數，我們已經知道總共的音框數量，設爲 *TA* 個音框。所以接下來我們就沿用前人的公式[12]:

$$T_k = D_k + \rho \cdot V_k \tag{2}$$

$$\rho = \left(TA - \sum_{i=1}^{8} D_i \right) \Big/ \sum_{i=1}^{8} V_i \tag{3}$$

來作各狀態的音框數分配，其中 D_k 表示 HMM 訓練時第 k 個狀態被駐留的平均音框數，

而 V_k 表示 D_k 的變異數。

確定各狀態被駐留的音框數之後，接著便可考慮如何產生各音框的 DCC 係數。除了考慮前人提出的最大似然法(maximum likelihood estimation, MLE)[11, 12]之外，我們還另外研究了二種逼近作法，分別稱爲線性內插法(LIA, linear interpolation approximation)與加權式線性內插法(WIA, weighted-linear interpolation approximation)，用以以加快 DCC 係數產生的速度，以便達成即時合成國語語音的目標。

(A) 線性內插法(LIA)

以圖 6 爲例，假設狀態 S_l 佔有 4 個音框，而狀態 S_r 佔有 7 個音框，t_l 表示狀態 S_l 所佔音框的中心時間位置，t_r 則是狀態 S_r 所佔音框的中心時間位置，則時刻 t 音框對應的比例值 $\eta(t)$ 的計算方式爲:

$$\eta(t) = (t - t_l) \,/\, (t_r - t_l) \tag{4}$$

依據比例值 $\eta(t)$，時刻 t 音框對應的第 j 維頻譜係數 $c_j(t)$ 就可以如下公式來算出:

$$c_j(t) = (1 - \eta(t)) \cdot cm_j^l + \eta(t) \cdot cm_j^r, \quad j = 0, ..., 38 \tag{5}$$

其中 cm_j^l 表示狀態 S_l 上平均頻譜向量的第 j 維，而 cm_j^r 表示狀態 S_r 上平均頻譜向量的第 j 維。

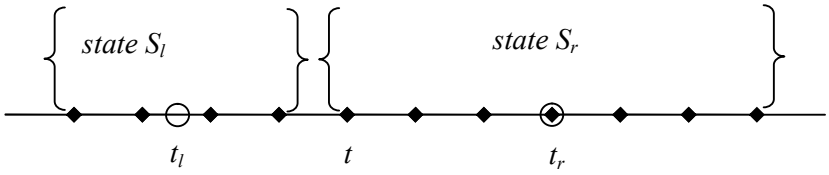


圖 6 線性內插例圖

(B) 加權式線性內插法(WIA)

此方法爲使用動態頻譜參數的另一種線性內插作法，此方法的細節配合圖 7 說明如下。假設我們要在兩相鄰狀態 S_i 與 S_{i+1} 之間插入數個音框，那麼首先計算兩狀態 S_i 與 S_{i+1} 的平均頻譜向量在各維之間的差異量(亦稱爲目標距離)，即

$$DF_j^i = cm_j^{i+1} - cm_j^i, \quad j=0, 1, 2, ..., 38, \tag{6}$$

其中 cm_j^i 表示狀態 S_i 上的平均頻譜向量的第 j 維。接著計算 AD_j^i，AD_j^i 表示在一般說話速度下狀態 S_i 會在第 j 維頻譜參數導入的距離的一半，即

$$AD_j^i = \Delta cm_j^i \times \frac{L_i}{2}, \quad j=0, 1, 2, ..., 38, \tag{7}$$

其中 Δcm_j^i 表示狀態 S_i 上平均差分頻譜向量的第 j 維，L_i 表示狀態 S_i 所佔據的音框個數。如此狀態 S_i 與 S_{i+1} 之間，在一般說話速度下會在第 j 維頻譜參數導入的距離就是

$$BD_j^i = AD_j^i + AD_j^{i+1}, \quad if \ \Delta cm_j^i \times \Delta cm_j^{i+1} \geq 0. \tag{8}$$

不過，相鄰狀態之間的平均差分頻譜參數 Δcm_j^i，可能會具有不同的正負號，因此當差分頻譜參數同號時，才以公式(8)計算狀態 S_i 與 S_{i+1} 之間導入的距離；而當差分頻譜參數不同號時，我們就簡化成前一小節的基本線性內插作法，所以直接令

$$BD_j^i = (L_i + L_{i+1}) / 2, \quad if \ \Delta cm_j^i \times \Delta cm_j^{i+1} < 0. \tag{9}$$

