The 2014 Conference on Computational Linguistics and Speech Processing
ROCLING 2014, pp. xx-xx


# 基於發音知識以建構頻譜 HMM 之國語語音合成方法

# A Mandarin Speech Synthesis Method Using Articulation-knowledge Based Spectral HMM Structure

古鴻炎*、賴名彥*、洪尉翔*、陳彥樺*

**Hung-Yan Gu, Ming-Yen Lai, Wei-Siang Hong, and Yan-Hua Chen**

**摘要**

在有限語料的情況下，本論文提出一種 HMM 的結構設計，來掌握各個語音單元之文脈相依的頻譜特性，以便改進合成語音的流暢度。此外，在決策樹之文脈分群方法之外，我們依據音素的發音知識，來作文脈分群而大幅降低文脈組合數量。為了評估所提出的 HMM 結構，我們使用三種不同的 HMM 結構方式去建造對應的國語語音合成系統，以作相互的比較。在這些系統裡，使用的韻律參數值是一樣的，都是使用之前研究的 ANN 模組來產生；但是頻譜係數則是使用各自的 HMM 模型來產生；至於信號波形的合成，則都是使用之前研究的基於諧波加雜音模型(HNM)的信號合成模組。聽測實驗的結果顯示，使用本論文提出的 HMM 結構所合成出的語音，比用其它 HMM 結構所合成的明顯地更為流暢；此外，依據錄音語句與合成語句之間的平均頻譜距離的量測結果，也顯示本論文的 HMM 結構，比其它 HMM 結構更能夠降低頻譜距離。

**關鍵詞:** 語音合成、HMM 結構、發音知識、頻譜流暢度、離散倒頻譜係數

**Abstract**

In this paper, a new HMM structure is proposed to work with a limited training corpus in order to obtain improved synthetic-speech fluency. Spectral fluency is improved because this HMM structure can model the context-dependent spectral characteristics of a speech unit. In addition, instead of using a decision tree to cluster contexts, the knowledge of phoneme articulation is based to cluster contexts and reduce the enormous quantity of context combinations. To evaluate the proposed HMM structure, we construct three Mandarin speech synthesis systems each uses one different HMM structure for comparisons. In these systems, the prosodic parameters are all generated with same ANN modules studied previously

*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, M9615074, M10115035, M10215005}@mail.ntust.edu.tw

but the spectral coefficients are generated with different HMM adopted by its corresponding system. As to the synthesis of signal waveform, the signal model, harmonic plus noise model (HNM), studied previously is commonly adopted in the three systems. According to the results of listening tests, the speech synthesized by the system using the proposed HMM structure is indeed more fluent than the speeches synthesized by the other two systems. In addition, average spectral distances are measured between recorded sentences and synthetic sentences. The results show that the HMM structure proposed here also obtains smaller average spectral distance than the other two HMM structures.

**Keywords:** Speech Synthesis, HMM Structure, Articulation Knowledge, Spectral Fluency, Discrete Cepstral Coefficients.

## 1. 緒論

近年來許多研究者早已利用隱藏式馬可夫模型(hidden Markov model, HMM)，來建造語音單元(如音素、音節等)之頻譜演進(spectrum progression)模型(Yoshimura *et al.*, 1999; Zen *et al.*, 2007; Yan *et al.*, 2009; Gu *et al.*, 2010)，之後在合成一個語句時，就會使用訓練得到的 HMM 模型來產生一序列的頻譜特徵向量，然後使用所產生的頻譜特徵向量序列去合成出語音信號。使用 HMM 來作語音信號的合成，通常能夠獲得增進的可理解性(intelligibility)與流暢性(fluency)。更好的是，Tokuda 等人基於 HTK (HMM tool kits)所發展的 HTS 語音合成軟體(Zen *et al.*, 2007)，提供公開的程式原始碼、並且可供下載，所以研究語音合成時，使用 HTS 能夠減少許多時間與氣力。不過，當未使用全域變異數(global variance, GV)匹配時，HTS 軟體所產生的頻譜包絡會發生過於平滑的現象，使得合成出的語音變得悶悶的(muffled) (Toda & Tokuda, 2005)。

