國 立 交 通 大 學

電信工程研究所

碩士論文

以階層式韻律模型為基礎之中文半隱藏式馬可夫模型語音合成器

A HSMM-based Mandarin Speech Synthesizer Based on Hierarchical Prosody Model

研 究 生：吳文良

指導教授：陳信宏 博士

中華民國一百年八月

以階層式韻律模型為基礎之中文半隱藏式馬可夫模型語音合成器

A HSMM-based Mandarin Speech Synthesizer Based on Hierarchical Prosody Model

研 究 生：吳文良　　Student：Wen-Liang Wu

指導教授：陳信宏 博士　　Advisor：Dr. Sin-Horng Chen

A Thesis
Submitted to Institute of Communication Engineering
College of Electrical and Computer Engineering
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master
In
Communication Engineering

August 2011

Hsinchu, Taiwan, Republic of China

中華民國一百年八月

以階層式韻律模型為基礎之中文半隱藏式馬可夫模型語音合成器

研 究 生：吳文良　　指導教授：陳信宏 博士

國立交通大學電信工程研究所碩士班

中文摘要

本論文目標為引入階層式韻律模型，進一步提升以馬可夫模型為基礎之合成器表現。首先引入韻律模型相關之韻律標記-音節邊界停頓標記與音節韻律狀態，將其運用到頻譜模型訓練過程，在決策樹分群階段改以韻律標記取代傳統語言資訊，改以介於上層語法資訊與下層音節資訊間的中層韻律資訊供決策樹分群使用，韻律標記除考量語言資訊外，更同時考量了聲學上的資訊，故應比語言資訊與頻譜更加相關，經實驗證實，韻律標記確實可提供勝過語言資訊的分群能力，訓練出更好的頻譜模型。接著進一步考慮合成時韻律模型的運用，因合成階段僅有文字，但欲取得標記需同時具有聲學與語言資訊，故本論文提出以條件式隨機域的方式訓練以文字預估韻律標記的模型，由於其可同時考量全域觀察序列之影響，並且利用前後狀態相關性進行模型學習，對於具時間相關性的參數預估應極有幫助，從實驗結果可發現，預估得到的韻律狀態，大多皆能符合音節邊界停頓對應的轉移特性。最後結合頻譜模型、韻律模型與預估得到之韻律標記，即為一完整合成系統，此系統具韻律變化豐富之優點，但因音節邊界停頓預估仍不夠好，導致部分合成語音的自然度欠佳，此有待未來繼續努力。

A HSMM-based Mandarin Speech Synthesizer Based on Hierarchical Prosody Model

Student：Wen-Liang Wu Advisor：Dr. Sin-Horng Chen

Institute of Communication Engineering
National Chiao Tung University

Abstract

In this thesis, we introduce the hierarchical prosody model to further improve the HMM-based synthesis system performance. First, we apply two types of prosodic tags, prosodic breaks and prosody states, to the spectral model training process. In the process of decision tree clustering, we replace the high-level linguistic features with the middle-level prosodic tags to cluster context dependent model. For the prosodic tags labeling, we consider not only linguistic features but also acoustic features. We suggest it be more related to spectrum than considering linguistic features only. The experiment confirms that our proposed method is better than the conventional method considering linguistic features only in the clustering process.

Second, in the synthesis stage, there is no way to label the prosodic tags of the text with the prosody model owing to the lack of acoustic features. As a result, we propose the conditional random fields(CRFs) method to estimate two types of prosodic tags according to the input text information. Because during the CRF model training process, it considers all the observation sequences and the neighboring output states, it is contributive to estimate the time-dependent parameter. The results of experiment show the transition of prosody states matches the corresponding prosodic breaks.

Last, we build our proposed complete synthesis system by combining the training spectral model, the prosody model and the estimating prosodic tags, which has the advantage of prosodic diversity. Nevertheless, it is still not good enough for the prosodic break prediction. The prediction results degrade the naturalness of synthesis speech, thus improving the prosodic break prediction will be the future work.

誌謝

完成這篇論文首先要感謝陳信宏老師從專題以來多年的指導，因為老師一直以來對我研究的關心與指點，今天才能順利完成這篇論文；當然也要感謝王逸如老師從碩一來的指導，有您不斷的提醒並點出我們研究上的盲點，我才能從大學生變成研究生，然後以個研究生的身分從交大畢業。

除了感謝兩位老師之外，當然也要感謝實驗室各位學長姐、同學與學弟妹，首先感謝我研究的啟蒙學長阿德，很遺憾最後沒有跟你一起做出很屌的方法幹掉 Toda，這個目標就靠你一個人完成了；接著要感謝性獸，在後來這一年對我研究的大力幫助，沒有你就沒有今日的我啊；也感謝合哥幫我解決了很多 HTK 與 Linux 的相關問題，另外還有輝哥、希群學長這兩年的幫忙；感謝完博班學長後，接下來要感謝的是常找我去吃宵夜的小宋和普屋，雖然你們都畢業兩年了，還是偶而會關心我給我研究上的建議；然後是去年畢業的各位學長姐，最關心我的承燁學長，還好有你不斷的提醒我寫論文，不然我到現在論文說不定還沒寫完，PUMA 在口試前一天還為了幫我看投影片不跟妹聊天，也感謝一哥宥余、未來的前輩小卡、快變明星的 jolin 與 NOVA 一姐舒舒這一年來對我的鼓勵；再來是一起奮鬥的各位夥伴們，第一梯先走一步的銘傑與勁竹，激勵剩下的人該好好認真不然就畢不了業；然後感謝一起口試的其他三人，感謝冠驛幫大家跑公文，從公民課本走出來的大胖讓我打球不會被電，玩遊戲時要特別小心的豆腐讓我打電動不會被電；最後是還在奮鬥的小蝦跟帥氣的智障，接下來就看你們表現了!!!實驗室的學弟妹，一起打球的三人組 DD、權哥、囂張的 KIWI(改天有空再回來電你們啊)，還有小邱、企鵝、雅婷跟昂星你們七個，實驗室的未來就託付給你們了。

最後的最後當然還要感謝生下我的父母，與各位大學同學們，還有一路上任何幫助過我的你、你、你、你、你...，然後感謝上天的保佑，我終於畢業了!!!!!!!!!!