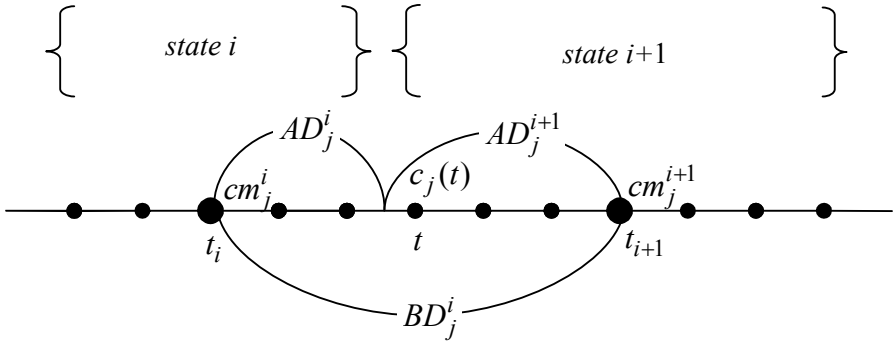


圖 7 加權式線性內插例圖

此時，目標距離 DF_j^i 與一般說話速度下所導入的距離 BD_j^i 的比例，就可以下式作計算，

$$R_j^i = DF_j^i \,/\, BD_j^i \ . \tag{10}$$

之後，依據比例值 R_j^i ，狀態 S_i 所管轄的在 t 時刻的音框，其第 j 維的頻譜參數 $c_j(t)$，便可以如下公式來計算出來。

$$c_j(t)=\begin{cases} cm_j^i+(t-t_i)\cdot\Delta cm_j^i\cdot R_j^{i\text{-}1}, & if\ \ t<t_i\ \ and\ \ \Delta cm_j^{i\text{-}1}\cdot\Delta cm_j^i\ge 0 \\ cm_j^i+(t-t_i)\cdot R_j^{i\text{-}1}, & if\ \ t<t_i\ \ and\ \ \Delta cm_j^{i\text{-}1}\cdot\Delta cm_j^i<0 \\ cm_j^i+(t-t_i)\cdot\Delta cm_j^i\cdot R_j^i, & if\ \ t\ge t_i\ \ and\ \ \Delta cm_j^i\cdot\Delta cm_j^{i+1}\ge 0 \\ cm_j^i+(t-t_i)\cdot R_j^i, & if\ \ t\ge t_i\ \ and\ \ \Delta cm_j^i\cdot\Delta cm_j^{i+1}<0 \end{cases} \tag{11}$$

(C) 逼近誤差、運算速度之比較

我們以最大似然法求解出的 DCC 係數作爲基準，來比較兩種內插法(LIA 與 WIA)的逼近誤差。實驗時分別使用 LIA 與 WIA 法來產生一篇文章共 131 個音節 3,506 個音框的 DCC 係數，並且細分成無聲部分與有聲部分，結果量測得到如表 5 所示之平均幾何距離之誤差數值。由表 5 可發現，使用了差分頻譜參數的 WIA 法有較小的幾何距離，也就是較能逼近最大似然法產生的 DCC 係數。此外，無聲部份之誤差距離，普遍會比有聲部分的大，其原因應是無聲音框的頻譜比較會有隨機的起伏，而較難以線性內插方式去逼近。

關於 DCC 係數的產生速度，我們以兩篇文章共 257 個音節、6,869 個音框(一個音框代表 11.61ms 的語音信號)來作測試，以比較三種方法(MLE, LIA, WIA)產生 DCC 係數的時間花費，量測結果如表 6 所示，由表 6 可發現 MLE 法產生 DCC 係數所花費的時間，遠遠大於 LIA 與 WIA 法，並且無法達到即時處理(81.6s > 6,896*11.61ms=79.75s)。

表 5 逼近誤差--平均幾何距離

	MLE	**LIA**	**WIA**
All frames	0	0.1740	0.1113
Voiced frames	0	0.1687	0.1051
Unvoiced frames	0	0.2049	0.1471