在本論文中，我們並不打算沿用、修改 HTS 的程式碼，因此必須自行發展 HMM 模型訓練的程式、及頻譜特徵向量序列之產生程式，這樣的決定是因為，我們想要研發一個具有彈性(flexibility)的語音合成系統，能夠容易地擴增額外的功能。例如音色轉換(timbre transformation)之功能，能夠把合成語音的音色從成年女性轉變成男孩(Gu & Tsai, 2013)；另一項預計擴增的功能則是，同步地播放合成出的語音信號及其對應的拼音符號，此功能可應用於人形機器人上，讓語音發聲和嘴型同步。除了擴增額外功能的原因之外，我們與其他研究者使用 HTS 的經驗(Hsia *et al.*, 2010)，發現 HTS 所產生的國語語音的基週軌跡(pitch contours)聽覺上並不令人滿意，因此本論文建造的國語語音合成系統，就決定使用不同的方法來產生各個音節的基週軌跡。

在先前的一次研究中(Gu *et al.*, 2010)，我們曾嘗試去建立音節單位的 HMM 模型，以掌握音節內的頻譜演進(spectrum progression)方式，但是此 HMM 模型所合成出的語音信號並不夠流暢，在音節邊界處的頻譜不連續(spectral discontinuities)情形經常會被聽出來，我們認為發生頻譜不連續的原因是，設定的語音單位"音節"太大，使得一個音節 X 和前後音節所組合出的不同文脈數量非常龐大，以至於無法為該音節 X 建立良好的文脈

相依 HMM 模型。因此，本論文決定使用較小的語音單位，即聲母和韻母，接著根據標記檔裡記載的拼音符號序列，去建構文脈相依的 HMM 模型，並且依據發音知識來對 HMM 模型作分組訓練。HMM 結構的設計方式將會在第二節詳細說明。

我們建造的國語語音合成系統，其合成階段之處理流程如圖 1 所畫，區塊(a)進行輸入文句之分析；區塊(b)使用類神經網路(Artificial Neural Network, ANN)模組來產生各個音節的基週軌跡參數和時長(duration)值，詳細作法可參考(Gu & Wu, 2009)；區塊(c)依據第二節介紹的挑選方法來為各個聲韻母選出對應的 HMM 模型；區塊(d)採用 Tokuda 等人提出的方法(Yoshimura *et al.*, 1999)，去計算 HMM 各狀態分配到的時長音框數；區塊(e)採用先前研究提出的加權線性內插法(Gu *et al.*, 2010)，去產生各音框的頻譜係數；區塊(f)依據基週軌跡參數去計算出各個有聲音框的音高頻率(pitch frequency)值；最後，區塊(g)使用諧波加雜音模型(Harmonic plus noise model, HNM)，去合成出信號波形，詳細作法可參考(Gu & Tsai, 2013)。