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究動機

二十一世紀是個科技爆炸的時代，隨著科技發展的日新月異，語音的實際應用也日益興盛，例如氣象查詢系統 (Jupiter) 及航空訂票系統 (ATIS) 等，語音向來為人與人間最自然、直接的溝通方式，因此使用語音取代其他輸入、輸出介面，做為人類與機器間的溝通橋樑，便成為科技研究的重要目標；與電子資訊相關產品結合更是語音研究的未來趨勢，伴隨著平版電腦、手機、GPS 及網路等的普及，我們期望以語音技術，為科技發展寫下嶄新的一頁。

近年來，語音合成系統的發展已經相當成熟，市面上無論是玩具或者高科技產品，都結合了語音合成的功能，但這些產品使用的合成技術大多有其限制存在，例如合成聲音不夠自然，或只能在特定情境下運用，至今尚未有人發展出可適用於任何情境，並具高品質聲音的合成器。因此本論文希望透過分析語者之韻律特性，將其運用到語音合成裡，進一步提升文字轉語音系統的聲音自然度與聲音品質。

1.2 文獻回顧

文字轉語音合成(Text-to-Speech, TTS)技術的發展已經有好幾十年的歷史，採用大語料為基礎的方法(Corpus-based method)如單元選取合成法(Unit selection method)大幅提升了 TTS 的合成聲音品質，其挑選的單元來自錄音語料，因此輸出語音的聲音品質與自然度都相當不錯；但因欲合成語句不可能都包含在語料庫中，故需從不同的句子裡挑選單元，單元銜接處易有不連續或不自然的情形，若能挑選到越大的單元輸出語音品質就越好，此方法的一大缺點便是需要有龐大的語料庫才能得到較佳的聲音品質，且受限於錄音語料，故輸出聲音較固定缺少變化。

另一較常見的方法為統計式參數語音合成法，其中包含近期最廣為人使用的基於隱藏式

馬可夫模型之語音合成器(HMM-based speech synthesis system)，HMM 因其適宜學習連續變化時間訊號的特性，故廣泛被使用在模型化語音信號，HMM-based 合成器可直接利用參數產生演算法，由 HMM 模型合成聲音，相較於單元選取的方法，不需蒐集大量語料庫即可合成出品質不錯的語音，雖然品質仍不如最好的串接式方法，但統計式參數語音合成法擁有方便延伸應用這個最大優點，如利用調適方法【1-2】轉換合成聲音特性或利用內插等方法改變語速等，且其相對於串接式合成法，具不需龐大空間儲存語料的優點，在實際應用上有極大優勢。

綜合上述優點，HMM-based 合成器為目前公認相當不錯的一種合成方法，但其仍存在許多待改善的問題，如基頻(F0)的求取及有聲與無聲(Unvoiced/Voiced, U/V)判定的問題，Tokuda【3】提出以連續與離散共存的機率分佈表示方法(Multi-Space Distribution, MSD)，解決了 F0 同時包含 U/V 部分難以模型化的問題，且有不錯的合成結果，但一些不正確的 U/V 判定仍會破壞合成聲音品質；另一造成聲音品質下降的重要因素則為過度平滑(Over-smoothing)之現象，參數間過度平滑通常會使合成聲音有背景低鳴聲(Muffle)的現象；Toda【4】引入全域變異數(Global Variance, GV)的概念，大幅提升了合成聲音品質。然而合成聲音的清晰度依然不足以在日常生活中廣泛使用，故仍有許多為進一步提升 HMM-based 合成之聲音品質的研究，如 Wu【5】提出最小生成誤差(Minimum Generation Error , MGE)的方法，取代一般常用之最大概似函數(Maximum Likelihood , ML)的準則(criterion)訓練模型；Zen 【6】提出以半隱藏式馬可夫模型(Hidden Semi-Markov Model , HSMM)取代 HMM，解決音長模型訓練與合成不一致的問題等等。

1.3 研究方向

傳統 HMM-based 方法因為利用統計式參數還原的方法，會產生接近「平均」的聲音，雖然乍聽之下品質不差，但因合成的聲音接近統計上的平均值，導致聲音變異性較小、韻律固定，聲音清晰度稍嫌不足，且在聽感上並不夠自然。

本研究為進一步提升 HMM-based 語音合成器的聲音品質，同時對頻譜與韻律部分做改進，以交大電信工程研究所江振宇博士所提出之非監督式中文語音韻律標記及韻律模型(Prosody Labeling and Modeling,PLM)【7】為基礎，頻譜方面，希望藉由韻律資訊的運用，取代傳統利用上層語言資訊對頻譜參數做分群的方法，本研究所使用之韻律標記可視為一介於上層語法資訊與下層音節層次間的一中層資訊，且模型化韻律參數的過程同時考慮了聲學(Acoustic)與語言學(Linguistic)的相關資訊，故此參數應較單純上層語言資訊更符合真實頻譜分佈，且能提供下層音節資訊無法表示的韻律階層架構之影響，希望藉此資訊的引入提高模型分佈的集中度，降低統計方法聲音會過度平滑之影響；韻律方面，則利用 PLM 訓練得到之韻律模型，根據預估之韻律標記與音節語言資訊直接產生音節音高與音長序列，此方法一樣可避免掉統計方法韻律特性過度平滑，語調平淡、節奏固定的缺點，最後結合頻譜、韻律模型提出本研究之合成器架構，解決傳統 HMM-based 合成器聲音品質不佳的缺點。

1.4 語料庫簡介

本論文所採用的實驗語料庫，是由一位專業的女性播音員讀稿錄製而成之中文語料庫，總計 418 個音檔，共 55525 個音節，平均一個音檔有 133 個字。音檔均為 16-bit PCM 格式，取樣頻率為 16kHz，錄製文本為 Sinica Treebank Version 3.0 語料庫中選出的短篇文章，Sinica Treebank Version 3.0 語料庫的檔案總共有 6 個類型(表 1.1)，本語料庫所選用的文本皆來自其中的新聞語料(news.check)；文本解析的內容是由自動標記後再經人工修正得到，音調與音節類型是經由 130000 萬詞的字典標記而成，所有音節的切割位置和基頻軌跡(F0)的偵測則分別由 Hidden Markov Model Tookit(HTK)【8】和 WaveSurfer【9】求取，再經過人工修正而成。而為配合實驗需要，本研究進一步將語料庫分成 375 句的訓練語料共 51708 個音節，與測試語料 43 句共 3817 個音節。

表 1.1 Sinica Treebank 語料庫內容

File name	Content
news.check, travel.check	News papers, books, or internet articles
ko.check, ev.check	Elementary school text books
oral.check	Text from phonetic balanced speech corpus
sino.check	Text from Taiwan Panorama

1.5 章節概要說明

本論文的內容共分為六章：

第一章：緒論，介紹本論文之研究動機、研究方向、及語料庫說明。

第二章：HSMM-based 中文語音合成器，簡介傳統 HTS 系統與本研究提出之系統架構。

第三章：以韻律模型為基礎之 HSMM，介紹本研究頻譜訓練方法，並分析模型訓練結果。

第四章：韻律產生器：介紹本研究所提出之韻律預估與產生方法。

第五章：合成系統實作與評估：整體合成系統比較與結果分析。

第六章：結論與未來展望。

第二章 HSMM-based 語音合成器

本章描述本論文之所使用之 HSMM-based 語音合成系統。2.1 節先介紹基於隱藏式馬可夫模型的語音合成系統(HMM-based Speech Synthesis System, HTS)，2.2 節介紹基於半隱藏式馬可夫模型的語音合成系統(HSMM-based Speech Synthesis System)，2.3 節介紹本論文所提出之合成系統架構。

2.1 HMM-based 語音合成系統

過去隱藏式馬可夫模型被大量應用在語音辨識系統中，利用機率模型來描述口腔的各種發音現象。近年來，此技術開始被廣泛應用到語音合成上，在目前眾多語音合成系統中，因合成品質可達到不錯水平被大眾廣為使用，並在 2005 年的語音合成比賽 Blizzard challenge【10】中，獲得相當好的成績，且因統計式參數語音合成法易於進一步延伸應用的特性，一躍成為當今語音研究中最受矚目的合成方法。