表 6 頻譜參數產生之時間花費

	MLE	**LIA**	**WIA**
花費時間	81.6s	0.67s	0.56s

3.4 音高頻率計算

在合成一個音節的語音信號時，有聲的部分才需要給予對應的基週軌跡，因此當一個音節含有無聲聲母時，我們必須先判斷它所對應的頻譜 HMM 裡，無聲、有聲部分各佔據多少個狀態。對於 HMM 的第 i 個狀態是否爲無聲或有聲的判斷，在此我們依據該狀態上的平均頻譜向量之最後一維(即 cm_{39}^{i})的數值，當 $cm_{39}^{i} > 0.5$ 時就判定爲有聲，不然則判定爲無聲，這樣的判定方式，很明顯地是基於參數 c_{39} 的定義(在 2.4 節)。當一個無聲聲母開頭之音節的頻譜 HMM 的狀態被逐一檢查時，如果檢查到第 j 個狀態被判定爲有聲，則很明顯地編號 j 之後的狀態也應都是有聲的，此時，我們就可將編號 j 及其後的各個狀態所分配到的音框數加總起來，而得知有聲部分佔據多少個音框。

決定有聲音框的數量之後(設有 Lv 個)，對於第 k 個有聲音框，可算出它的正規化時間爲 $x = (k\text{-}1) / (Lv - 1)$，另外令此音節的 16 個基週軌跡參數爲 $y_0, y_1, \ldots, y_{15}$，並且分別座落於正規化時間 $x_0, x_1, \ldots, x_{15}$ 上(即 $x_i = i / 15$)，依據 x，首先找出最靠近 x 的四個連續的正規化時間點，設爲 $x_n, x_{n+1}, x_{n+1}, x_{n+2}$ 等，接著就使用三階 Lagrange 內插，來算出第 k 個有聲音框的音高頻率值 $F(x)$，公式爲

$$F(x) = \sum_{j=0}^{3} \mathrm{y}_{n+j} \cdot \prod_{i=0, i\neq j}^{3} \frac{x - x_{n+i}}{x_{n+j} - x_{n+i}} \tag{12}$$

3.5 HNM 信號合成

首先依據各音框的 DCC 係數來算出頻譜包絡曲線，然後對於有聲的音框，直接將 MVF (maximum voiced frequency)訂爲 6,000Hz，即 6,000Hz 之前視爲諧波(harmonic)部分，之後視爲雜音(noise)部分[26]；此外設定兩相鄰音框之間的間隔爲 256 個樣本點，而取樣率則設爲 22,050Hz。接著，依據第 i 個音框的頻譜包絡 $S(f)$及音高頻率 F_0，求出諧波部分各個諧波的頻率 $F_k^i = k \cdot F_0$ 與振幅 $A_k^i = S(k \cdot F_0 \,/\, 22{,}050)$, k=1, 2, …, M_i，進一步再根據能量係數 c_0 來調整各諧波的振幅，其作法如下:

$$\overline{A}_k^i = R \cdot A_k^i \quad , \quad k=0, 1, ..., M_i \quad , \tag{13}$$

$$R = \sqrt{\exp(c_0) \Big/ \sum_{j=1}^{M_i} (A_j^i)^2} \quad .$$

之後，當要合成第 i 和第 i+1 音框之間時刻 t 的諧波信號樣本 $h(t)$時，我們先以如下公式作線性內差，

$$\begin{aligned} f(k,t) &= F_k^i + \frac{t}{256}(F_k^{i+1} - F_k^i), \quad k = 1,2,...,M \\ a(k,t) &= \overline{A}_k^i + \frac{t}{256}(\overline{A}_k^{i+1} - \overline{A}_k^i), \quad k = 1,2,...,M \end{aligned} \tag{14}$$

以求取時刻 t 時各諧波的頻率與振幅，其中 256 表示相鄰音框之間的樣本點數，M 是 M_i 和 M_{i+1} 的較大者，因此當 M_i 小於 M_{i+1} 時，就要把 $\overline{A}_k^i$, k= M_i +1,… , M_{i+1} 設爲零值。然後，以如下公式計算 $h(t)$，

$$h(t) = \sum_{k=1}^{M} a(k,t) \cdot \cos\big(\phi(k,t)\big), \quad 0 \le t < 256 \tag{15}$$

$$\phi(k,t) = \phi(k,t-1) + 2\pi \cdot f(k,t) / 22,050$$

其中$\phi(k,t)$表示第 k 個諧波累積到時刻 t 時的相位量，關於初始值$\phi(k,-1)$，我們令其等於前一音框裡的$\phi(k, 255)$以便維持相位的連續性，而當音框編號 i 爲 0 時就以亂數來設定。