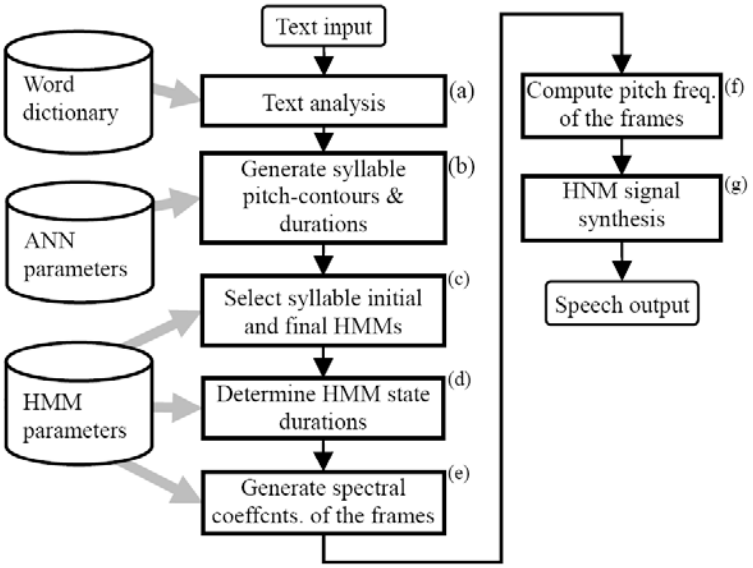


***圖 1. 合成階段之主要處理流程***

## 2. 聲韻母分類及文脈組合

在語音辨識領域，基本的 HMM 結構就如圖 2 所示之左至右(left to right)結構，然而此結構並未處理左右兩邊界的文脈相依關係，因此其效能(辨識率)會顯現衰退的情形。

在傳統上，一個國語音節被分割成聲母(開頭子音)和韻母(結尾母音群)兩個部分，若採取聲、韻母作為語音單位，則一個語音單位的左右文脈之組合量可以比音節單位時的文脈組合量大幅減少。但是採用聲、韻母作為語音單位時，可能組合出的文脈量仍是非常龐大，如韻母的前後文脈組合量約有十二萬四千個，詳細數目是(21+37) × 37 × (21+37)

= 124,468，其中 21 表示 21 種聲母，37 表示 37 種韻母，因此有需要再對聲、韻母作進一步的分類。

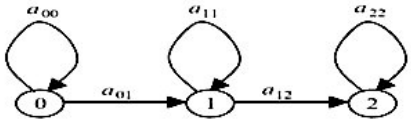


*圖 2. 左至右 HMM 結構*

在一個聲母的左邊，可能遇到前一音節的韻母或靜音，在此我們把韻母尾端之可能發音口形分類成 11 類，詳細分類方式如表 1 所列(含靜音"sil")。相對地，在一個聲母的右邊只會遇見韻母，在此我們把韻母開頭之可能發音口形分類成 8 類，詳細分類方式如表 2 所列。當把聲母前後可能遇到的韻母發音口形作了分類之後，再考慮國語共有 21 個聲母，由此可推算出聲母可能組合出的文脈數量是(11+1)×21×8=2,016 個。

*表 1. 韻母結尾之發音口形分類*

| Index | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gesture class | a | o | ə | e | i | u | yu | ii | er | n | ng | sil |

*表 2. 韻母開頭之發音口形分類*

| Index | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Gesture Classes | a | o | ə | e | i | u | yu | ii |

類似於前一段的敘述，如果目前音節沒有聲母時，則在此音節韻母的左邊，將會遇到前一音節的韻母或靜音，在此也把前一音節韻母尾端之可能發音口形分成 11 類，如表 1 所示。另一種情況，當韻母左邊遇到的是本音節的聲母時，在此我們根據聲母的發音位置把可能遇到的聲母分成 6 類，詳細分類方式如表 3 所列。舉例來說，/b/, /p/, /m/, /f/ 的發音位置皆在嘴唇，所以它們都被分類至"b"類別；同樣道理，/d/, /t/, /n/, /l/的發音位置皆在齒槽，所以把它們都分類至"d"類別。

*表 3. 聲母依發音位置之分類*

| Index | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Consonant Classes | b<br>ㄅ | d<br>ㄉ | z<br>ㄗ | zh<br>ㄓ | j<br>ㄐ | g<br>ㄍ |

考慮目前音節韻母的右邊可能遇到的語音單元，第一種情況是，後接的音節沒有聲母，在此我們把後接音節的韻母開頭之可能發音口形分成 9 類，詳細分類方式如表 1 中的 9 個類別，但不包含/n/和/ng/；第二種情況是，後接的音節具有聲母，在此我們把後

接音節聲母的可能發音口形分成 6 類，詳細分類方式如表 3 所列；第三種情況是，後接音節不存在(即目前音節是語句的最後音節)，相當於後面銜接的是靜音。當把韻母前後可能遇到的聲、韻母發音口形作了分類之後，再考慮國語共有 37 個韻母，由此可推算出國語韻母可能組合出的文脈數量是(11+6+1) × 37 × (9+6+1)=10,656 個。