本研究使用的 HTS 為日本名古屋大學資工研究所開發出來的 HTS 2.1【11】，此系統為基於 HTK 技術開發出的一套合成系統，專為使用隱藏式馬可夫模型語音合成系統建構。HTS 語音合成系統架構圖如圖 2.1 所示：

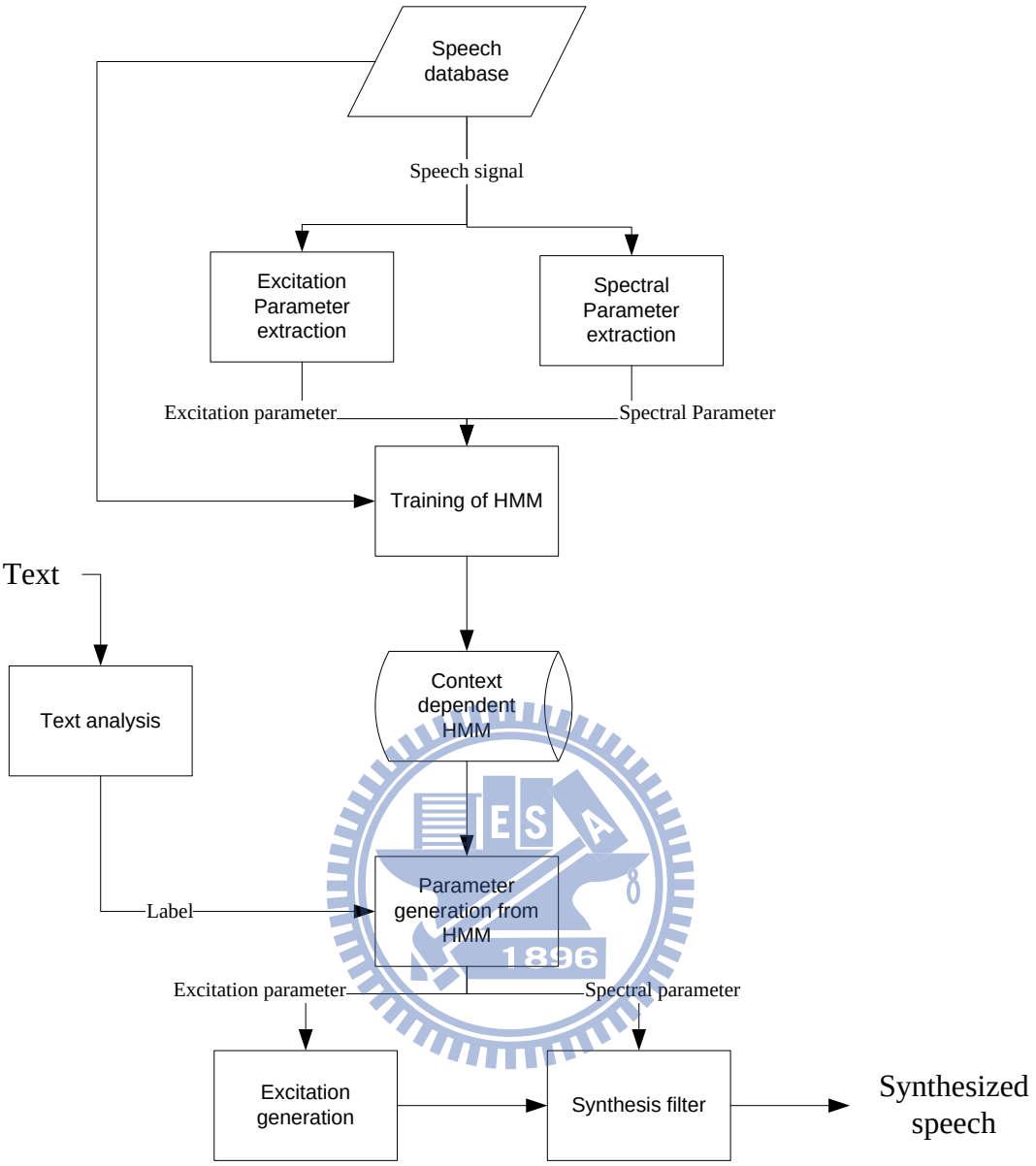

圖 2.1 HMM-based 語音合成系統架構圖【12】

如圖 2.1 所示，HTS 系統可分為訓練部分與合成部分，訓練部分，首先由語料庫抽取廣義梅爾倒頻譜參數(Mel-generalized Cepstrum, MGC)與激發訊號參數(log F0)，另一方面，根據文字分析器的文字分析結果產生對應的文脈相關之文本標示，搭配適當的文脈相關問題集，訓練狀態(state)合併分裂樹，如圖 2.2 所示，首先收集所有屬於同一狀態的資料，由根節點根據文脈相關問題集進行分裂，頻譜與音高模型的決策樹為獨立分開訓練，如圖 2.2 的 HMM 有三個狀態，便會針對每個狀態的頻譜與音高參數個別產生一顆決策樹，共六棵樹，而音長

模型則定義僅有一個狀態，參數維度同於頻譜與音長模型的狀態數，這邊即為三維，因為只有一個狀態所以只訓練一顆決策樹，HTS 即透過決策樹之分群產生文脈相關 HMM，包含音高模型、頻譜模型及音長模型。

合成部分，輸入文字後，透過文字分析器產生文脈相關的文本標示，利用分類與回歸樹(CART)演算法，挑選對應的 HMM 模型序列，並藉由參數產生演算法，產生頻譜參數與激發訊號參數，最後以 MLSA 濾波器(Mel Log Spectrum Approximation filter))產生語音信號。