關於雜音信號的合成，我們採取 HNM 文獻上提到的一個作法[26]，就是把雜音當作是 MVF 之後頻率間隔固定爲 100Hz、但振幅會隨時間改變之一些弦波的加總。先依 MVF 決定頻率下標之下限 ML = MVF / 100, 而其上限明顯地是 MU = 11,025 / 100，如此，對於第 i 和第 i+1 音框之間時刻 t 的雜音信號樣本 $g(t)$，我們以如下公式來計算，

$$\begin{aligned} g(t) &= \sum_{k=ML}^{MU} b(k,t) \cdot \cos\big(\psi(k,t)\big), \quad 0 \le t < 256 \\ \psi(k,t) &= \psi(k,t-1) + 2\pi \cdot k \cdot 100 / 22,050 \end{aligned} \tag{16}$$

其中$b(k,t)$表示時刻 t 時第 k 個弦波的振幅，其值也是以類似公式(10)之線性內差來求得，$\psi(k,t)$ 表示第 k 個弦波累積到時刻 t 時的相位量，其初始值也是以亂數來設定。最後，將 $h(t)$與 $g(t)$相加，即可得到時刻 t 的合成信號樣本。

四、語音合成與聽測實驗

4.1 語音合成實驗

我們拿一句參加頻譜 HMM 訓練的內部語句作參考，該語句的內容是“餐館早已打烊”，其錄音得到的信號波形與聲譜(spectrogram)如圖 8 所畫。另外，拿這個文句作輸入給我們的系統作合成處理，合成得到的信號波形及聲譜則如圖 9 所示。比較兩圖上半的聲譜部分，可發現圖 8 中的共振峰軌跡比圖 9 中的細瘦(即頻寬窄)、清楚，所以我們系統合成出的語音，在清晰度上仍然還不能和原始錄音的語音作比較。不過，圖 9 中在前四個字“餐館早已”的部分，各個共振峰軌跡的高度與走勢，大體上和圖 8 中的近似，所以合成出的語音，是可以清楚聽出說話內容的，而最後二字“打烊”部分，我們系統尚不能把共發聲(coarticulation)現象(音節邊界共振峰軌跡平順銜接)表現出來。

雖然我們系統具有前述的缺點，但是以聽覺的感受來講，其實我們系統所合成出的語音，說話內容是聽得清楚的，也就是理解度(intelligibility)應不是問題；並且具有不錯的韻律流暢度，可表現出整句性的音調高低變化；此外也具有音節內的聲學流暢度，這應是頻譜 HMM 發揮了功用。爲了讓有興趣的讀者下載、試聽所合成出的音檔，我們準備了一個網頁以供瀏覽，網址爲 http://guhy.csie.ntust.edu.tw/hmmsyn/ .

爲了測試我們系統作合成處理的速度是否能達到即時處理，在此以一篇 131 個字的文章作爲輸入，並且使用加權式線性內插法來產生 DCC 係數，測試的平台是一台使用 Intel T5600 1.83GHz CPU 之筆記本電腦，結果共花費 29.07 秒之處理時間，而合成出的音檔長度爲 49.21 秒，所以我們系統的確可達成即時處理。

4.2 聽測實驗

我們在不同條件下合成出語音音檔，用以作合成語音的聽測評估，但是輸入到系統作合成的文字內容都一樣，內容爲節錄一篇文章開頭的一小段：“因爲不知道你的名字, 就讓我叫你白花樹, 春天, 當你的花朵盛開時, 就像點亮了滿樹白蠟燭.”。聽測時，由受測