如果我們依據前述的文脈組合方式去建造 HMM 模型，則需要訓練的 HMM 模型會有 2,016 + 10,656 個，這意味著在準備訓練語料時，必須錄製數倍於 10,656 個音節數量的音節發音。然而準備大量的訓練語料需要付出昂貴的費用，這暗示使用前述之文脈相依 HMM 模型是不切實際的。對於此問題的解決辦法，先前研究者提出使用決策樹(decision tree)來對 HMM 模型作分群，再對各群分別去訓練一個共用的 HMM，如此就可大幅降低所需的訓練語料數量，如 HTS 軟體就是採取此種作法。不過，本論文採取另外一種研究方向(approach)，就是先依據發音知識來對聲、韻母作分類(這相當於對 HMM 作分群)，再研究新的 HMM 結構之設計，以便解決前述的需求大量訓練語料的問題。

## 3. 建構新的HMM結構

如圖 3 所畫的是擁有 6 個狀態之韻母 HMM 的原本結構，此結構表示該韻母 HMM 是左右文脈相依的，符號 FY_XY 表示此 HMM 是韻母 Y 的模型(F 表示左右文脈相依)，並且 CX 表示韻母 Y 前接的文脈樣式(context type)，CZ 則表示 Y 後接的文脈樣式。根據表 1 和表 3 得知文脈樣式 CX 共有 12+6=18 種；相對地，從表 1(除了/n/與/ng/之外)及表 3 得知文脈樣式 CZ 共有 10+6=16 種。此外，國語有 37 種韻母 Y，所以國語韻母的文脈相依之 HMM 模型總計多達 10,656 個。

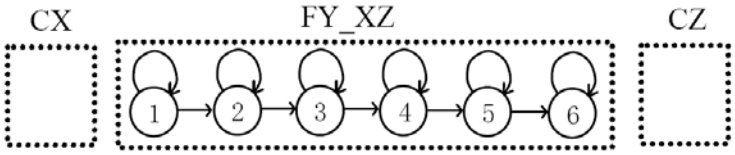


*圖 3. 左右文脈相依之 HMM 模型結構*

為了減少實作上需要投入的費用與人力(例如準備大量的訓練語料)，因此我們嘗試對圖 3 的 HMM 模型 FY_XZ 去重作結構安排。我們的解決方法是，假設圖 3 中前面半數的狀態(即狀態 1、2、3)會相依於前接文脈 CX，但是和後接文脈 CZ 不相關；類似地，我們也假設圖 3 中後面半數之狀態(即狀態 4、5、6)只相依於後接文脈 CZ，但是和前接文脈 CX 不相關。根據前述的兩項假設，圖 3 裡左右文脈依之 HMM 模型 FY_XZ，就可以被分解成圖 4 與圖 5 之半段式 HMM 模型 GY_X 和 HY_Z，亦即我們要以半段式 HMM 模型 GY_X 和 HY_Z 之串接來取代 HMM 模型 FY_XZ。