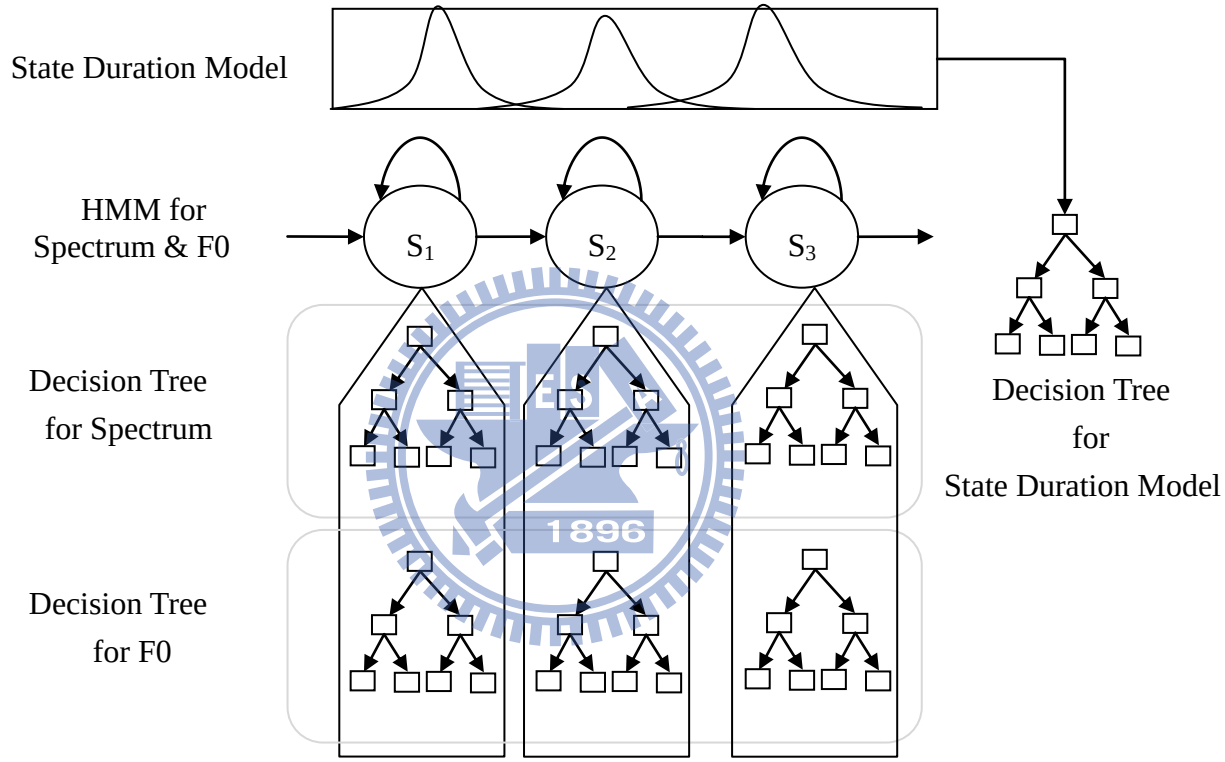

圖 2.2 HTS 系統之決策樹【12】

2.2 HSMM-based 語音合成系統

HSMM-based 的語音合成系統，基本架構與上一節中介紹的 HMM-based 語音合成系統相同，差別僅在模型訓練階段，HMM-based 使用的 HMM 如圖 2.3 所示，而 HSMM-based 使用的 HMM 如圖 2.4 所示，比較兩圖可發現，圖 2.3 中標準 HMM 的狀態轉移機率，在圖 2.4 中被一機率分佈模型給取代，簡言之，HSMM 即以狀態持續時間模型取代 HMM 的轉移

機率，可視為加入狀態持續時間模型的 HMM。傳統 HMM-based 合成器的狀態持續時間模型由訓練階段模型最後一次估算的網狀結構(Trellis)中計算產生，然而在合成階段，預估頻譜參數序列卻得同時考慮狀態持續時間模型與頻譜模型，但狀態持續時間模型在訓練階段並未與頻譜模型一同訓練，導致訓練階段與合成階段所考慮之模型不一致，估出之最佳序列可能會不夠好，以致影響最後合成的品質；Zen【6】提出以 HSMM 取代 HMM 的方法，如此即可在訓練階段一併訓練狀態持續時間模型，解決訓練與合成時，考慮的模型不一致的問題。

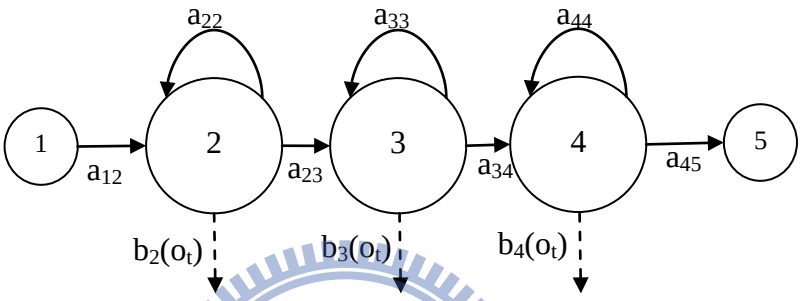

圖 2.3 三個狀態由左至右不允許狀態跳躍之 HMM

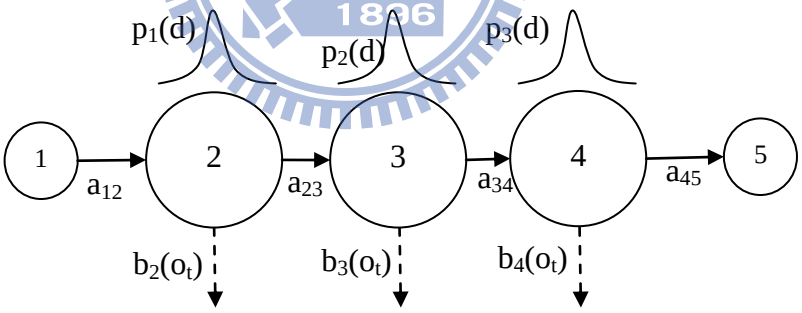

圖 2.4 三個狀態由左至右不允許狀態跳躍之 HSMM

本論文即以 HSMM-based 語音合成系統為基礎做延伸應用，因此比較的對象為使用 HSMM 架構的傳統 HTS 系統，詳細的模型訓練流程將在第三章中介紹，而因合成系統是利用決策樹對模型做分群，故重點在於決策樹對應的文本標記參數與問題集的選定，因此接下來即對傳統 HTS 使用的問題集與文本標記做介紹。

文本標記資訊為 HTS 相當重要的一環，採用哪些語言參數會直接影響到文脈相關模型

的狀態分裂合併結果。本論文所採用的語言參數，可粗分為五大類：音節層次(Syllable level)、詞層次(Word level)、片語層次(Phrase level)、句子層次(Sentence level)，詳細使用之文脈相關語言參數，如表 2.1 所示：

表 2.1 文脈相關資訊

level	ID	Description
Syllable level	Pr_Phn	Previous phone(Initial/Final)
	-Cur_Phn	Current phone(Initial/Final)
	+Fol_Phn	Following phone(Initial/Final)
	^Phn_in_Syl	Phone position in a syllable
	=Pr_Tone	Lexical tone of previous syllable
	@Cur_Tone	Lexical tone of current syllable
	#Fol_Tone	Lexical tone of following syllable
	&F_Syl_in_SubWrd	Syllable position in a sub-lexical word (SLW) (forward)
	|B_Syl_in_SubWrd	Syllable position in a SLW (backward)
	/p:F_Syl_in_Wrd	Syllable position in a lexical word (LW) (forward)
	/q:B_Syl_in_Wrd	Syllable position in a LW (backward)
Word level	/a:Pre3POS_SWL	47-type POS/word length (WL) of previous-previous-previous SLW
	/b:Pre2POS_SWL	47-type POS/WL of previous-previous SLW
	/c:Pre1POS_SWL	47-type POS/WL of previous SLW
	/d:CurPOS_SWL	47-type POS/WL of current SLW
	/e:Fol1POS_SWL	47-type POS/WL of following SLW
	/f:Fol2POS_SWL	47-type POS/WL of following-following SLW
	/g:Fol3POS_SWL	47-type POS/WL of following-following-following SLW
	/A:Pre3POS_WL	47-type POS/word length (WL) of previous-previous-previous LW
	/B:Pre2POS_WL	47-type POS/WL of previous-previous LW
	/C:Pre1POS_WL	47-type POS/WL of previous LW
	/D:CurPOS_WL	47-type POS/WL of current LW
	/E:Fol1POS_WL	47-type POS/WL of following LW
	/F:Fol2POS_WL	47-type POS/WL of following-following LW
	/G:Fol3POS_WL	47-type POS/WL of following-following- following LW
Phrase level	/H:F_Syl_in_Ph	Syllable position in a syntactic phrase (forward)
	/I:B_Syl_in_Ph	Syllable position in a syntactic phrase (backward)
	/J:CurPhType_PhL	Syntactic phrase type/length of current phrase
	/K:FolPhType_PhL	Syntactic phrase type/length of the following phrase
Sentence level	/L:Pr_PM	PM type preceding current syllable
	/M:Fol_PM	PM type following current syllable
	/N:F_Syl_in_Snt	Syllable position in a sentence (forward)
	/O:B_Syl_in_Snt	Syllable position in a sentence (backward)
	/P:CurSntL	Current sentence length in syllable