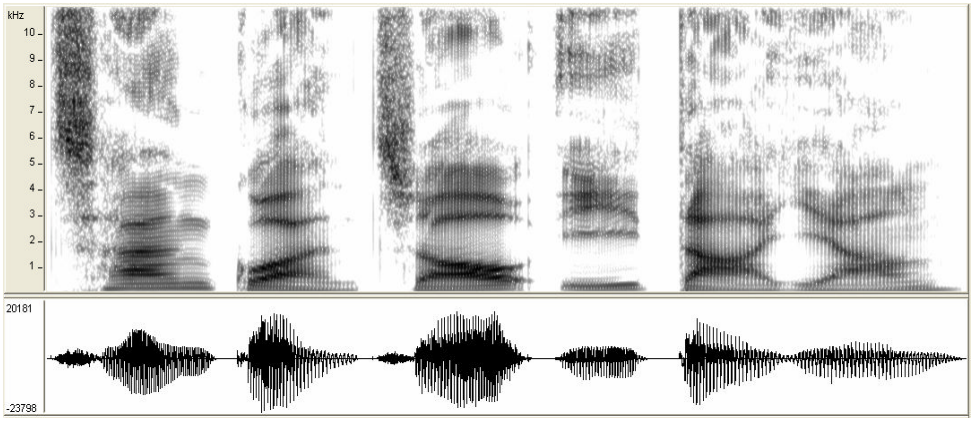


圖 8 錄音語句“餐館早已打烊”之聲譜圖與信號波形

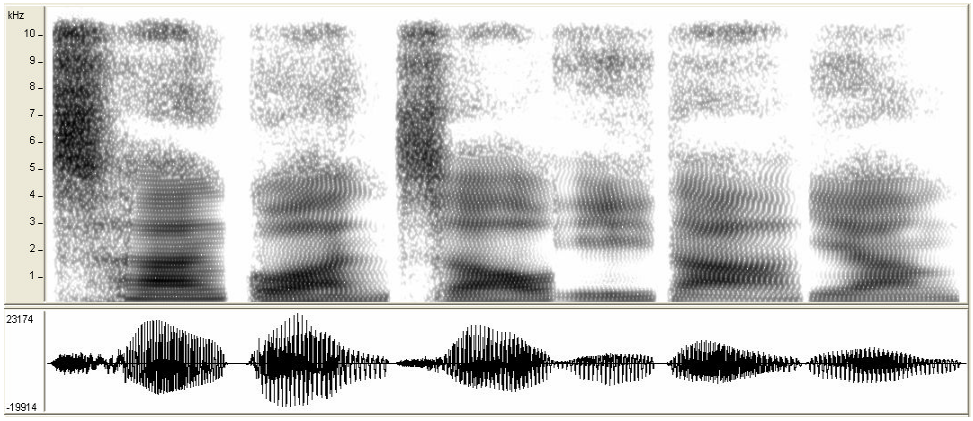


圖 9 合成語句“餐館早已打烊”之聲譜圖與信號波形

者主觀比較兩個合成語音檔 SX 與 SY 的自然度差異，自然度包含了合成語音的清晰性與流暢性，也就是要評估合成語音的感覺接近一般人說話的程度，在不同的聽測項目裡，SX 與 SY 將代表不同的合成音檔。在各個聽測項目的實驗裡，我們都請了 15 位受測者來作評分，其中多數人沒有語音合成方面的經驗。當一位受測者依序聽完 SX 與 SY 之音檔後，他會被要求給一個評分，作爲 SY 比 SX 的自然度差異情形，在此評分值的範圍是由-2 至 2，分別的意義是，2 (或-2) 表示 SY 比 SX 好很多(或差很多)，1 (或-1) 表示 SY 比 SX 好一些 (或差一些)，0 則表示二者不能被分辨。

第一個聽測項目是關於 DCC 係數產生之方法。在此 SX 表示使用 MLE 產生法所合成出的音檔，而 SY 則表示使用 WIP 產生法所合成出的音檔，其它因素則控制爲一致。聽測之後，我們對 15 位受測者所給的評分求取平均值與標準差，結果得到平均評分爲 0.80，而其標準差爲 0.528。由此可知，使用 WIP 法來產生 DCC 係數，反而可得到較高的評分。從受測者得到的意見是，MLE 法合成出的音檔聽起來過於平穩且略顯呆板，而不像一般人那樣具有活潑的說話方式，相對地加權式線性內插法合成出的音檔，在聽覺上會有比較活潑的感覺。

第二個聽測項目是關於音節音長的產生方法，以了解不同的音長產生方法對於聽覺上流暢性的影響。在此 SX 表示使用一個音節 HMM 各狀態的平均駐留音框數的加總作爲該音節的音長，這相當於把公式(2)的 ρ 係數設成 0 值，而 SY 表示使用音長 ANN 模型來得到各音節的音長，至於其它因素則控制爲一致。聽測之後，將 15 位受測者所給的評分集中並求取平均值與標準差，結果得到平均評分爲 1.03，而其標準差爲 0.352。所以使用音長 ANN 模型所產生的音節音長，在聽覺上可得到較佳的流暢性表現。