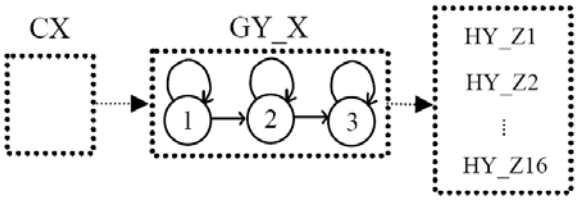


*圖 4. 韻母 Y 前半段之左文脈相依 HMM 結構*

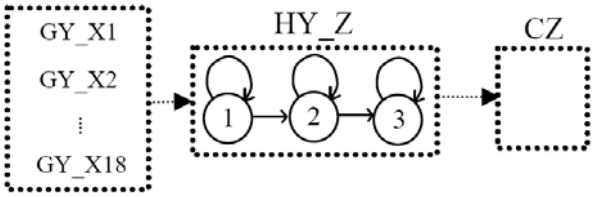


*圖 5. 韻母 Y 後半段之右文脈相依 HMM 結構*

由於一個韻母前接的文脈被分類成 18 種文脈樣式，所以我們需要建立 18 個半段式 HMM 模型 GY_X1, GY_X2, …, GY_X18 (如圖 5 裡列出)，來掌握韻母 Y 的前半段部分，符號 GY 之 G 表示半段式 HMM 之前半段。類似地，一個韻母後接的文脈被分類成 16 種文脈樣式，所以我們需要建立 16 個半段式 HMM 模型 HY_Z1, HY_Z2, …, HY_Z16 (如圖 4 裡列出)，來掌握韻母 Y 的後半段部分，符號 HY 之 H 表示半段式 HMM 之後半段。如此，一個韻母需要建立的半段式 HMM 模型數量是 18+16=34 個，而國語有 37 種韻母，所以總共需要建立的半段式 HMM 的數量是，34 × 37 = 1,258 個，1,258 比起 10,656 個左右文脈相依之韻母 HMM 少了許多。

關於國語 21 個聲母的 HMM 模型的建立，我們把前述之韻母 HMM 模型的假設套用進來。由於一個聲母前接的文脈被分類成 12 種文脈樣式，所以我們需要為一個聲母建立 12 個半段式 HMM 模型，來掌握該聲母的前半段部分；此外，一個聲母後接的文脈被分類成 8 種文脈樣式，所以我們需要為一個聲母建立 8 個半段式 HMM 模型，來掌握該聲母的後半段部分。如此，一個聲母需要建立的半段式 HMM 模型數量是 12+8=20 個，而國語有 21 種聲母，所以總共需要建立的半段式 HMM 的數量是，20 × 21 = 420 個，420 比起 2,016 個左右文脈相依之聲母 HMM 少了許多。

## 4. 實驗評估

### 4.1 訓練階段

我們邀請了一位成年男性至隔音錄音室錄製 1,208 個語句的發音，這些語句的腳本(script)是隨機地從無關的文章中挑選出，總計有 10,173 個音節，而錄音的取樣率為 22,050 Hz。為了標記這些語句中各音節的拼音符號，我們首先使用 HTK 套件來作 forced alignment 處理，而得到初步的音節邊界之標記，然後以人工操作 WaveSurfer 軟體去更正錯誤的音節邊界標記。

我們將各個音節的音檔切割成一序列的音框，音框長度設為 512 個樣本點，而音框位移設為 128 個樣本點。每個音框經分析計算後擷取出 39 個頻譜參數 $c_0, c_1, \ldots, c_{38}$，實際上是離散倒頻譜係數(discrete cepstral coefficients, DCC) (Cappé & Moulines, 1996)，DCC 係數的擷取方法，請參考我們先前的研究論文(Gu & Tsai, 2009)。此外，一個音框的週期性資訊也會被記錄在另一個維度中，即存入 $c_{39}$，如果一個音框被偵測出是週期性的，則設定 $c_{39}$ 的值為 1，反之則設定 $c_{39}$ 的值為 0。一個 HMM 經過訓練之後，我們就可以依據平均向量的 $c_{39}$ 的值，來判斷各個 HMM 狀態是否為有聲(voiced)或無聲(unvoiced)之狀態，然後就可為有聲的 HMM 狀態去產生基週軌跡。另外，頻譜參數的差分值也是有用的，所以我們把頻譜特徵向量增加 40 維，以儲存 DCC 係數的一階差分值。在訓練完一個 HMM 之後，除了記錄 HMM 的模型參數之外，也要記錄各個 HMM 狀態被駐留的音框個數之平均值與變異數。

在本論文中，一個韻母的前半與後半部分之半段式 HMM 模型皆是由 3 個狀態建造而成，並且狀態移轉方式都是由左至右，就如圖 4 和圖 5 所示。不過，對於一個聲母的半段式 HMM 模型，我們僅使用 2 個狀態去建造。關於高斯混合組件(Gaussian mixture component)的數量，在每一個 HMM 狀態上，我們只設置一個高斯混合。對於半段式 HMM 之訓練，我們依據分段式 K 中心(segmental K-means)演算法(Rabiner & Juang, 1993)，去發展了訓練程式。