建立好文脈標示後，接著根據表 2.1 之參數設計相關問題集，可分為下列五大類問題集：

1. 音節層次(Syllable level)：
   - i. 考慮當前音素與前、後音素:
     - ➢ 聲母發音類別：爆破音、摩擦音、鼻音、邊音、塞擦音等等。
     - ➢ 韻母發音類別：單元音韻母、複合元音韻母、鼻尾音韻母等等。
   - ii. 考慮當前音節聲調與前、後音節聲調。
   - iii. 考慮音節在詞中位置：由前面數來第幾個字，由後面數來第幾個字，詞中不同位置都可能影響最後聲音的韻律特性，此處將詞彙詞(Lexical word)與次詞彙詞(Sub-lexical word)分開考慮。
2. 詞層次(Word level)：
   - i. 考慮當前詞(±0)與前後三個詞(±1、±2、±3)的詞類，依中研院 46 類詞類依實詞、虛詞、八大詞類及特殊詞類集合合併，產生問題集。
   - ii. 考慮當前詞(±0)與前後三個詞(±1、±2、±3)的詞長。

   以上同樣分詞彙詞與次詞彙詞兩類考慮。
3. 片語層次(Phrase level)：
   - i. 考慮當前音節在片語中位置：由前面數來第幾個字，由後面數來第幾個字。
   - ii. 考慮當前片語與後一個片語的類別。
   - iii. 考慮當前片語與後一個片語的長度。
4. 句子層次(Sentence level)：
   - i. 考慮當前音節的前、後音節邊界是否存在標點符號
   - ii. 考慮當前音節位在句子中第幾個字：由前面數來，由後面數來。
   - iii. 考慮當前句子與後一個句子的長度。

訂定好問題集與文本標記後，只需對訓練與欲合成的文本做好文脈相關文本標記，即可利用 HTS 系統，如圖 2.1 之系統架構圖訓練模型並合成聲音。

2.3 結合韻律模型之 HSMM-based 語音合成系統

本研究的模型訓練階段，與傳統 HTS 系統使用相似的訓練流程，但因合成階段架構不同的緣故，將原本合併訓練的頻譜與音高參數(mgc+lf0)修改成單純訓練頻譜參數，並根據客觀實驗結果，對模型結構做調整，取代傳統的狀態模型結構(State-based)根節點結合相同狀態的所有音素模型，進行決策樹分裂(Tree splitting)，訓練本研究之文脈相關模型(Context-dependent model)。另一方面，以江振宇博士所提出之非監督式中文語音韻律標記及韻律模型演算法訓練韻律模型，並使用此韻律模型所定義之兩類韻律標記幫助頻譜模型訓練。

合成階段可參考下方之語音合成系統架構圖，當一段文字進來，首先透過文字分析器產生所需的各項語言資訊，並利用此資訊預測兩類韻律標記，預測完成後即可透過韻律產生器，利用聲調、音節類型與韻律標記，直接從訓練的韻律模型產生音節音高與音節音長序列；另一方面，當具備語言資訊與韻律標記，即可從文脈相關 HSMM 中挑選合成單元串接成頻譜與音長 HMM 序列；另一方面，韻律產生器產生之音節音長，配合音長 HMM 序列，預估每個音素的狀態持續時間(State duration)，有了頻譜 HMM 序列與狀態持續時間即可運用參數產生演算法生成頻譜參數序列，音節音高根據有聲部分長度還原音節基頻軌跡(Syllable pitch contour, F0)，經簡單轉換後可得激發訊號(Excitation)，結合頻譜參數與激發訊號便可使用 MLSA 合成器合成語音。

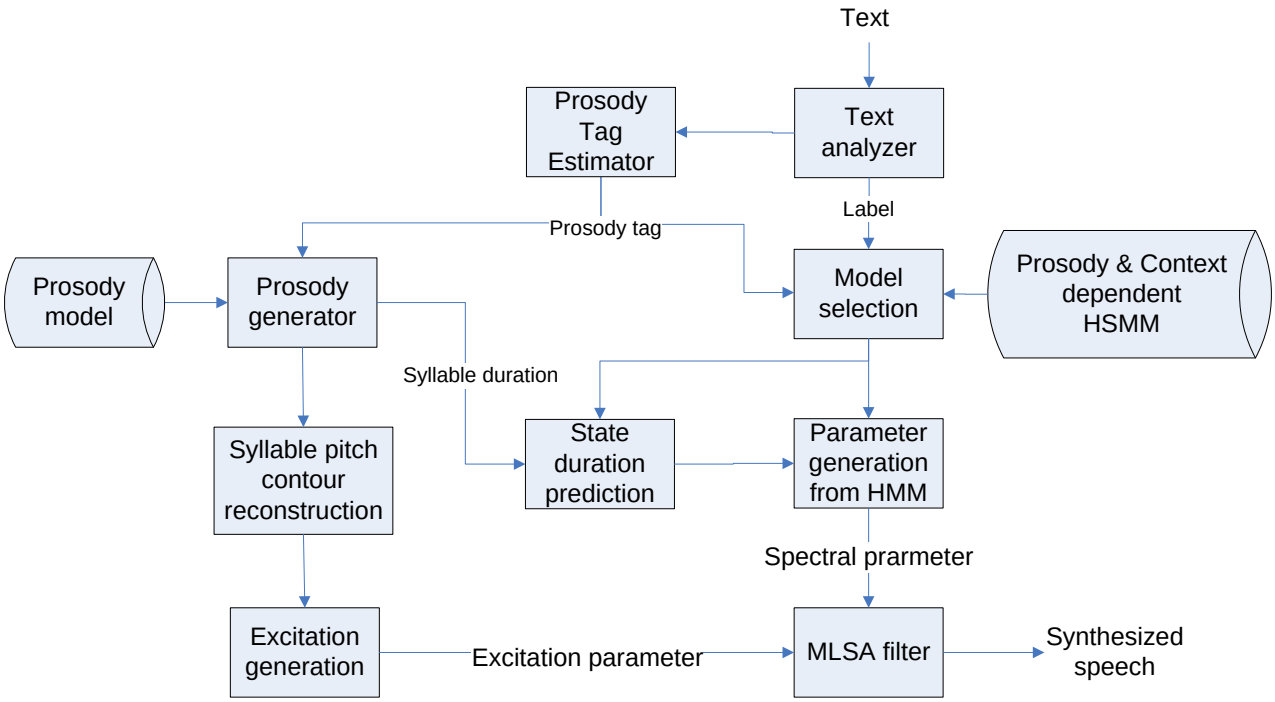

圖 2.5 語音合成系統架構圖

第三章 以韻律模型為基礎之 HSMM

本論文以江振宇博士所提出之中文韻律模型【7】為基礎，引入其所定義之兩類韻律標記，幫助訓練 HSMM。3.1 節將介紹中文語音之階層式韻律架構；3.2 節簡介中文韻律模型；3.3 節介紹本論文的 HSMM 訓練過程；3.4 節則對引入韻律標記所建立之模型做簡單分析。