第三個聽測項目是比較本論文系統和先前製作的國語語音合成系統[10]。先前系統是以 DTW 匹配分離發音的參考音和語句發音的訓練音，來取得頻演參數用以訓練頻演 ANN 模型，再依據頻演 ANN 模型產生出的頻演參數，去控制 HNM 信號模型作時間軸對映，細節部分可參考我們先前的論文 [10]。本論文系統在此的設定是，使用音長 ANN 模型來產生音節的音長，再搭配加權式線性內插法去產生出 DCC 係數。作聽測比較時，以先前系統合成出的音檔作爲 SX，而以本論文系統所合成出的音檔作爲 SY。聽測之後，依據 15 位受測者給的評分所算出的平均分數是 1.567，而標準差則是 0.530。所以，本論文系統以音節 HMM 模型來產生 DCC 頻譜參數，再據以算出頻譜包絡的作法，比起先前系統使用頻演 ANN 模型來產生頻演參數，再據以作時間軸對映的作法，會得到顯著的自然度改進。

五、結語

本論文研究提出一種結合 ANN 韻律模型、HMM 頻譜模型及 HNM 信號模型之國語合成系統之架構，這種結合代表的意義是，我們可以使用優秀的韻律模型(也許是別人建造的)來產生韻律參數(如音節音長、基週軌跡)，而不必限制於只使用 HMM 來含蓋語音的各種特性(如韻律流暢性、聲學流暢性)。然而使用 HMM 模型來掌握音節內部的頻譜演進方式，的確可讓 HMM 發揮專長。當以兩種 ANN 分別產生出音節音長和基週軌跡參數後，再使用 HMM 模型來產生出一音節各音框的頻譜參數，然後用以控制 HNM 作語音信號的合成，如此將可同時提升國語合成語音的**韻律**與**聲學**流暢性。

關於 HMM 頻譜模型所使用的頻譜參數，我們使用離散倒頻譜爲基礎的頻譜包絡估計方法，來計算出訓練語料各音框的 DCC 係數，並且以音框能量取代 DCC 的 c_0 係數、以基頻偵測的週期性參數來作爲 DCC 的 c_{39} 係數，實驗顯示這是可行的頻譜參數選擇。對於欲合成音節各音框的 DCC 係數的產生，首先以音長 ANN 模型產生出音節音長值，再據以決定頻譜 HMM 各狀態應停留的音框個數；然後，除了考慮以 MLE 法來產生 DCC 係數之外，我們也另外研究了二種內插方法(LIP 與 WIP 法)來加快產生各音框的 DCC 係數；實驗的結果顯示，WIP 法比 LIP 法具有較小的逼近誤差，並且 WIP 法比 MLE 法可獲得非常多的速度改進，而達成即時處理的目標，此外聽測實驗也顯示，使用 WIP 法所合成出的語音，其自然度反而是會比 MLE 法的好。

雖然我們系統能夠合成出自然度還不錯的國語語音信號，但是從聲譜圖上作觀察，合成語音的共振峰軌跡和真實說話語音的共振峰軌跡之間，仍然存在明顯的差異，也就是合成語音的共振峰軌跡比較粗胖而不夠細瘦，並且相鄰音節邊界上的共發聲現象，也未能表現出來，所以未來可再考量如何改進這樣的缺點。此外，前人文獻中提到的二次差分之頻譜參數($\Delta\Delta c_i$)，將來也可考慮加以應用。

參考文獻

[1] J. D. Markel and A. H. Gray Jr., *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

[2] D. O'Shaughnessy, *Speech Communications: Human and Machine*, IEEE Press, Piscataway, NJ, 2000.

[3] Y. Sagisaka, *et al.*, "ATR v-talk speech synthesis system," *Int. Conf. Spoken Language Processing*, Banff, Alberta, Canada, pp. 483-486, 1992.