## 4.2 語音合成處理

本論文製作的國語語音合成系統，其合成階段之處理流程如圖 1 所示，區塊(a)作 文句分析(text analysis)，每次會從輸入的檔案讀取一個文句進來，然後以查詞典及檢查數個構詞規則的方式去作剖析，把讀入的文句切割成一序列的詞語(words)，並且每一個詞語經由查詞典也可得知它的拼音符號。接著，在區塊(b)產生各音節之音高軌跡(pitch contour)及時長(duration)值，對於每一個音節，先準備好它的文脈資料項，再將文脈資料輸入至兩個類神經網路(ANN)，以分別預測出音高軌跡參數和時長參數的值，關於 ANN 的輸出/輸入資料項、及結構的細節，請參考我們先前的研究論文(Gu & Wu, 2009)。

在區塊(c)挑選聲母、韻母之 HMM 模型，首先依據各個音節的拼音符號去查詢出對應的聲母和韻母之拼音符號與編號，並且決定聲、韻母在表 1、表 2、與表 3 的分類編號，然後依據各個單元(聲母或韻母)的編號和其前後文脈的分類編號，從訓練出的半段式 HMM 模型中，找出一個單元(聲母或韻母)對應的前、後兩個半段式 HMM，接著把聲、韻母的四個半段式 HMM 依序串接成一個音節的完整 HMM 模型。在區塊(d)決定各個

HMM 狀態的駐留音框數，我們採用 Tokuda 等人提出的方法(Tokuda *et al.*, 2004)，依據 ANN 產生的音節時長值，去計算一個音節 HMM 之各個狀態所應分配到的時長音框數。接著，在區塊(e)產生各音框之頻譜係數(即 DCC 係數)，一個常被使用的方法是最大似然估計法(Maximum likelihood Estimate, MLE) (Tokuda *et al.*, 2004)，不過，本論文使用的是先前我們提出的加權式線性內插法(weighted linear interpolation, WLI) (Gu *et al.*, 2010)，來計算各個音框的 DCC 係數。

在區塊(f)計算各個音框的音高值，一個有聲單元(例如聲母/m/和韻母/a/)的各個音框都必須被指派一個音高頻率值(單位 Hz)，在此我們拿 ANN 產生的基週軌跡參數去作拉格蘭日內插(Lagrange interpolation)，以求得各個音框的音高頻率值。接著，在區塊(g)採用 HNM 信號模型作語音信號合成，我們把一個單元各音框的 DCC 係數與音高頻率值，按照音框次序逐個音框送給 HNM 語音信號合成模組，去合成出語音信號，關於 HNM 信號合成之細部處理方法，可參考我們先前的研究論文(Gu & Tsai, 2013; Gu & Tsai, 2009)。

## 4.3 頻譜距離量測

在此以代號 SYC 與 SYD 表示本研究所建造的兩個國語語音合成系統，SYC 表示完整的系統，對於一個音節的前接與後接文脈都納入考慮，這意味聲母的前半段 HMM 的左文脈有作區分(例如圖 4 之 GY_X)，並且韻母的後半段 HMM 的右文脈也有作區分(如圖 5 之 HY_Z)，所以對於各個不同的文脈樣式分類(如 CX)，就需要去建造一個對應的半段式 HMM (如 GY_X)。相反地，在簡化的系統 SYD 裡，對於一個音節的前接與後接文脈就不去作區分了，亦即聲母的前半段 HMM 不去區分它的左文脈，如此一個聲母就僅需訓練出一個前半段之 HMM，然而聲母後半段之 HMM，則仍然需訓練出數個半段式 HMM，以區分右文脈；類似道理，韻母的後半段 HMM 就不去區分它的右文脈，如此一個韻母就只需訓練出一個後半段之 HMM，然而韻母前半段之 HMM，則仍然需訓練出數個半段式 HMM，以區分左文脈。此外，以 SYP 表示先前研究裡所建造的國語語音合成系統(Gu *et al.*, 2010)，在 SYP 系統裡，我們對於每一種國語音節都建造了一個或數個音節寬度的 HMM，至於一種音節所建造的 HMM 個數，則和該音節在訓練語句中的發音次數有關。在本研裡，我們使用相同的訓練語句，來訓練這三個系統: SYC、SYD 和 SYP。