3.1 中文語音階層式韻律架構

據韻律相關研究結果【13】，中文語音的韻律結構由階層式韻律架構(Hierarchical structure)組成，傳統定義韻律架構由底層至上層分別為音節(Syllable, SYL)、韻律詞(Prosodic Word, PW)、韻律短語(Prosodic Phrase, PPh)及語調短語(Intonation Phrase, IP)。因中文一個音節一個字的特性，故最底層的韻律單元為音節，而相同音節不同聲調語義多不相同，且聲調強烈影響音節基頻軌跡走向(音高)，也影響音節長度與音節能量，可視為音節層次最重要的韻律影響因素；韻律詞則是由雙音節或多音節構成的詞組，通常在句法或語意上緊密相關，因此易將其視為一個發音單元；韻律短語則是由一或多個韻律詞所組成，結尾通常有可察覺但不明顯的停頓；語調短語則是中文韻律架構的最上層，結尾會有明顯的停頓，由數個韻律短語組成的句子，音高變化亦受此層影響。基本上，一個句子中每個音節的音高和音長變化，皆可由此四層韻律架構詮釋。

鄭秋豫博士【14】提出韻律標記的概念並定義了一個韻律架構，如圖 3.1 所示。其將中文韻律結構分成五層，前三層(由下至上)和前述韻律架構相同，同樣為音節、韻律詞以及韻律短語。第四層則是將連續的韻律短語組合成呼吸群(Breath Group, BG)，代表一個有音高及音長變化的篇章或段落，藉此表示上層對韻律的貢獻，同時定義了第五層，由連續 BG 組成的韻律群(Prosody Group, PG)。此處的五層韻律架構共定義六種標記區分，其中 B0 和 B1 代表 SYL 的邊界，B0 表示 reduced syllable boundary，B1 則是 normal syllable boundary，通常 B0 及 B1 的位置聽不出停頓；B2 及 B3 分別代表 PW 和 PPh 的邊界；B4 和 B5 則是區分 BG 和 PG 的邊界，B4 代表呼吸停頓，而 B5 為一完整語音段落的結束，並有句尾音節長度拉長

(final lengthening)以及能量減弱的現象。

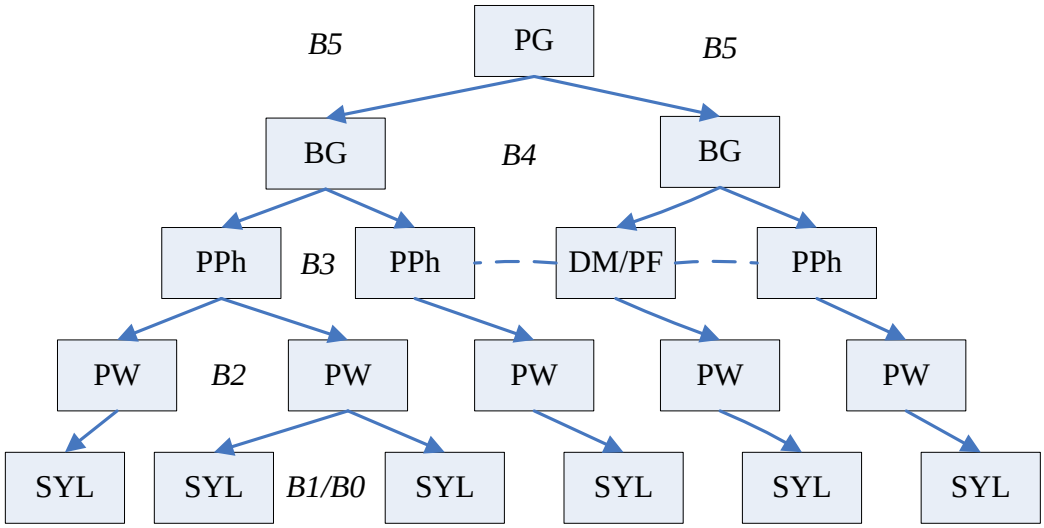

圖 3.1 階層式多短語韻律句群(Hierarchical Prosodic Phrase Grouping)架構。【14】

3.2 中文韻律模型

本節將介紹本研究使用之韻律模型定義的韻律架構，模型使用參數，與訓練方法。

3.2.1 韻律架構

本研究使用的韻律模型以鄭秋豫博士提出的中文韻律架構為基礎，將其中的韻律詞邊界 B2 進一步細分為 B2-1、B2-2 及 B2-3，分別代表明顯音高重置(Pitch reset)、短停頓(Short pause)及有音節延長效應(Duration lengthening)的韻律詞邊界，這三類雖然同屬於韻律詞邊界，但因其對應的聲學特性不同，只用一類描述並不適當，故我們將其分成三類；並將最上層的 B4、B5 合併成 B4 一層，因為這兩類所對應的聲學特性相近，故可合併用一類韻律邊界停頓表示即可。整個架構由原本定義的 5 層變回 4 層，如圖 3.2 所示。綜合上述，此韻律模型採用了 7 種停頓標記(Break Type) $\mathbf{B}$ ={B0, B1, B2-1, B2-2, B2-3, B3, B4}，來標記音節、韻律詞、韻律短語及呼吸群/韻律群四種韻律單元。

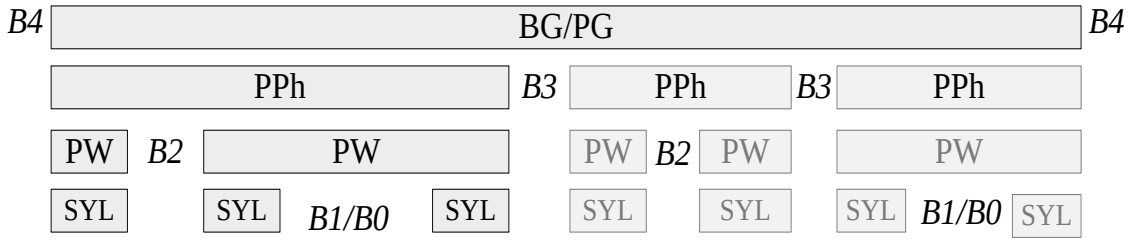

圖 3.2 本研究所用之階層式韻律架構

為了更詳盡描述這四層的階層式韻律架構，除了描述韻律邊界的停頓標記外，還需有描述韻律單元變化的其他韻律標記或參數。本研究利用帶有上層韻律資訊的標記來間接表示這些上層資訊對韻律架構的影響，此標記即為韻律狀態(Prosody state)，代表圖 3.2 架構中上面三層對韻律變化的貢獻。在本研究共使用了三種不同的韻律狀態，分別為正規化後之音高、音長和音節能量的量化值。正規化後的音高為扣除音節層次對音高的影響後的殘餘值，經量化後的殘餘值我們定義為音高的韻律狀態，代表韻律詞、韻律短語、呼吸群/韻律群這些上層架構對音高的貢獻；音長或音節能量同理可在扣除音節層次影響因素後，對其量化得到各自的韻律狀態。經由定義韻律狀態，可將音高、音長和音節能量在音節層次和高層次的影響分開，將複雜的高層次影響因素直接以韻律狀態表示之。