[4] J. H. Chen, *A study on Synthesis Unit Selection and Prosodic Information Generation in a Chinese Text-to-speech System*, Ph.D. Dissertation, Department of Electrical Engineering, National Cheng

Kung University, Tainan, Taiwan, 1998.
[5] F. C. Chou, *Corpus-based Technologies for Chinese Text-to-speech Synthesis*, Ph.D. Dissertation, Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, 1999.
[6] M. Chu, H. Peng, H. Y. Yang, and E. Chang, "Selecting Non-uniform Units from a Very Large Corpus for Concatenative Speech Synthesizer", *Int. Conf. Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, pp. 785-788, 2001.
[7] H. Y. Gu and W. L. Shiu, "A Mandarin-syllable Signal Synthesis Method with Increased Flexibility in Duration, Tone and Timbre Control," *Proc. Natl. Sci. Counc. ROC(A)*, Vol. 22, pp. 385-395, 1998.
[8] H. Y. Gu and C. C. Yang, "A Sentence-pitch-contour Generation Method Using VQ/HMM for Mandarin Text-to-speech", *Int. Symposium on Chinese Spoken Language Processing*, Beijing, China, pp. 125-128, 2000.
[9] H. Y. Gu and Y. Z. Zhou, "An HNM Based Scheme for Synthesizing Mandarin Syllable Signal", *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 13, No. 3, pp. 327-342, 2008.
[10] H. Y. Gu and C. Y. Wu, "Model Spectrum-progression with DTW and ANN for Speech Synthesis", in *Proc. ECTI-CON 2009*, Pattaya, Thailand, pp. 1010-1013, 2009.
[11] K. Tokuda, *et al.*, "Speech Parameter Generation Algorithms for HMM-based Speech Synthesis", *ICASSP*, Istanbul, Turkey, pp. 1315-1318, 2000.
[12] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based Approach to Multilingual Speech Synthesis", in Text to Speech Synthesis: New Paradigms and Advances, Editors: S. Narayanan and A. Alwan, Prentice Hall, NJ, pp. 135-153, 2004.
[13] Y. Qian, F. Soong, Y. Chen, and M. Chu, "An HMM-Based Mandarin Chinese Text-to-Speech System", *Int. Symposium on Chinese Spoken Language Processing*, Singapore, pp. 223-232, 2006.
[14] M. Zhang, J. H. Tao, H. B. Jia, and X. Wang, "Improving HMM Based Speech Synthesis by Reducing Over-smoothing Problems", *Int. Symposium on Chinese Spoken Language Processing*, Kunming, China, pp. 1-4, 2008.
[15] 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, pp. 226-239, 1998.
[16] M. S. Yu, N. H. Pan, and M. J. Wu, "A Statistical Model with Hierarchical Structure for Predicting Prosody in a Mandarin Text-to-speech System," *Int. Symposium on Chinese Spoken Language Processing,* Taipei, Taiwan, pp. 21-24, 2002.
[17] 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, pp. 309-324, 2004.
[18] W. H. Lai, *A Statistic Prosodic Modeling for Mandarin Speech*, Ph.D. Dissertation, Institute of Communication Engineering, National Chiao Tung University, Hsinchu, Taiwan, 2003.
[19] T. Galas and X. Rodet, "An Improved Cepstral Method for Deconvolution of Source Filter Systems with Discrete Spectra: Application to Musical Sound Signals", *International Computer Music Conference*, Glasgow, Scotland, pp. 82-44, 1990.
[20] 古鴻炎、蔡松峯，"基於離散倒頻譜之頻譜包絡估計架構及其於語音轉換之應用"，第 21 屆自然語言與語音處理研討會(ROCLING 2009)，台中，第 151-164 頁，2009。
[21] Kim, H. Y., *et al.*, "Pitch Detection with Average Magnitude Difference Function Using Adaptive Threshold Algorithm for Estimating Shimmer and Jitter", Proc. of the 20th Annual International Conference in Medicine and Biology Society, pp. 3162-3164, 1998.
[22] 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 16 屆自然語言與語音處理研討會(ROCLING XVI)，台北，第 325-334 頁, 2004。
[23] T. Calinski and J. Harabasz, "A Dendrite Method for Cluster Analysis", Communications in Statistics, Vol. 3, No. 1, pp. 1-27, 1974.
[24] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.
[25] H. Y. Gu, Y. Z. Zhou, and H. L. 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.
[26] Y. Stylianou, *Harmonic plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification,* Ph.D. thesis, Ecole Nationale Supèrieure des Télécommunications, Paris, France, 1996.