藉由這三個系統，我們就可以把錄音語句與合成語句相對應的音框拿去作頻譜距離的計算。詳細情形是，把 50 句測試語句的標記檔逐一輸入給 SYC、SYD 和 SYP 系統去處理，以取得三個系統對應於各句測試語句的的合成音檔，然後分別拿各句的錄音語句去和合成語句去作 DCC 分析，以計算出兩個 DCC 頻譜特徵向量的序列。接著，偵測兩 DCC 序列中各組對應音框是否都為有聲，若對應的音框都偵測為有聲，就拿該組音框去計算音框之間的 DCC 向量幾何距離，然後我們拿 50 句測試語句的所有有聲音框算出的幾何距離，計算出一個跨語句的平均距離。

表 4 所列出的數值，就是這三個系統的平均頻譜距離，從表 4 可知 SYC 系統所產生的音框 DCC 向量，最靠近於錄音語句分析出來的 DCC 向量，而 SYP 系統所產生的音

框 DCC 向量，最遠離錄音語句分析出來的 DCC 向量。此外，只要一個音節中的聲母和韻母之間的文脈相依性有被掌握(modeled)，即 SYD 系統的情況，則量測出的 DCC 頻譜距離，就會比 SYP 系統的好很多，這表示圖 4 和圖 5 所列出的半段式 HMM 結構，的確可幫忙掌握兩個相鄰語音單元之間的文脈相依之頻譜特性。

*表 4. 平均 DCC 頻譜距離*

| System | SYC | SYD | SYP |
|---|---|---|---|
| Avg. dist. | **0.633** | **0.640** | **0.732** |

## 4.4 主觀聽測實驗

聽測實驗使用一篇訓練語句沒有用到的短文，該短文包括 70 個音節，我們將它分別輸入到三個系統去合成出語音音檔，在此分別以 WC、WD 和 WP 來表示 SYC、SYD 和 SYP 這三個系統所合成的音檔，這些音檔可到如下的網址去下載與試聽: http://guhy.csie.ntust.edu.tw/hmmhalf/。

透過 WC、WD 和 WP 三個音檔，我們進行流暢度比較的聽測實驗，一共邀請了 12 位受測者，在第一次聽測實驗裡，受測者以隨機次序聆聽 WC 和 WD 音檔；在第二次聽測實驗裡，受測者以隨機次序聆聽 WC 和 WP 音檔；在第三次聽測實驗裡，受測者則以隨機次序聆聽 WD 和 WP 音檔。在各次聽測實驗中，當一位受測者聽完兩個音檔後，我們要求他給一個分數來顯示流暢度的聽測結果，評分的範圍為-2 到 2 分，2(-2)分表示後者(前者)比前者(後者)流暢很多，1(-1) 分表示後者(前者)比前者(後者)稍微流暢，0 分表示分辨不出來。

三次聽測實驗之後，我們依音檔播放次序來調整分數的正負號，然後計算出各次實驗的平均分數和標準差，結果得到如表 5 所示的數值。從表 5 可知第一次和第二次實驗的平均分數為-0.833 和-0.417，負的分數表示音檔 WC 比 WD 和 WP 較為流暢，如果全部受測者都具有語音合成研究的背景，我們認為兩負分的絕對值應會更大，所以半段式 HMM 結構的確可以有效的提升合成語音的流暢度。至於第三次實驗的平均分數 0.250，該分數的絕對值最小，我們覺得這表示 WD 和 WP 音檔的流暢度應無明顯的差異。