3.2.2 模型設計

本研究所採用韻律模型為江振宇博士所提出之中文韻律模型(PLM)，其依據圖 3.2 表示之中文階層式韻律架構，可針對一未經人工事先標記好的語料，利用語言參數和聲學參數，自動標記出停頓標記及韻律狀態。此演算法具備兩大優點：1.可自動標記，解決傳統上韻律標記多為人工標記，既耗時又耗力，且易有不一致的問題。2.透過此模型可清楚分析韻律詞層次以上的韻律變化趨勢。

韻律標記問題可視為，在給定語料庫之語音聲學參數集合 **A**，和對應的語言參數集合 **L** 下，求取最佳韻律標記集合 **T** 的過程，即

$$\mathbf{T}^{*} = \arg\max_{\mathbf{T}} P(\mathbf{T}|\mathbf{A},\mathbf{L}) = \arg\max_{\mathbf{T}} P(\mathbf{T},\mathbf{A}|\mathbf{L}) \tag{3-1}$$

韻律標記集合**T**={**B**, **PS**}包含了兩類重要的語音韻律資訊，第一類是音節邊界停頓標記(Break type)，本論文使用音節邊界停頓標記集合**B** = {B0, B1, B2-1, B2-2, B2-3, B3, B4}；另一類的韻律標記是音節的韻律狀態，在本方法中韻律狀態有 3 種**PS**={**p**, **q**, **r**}，代表的意義分別是經過量化和正規化音節基頻韻律狀態**p**、音長韻律狀態**q**和音節能量韻律狀態**r**。正規化後的基頻韻律狀態扣除掉音節層次對基頻的貢獻，即扣除聲調和連音的影響因素，此時音節基頻的韻律狀態代表的是韻律詞、韻律短語、呼吸組/韻律句組對基頻的貢獻；至於音長或能量強度則分別扣除語句、聲調、基本音節類型或韻母類型的影響因素。

聲學參數也可分為兩類，其中一類的聲學參數和韻律狀態標記有很大的相關性，而與音節邊界停頓標記的相關性則非常小，屬於這類的聲學參數有音節基頻軌跡、音長和音節能量；另一類的聲學參數則特性相反，和音節邊界停頓標記有很大的相關性，而與韻律狀態標記的相關性小，屬於這類的聲學參數有音節邊界的停頓時長(Pause duration)、音節邊界的能量低點(Energy-dip level)、正規化的能量差、正規化的基頻差(Normalized pitch jump)以及正規化的音節長度拉長因子(Normalized duration lengthening factor)等。因此我們定義 **A** 包含音節基頻軌跡序列 **sp**、停頓時長序列 **pd**、音節能量低點(Energy-dip level)序列 **ed**、音節長度序列 **sd**、音節能量序列**se**、正規化的音節內基頻差序列**pj**及正規化的音節長度拉長因子序列**dl**和**df**，其中 **pj** 定義為：

$$pj_n = \left(\mathbf{sp}_{n+1}(1) - \boldsymbol{\beta}_{t_{n+1}}(1)\right) - \left(\mathbf{sp}_n(1) - \boldsymbol{\beta}_{t_n}(1)\right), \tag{3-2}$$

上式括號中的 1 代表參數的第一維，下標 n 表示此為第 n 個音節，$\boldsymbol{\beta}_{t_n}$ 為聲調影響因素 t_n 的 Affecting patterns(APs)，而正規化的音節長度拉長因子序列 **dl** 和 **df** 定義為：

$$dl_n = \left(sd_n - \gamma_{t_n} - \gamma_{s_n}\right) - \left(sd_{n-1} - \gamma_{t_{n-1}} - \gamma_{s_{n-1}}\right) \tag{3-3}$$

和

$$df_n = \left(sd_n - \gamma_{t_n} - \gamma_{s_n}\right) - \left(sd_{n+1} - \gamma_{t_{n+1}} - \gamma_{s_{n+1}}\right) \tag{3-4}$$

其中 γ_t 和 γ_s 分別表示聲調與基本音節類型影響因素的 APs，因此聲學參數集合 **A**={**sp**,**sd**,**se**,**pd**,**ed**,**pj**,**dl**,**df**}。為了更清楚的說明這些聲學參數，將 **A** 進一步細分為三個類別：音節韻律參數(Syllable Prosodic Feature) **X**={**sp**,**sd**,**se**}，音節內韻律參數(Inter-syllabic Prosodic Feature) **Y**={**pd**,**ed**}以及音節差韻律參數(Differential Prosodic Feature) **Z**={**pj**,**dl**,**df**}。

至於語言參數方面，則用 **L** 來表示所有的語言參數集合。其中特別將音節聲調、基本音節類型與韻母類型從 **L** 獨立出來，因為這三個參數分別對音節基頻軌跡、音長與音節能量有顯著的影響；此外考慮到不同語句時，說話速度的變動會造成音長的變化以及音量變動會造成能量的變化，因此再把兩個語句層次的正規化因子獨立出來；扣除從 **L** 中獨立出來的語言參數後，剩餘的語言參數，則統一定義為 **l** (Reduced linguistic feature set)。為了清楚的表示這些符號定義，將其整理在表 3.1。

表 3.1 韻律標記、韻律參數和語言參數的表示法

T : prosodic tag	**B** : break type	
	PS : prosodic state	**p** : pitch prosodic state **q** : duration prosodic state **r** : energy prosodic state
A : prosodic feature	**X** : syllable prosodic feature	**sp** : syllable pitch contour **sd** : syllable duration **se** : syllable energy level
	Y : inter-syllabic prosodic feature	**pd** : pause duration **ed** : energy-dip level
	Z : differential prosodic features	**pj** : normalized pitch jump **dl** : normalized duration lengthening factor 1 **df** :normalized duration lengthening factor 2

L : linguistic feature	**l**: reduced linguistic feature set
	t: syllable tone sequence
	s: base-syllable type sequence
	f: final type sequence
	u: utterance sequence

綜合上述之討論，可將 3-1 式改寫為

$$\begin{aligned} P(\mathbf{T},\mathbf{A}|\mathbf{L}) &= P(\mathbf{A}|\mathbf{T},\mathbf{L})P(\mathbf{T}|\mathbf{L}) = P(\mathbf{X},\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{PS},\mathbf{L})P(\mathbf{B},\mathbf{PS}|\mathbf{L}) \\ &\approx P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})P(\mathbf{PS}|\mathbf{B})P(\mathbf{B}|\mathbf{L}) \end{aligned} \tag{3-5}$$