*表 5. 聽測實驗之平均評分*

| | WC vs. WD | WC vs. WP | WD vs. WP |
|---|---|---|---|
| AVG | **-0.833** | **-0.417** | **0.250** |
| STD | **0.718** | **0.900** | **0.866** |

## 5. 結語

在本論文中，我們應用音素的發音知識於取代決策樹，來對一個語音單元(聲母或韻

母)的左右文脈作分類，以降低文脈組合之數量；此外，更進一步研究提出文脈相依之半段式 HMM 結構，以便在有限語料的情況下，掌握一個語音單元的文脈相依頻譜特性。如此結合兩者，用以改進合成語音的流暢度。

為了評估本論文所提出的半段式 HMM 結構，我們進行了兩種實驗，即頻譜距離量測和流暢度聽測，量測出的平均頻譜距離顯示，使用半段式 HMM 結構所合成出的語句，和原始錄音語句之間的頻譜距離，可從 0.732 減少到 0.633；此外，聽測實驗的結果顯示，使用半段式 HMM 所合成出的語音，比使用另外兩種 HMM 結構的較為流暢，所以在訓練語句不充足的情況下，半段式 HMM 結構確實可改進合成語音的流暢度。

未來我們可在相同訓練語料的情況下，比較我們系統的合成語音與 HTS 軟體的合成語音，觀察兩者在客觀頻譜距離和主觀聽測上的差異。另外，本論文著重於改進語音單元之間頻譜銜接上的流暢度，未來可再考慮去改進韻律方面的流暢度，以更為提升系統整體的流暢度。

## 致謝

感謝國科會計畫之經費支援，國科會計畫編號: NSC 102-2221-E-011-129。

## 參考文獻

Cappé, O., & E. Moulines (1996). Regularization techniques for discrete cepstrum estimation. *IEEE Signal Processing Letters*, 3(4), 100-102.

Gu, H. Y., & S. F. Tsai (2009). A discrete-cepstrum based spectrum envelope estimation scheme and its example application of voice transformation. *International Journal of Computational Linguistics and Chinese Language Processing*, 14(4), 363-382.

Gu, H. Y., & C. Y. Wu (2009). Model spectrum-progression with DTW and ANN for speech synthesis. In *Proc. ECTI-CON*, Pattaya, Thailand, 1010-1013.

Gu, H. Y., M. Y. Lai, & S. F. Tsai (2010). Combining HMM spectrum models and ANN prosody models for speech synthesis of syllable prominent languages. In *Proc. ISCSLP*, Tainan, Taiwan, Special Session 1.

Gu, H. Y., & C. L. Tsai (2013). Integrating speaker-nonspecific timbre transformation to an HNM based speech synthesis scheme. *Journal of the Chinese Institute of Engineers*, 36(3), 371-381.

Hsia, C. C., C. H. Wu, & J. Y. Wu (2010). Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in HMM-based speech synthesis. *IEEE trans. Audio, Speech, and Language Processing*, 18(8), 1994-2003.

Rabiner, L. & B. H. Juang (1993). *Fundamentals of Speech Recognition*, Prentice Hall.

Toda, T., & K. Tokuda (2005). Speech parameter generation algorithm considering global variance for HMM-based speech synthesis. In *Proc. Eurospeech*, Lisbon, Portugal, 2801-2804.

Tokuda, K., H. Zen, & A. W. Black (2004). 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, 135-153.

Yan, Z. J., Y. Qian, & F. K. Soong (2009). Rich context modeling for high quality HMM-based TTS. In *Proc. INTERSPEECH*, Brighton, UK, 1755-1758.

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi, & T. Kitamura (1999). Simultaneous modeling of spectrum, pitch and duration in HMM based speech synthesis. In *Proc. Eurospeech*, Budapest, Hungary, 2347-2350.

Zen, H., T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, & K. Tokuda (2007). The HMM-based speech synthesis system (HTS) version 2.0. In *Proc. 6th ISCA Workshop on Speech Synthesis*, Bonn, Germany, 294-299.