其中 $P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})$ 稱為音節韻律模型(Syllable Prosodic Model)，$P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$ 稱為停頓聲學模型(Break-acoustic Model)，$P(\mathbf{PS}|\mathbf{B})$ 稱為韻律狀態模型(Prosodic State Model)，$P(\mathbf{B}|\mathbf{L})$ 稱為停頓標記語言模型(break-syntax model)。進一步將音節韻律模型 $P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})$ 分解成三個模型，分別模擬音節基頻軌跡序列 **sp**、音長序列 **sd** 和音節能量序列 **se**，並且假設 **sp**、**sd** 和 **se** 的變化在此只受到以下幾個影響因素控制：音節聲調 **t**、基本音節類型 **s**、韻母類型 **f** 、語句 **u**、韻律狀態 $\mathbf{PS}=\{\mathbf{p},\mathbf{q},\mathbf{r}\}$ 和韻律邊界停頓 **B**，因此得到

$$\begin{aligned} & p(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L}) \approx p(\mathbf{sp}|\mathbf{B},\mathbf{p},\mathbf{t})\,p(\mathbf{sd}|\mathbf{q},\mathbf{t},\mathbf{s},\mathbf{u})\,p(\mathbf{se}|\mathbf{r},\mathbf{t},\mathbf{f},\mathbf{u}) \\ & \approx \prod_{n=1}^{N} p(\mathbf{sp}_n|B_{n-1}^{n}, p_n, t_{n-1}^{n+1}) \prod_{n=1}^{N} p(sd_n|q_n,t_n,s_n,u_n) \prod_{n=1}^{N} p(se_n|r_n,t_n,f_n,u_n) \end{aligned} \tag{3-6}$$

其中 $\prod_{n=1}^{N} p(\mathbf{sp}_n|B_{n-1}^{n}, p_n, t_{n-1}^{n+1})$ 是在模擬音節基頻軌跡受的各類影響，式子代表的意思為第 n 個音節之基頻軌跡 $\mathbf{sp}_n$ 會受到目前基頻韻律狀態 p_n、目前聲調 t_n 以及給定韻律邊界停頓 B_{n-1} 和 B_n 情況下，前後相鄰音節聲調 t_{n-1} 和 t_{n+1} 造成的連音影響，此處 $B_{n-1}^{n}=(B_{n-1},B_n)$，$t_{n-1}^{n+1}=(t_{n-1},t_n,t_{n+1})$，而 $\mathbf{sp}_n$ 則代表第 n 個音節基頻軌跡，是將音節基頻軌跡進行正交展開(Orthogonal expansion)，投影到四個 Legendre 多項式基底所得到的四維正交參數，根據上面描述可將 $\mathbf{sp}_n$ 表示成

$$\mathbf{sp}_n = \mathbf{sp}_n^{r} + \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\beta}^{f}_{B_{n-1},tp_{n-1}} + \boldsymbol{\beta}^{b}_{B_n,tp_n} + \boldsymbol{\mu} \qquad \text{for } 1 \le n \le N \tag{3-7}$$

3-7 式的每項 $\boldsymbol{\beta}_x$ 表示音節基頻軌跡影響因素為 x 時的 AP，這裡的 tp_n 是 tone pair $t_n^{n+1}=(t_n,t_{n+1})$，$\boldsymbol{\beta}^f_{B_{n-1},tp_{n-1}}$ 和 $\boldsymbol{\beta}^b_{B_n,tp_n}$ 分別是第 n-1 個和第 n+1 個音節所貢獻的前後音節影響效應的 APs，$\boldsymbol{\mu}$ 則是全域平均值(Global mean)；而每個語句的韻律邊界都有開始與結束兩個特例，這兩個分別以 B_b 和 B_e 表示之，因此 $\boldsymbol{\beta}^f_{B_b,t_1}=\boldsymbol{\beta}^f_{B_0,tp_0}$，$\boldsymbol{\beta}^b_{B_e,t_N}=\boldsymbol{\beta}^b_{B_N,tp_N}$ 為兩個特例的連音效應 APs；另外為了限制韻律狀態只對目前音節的 log-F0 level 有影響，我們定義 $\boldsymbol{\beta}_{p_n}$ 為四維正交係數的第一維且都是非零值；$\mathbf{sp}_n^r$ 是正規化後的 $\mathbf{sp}_n$，為 $\mathbf{sp}_n$ 扣除 $\boldsymbol{\beta}_{t_n}$、$\boldsymbol{\beta}_{p_n}$、$\boldsymbol{\beta}^f_{B_{n-1},tp_{n-1}}$、$\boldsymbol{\beta}^b_{B_n,tp_n}$ 和 $\boldsymbol{\mu}$ 的殘餘值(residual)。圖 3.3 為 $\mathbf{sp}_n$ 與影響因素的關係表示圖，藉由假設 $\mathbf{sp}_n^r$ 是一平均值為零的高斯分佈(Normal distribution)，即 $N(\mathbf{sp}_n^r;\mathbf{0},\mathbf{R})$，可以得到

$$P(\mathbf{sp}_n \mid p_n, B_{n-1}^n, t_{n-1}^{n+1}) = N(\mathbf{sp}_n;\boldsymbol{\beta}_{t_n}+\boldsymbol{\beta}_{p_n}+\boldsymbol{\beta}^f_{B_{n-1},tp_{n-1}}+\boldsymbol{\beta}^b_{B_n,tp_n}+\boldsymbol{\mu},\mathbf{R}) \quad \text{for } 1\le n\le N \tag{3-8}$$

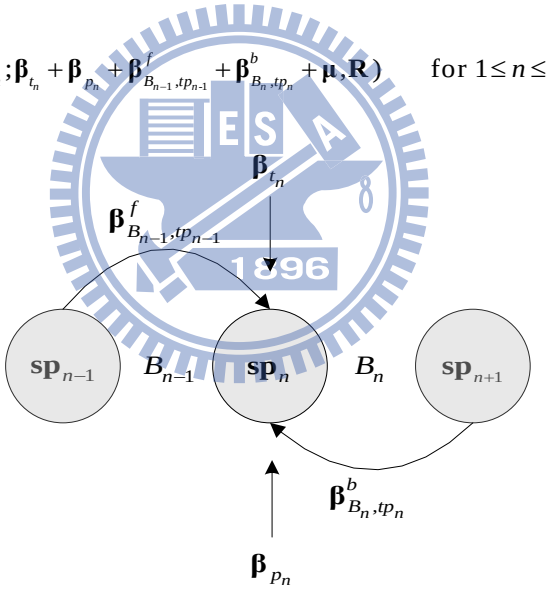

圖 3.3 觀察到的音節基頻軌跡與其影響因素的關係圖

第二個模型模擬音節長度 sd_n，可表示成：

$$P(sd_n \mid q_n, t_n, s_n, u_n) = N(sd_n;\gamma_{t_n}+\gamma_{q_n}+\gamma_{s_n}+\gamma_{u_n}+\mu_d, R_d) \tag{3-9}$$

其中 γ 表示各個影響因素的 AP，q_n 與 s_n 分別為第 n 個音節的音長韻律狀態與音節類別，u_n 則代表句子的影響因素，μ_d 與 R_d 分別表示全域平均值與音長殘餘值的共變異數矩陣；第三

個模型模擬了音節能量 se_n，可表示成：

$$P(se_n \mid r_n, t_n, f_n, u_n) = N(se_n; \alpha_{t_n} + \alpha_{r_n} + \alpha_{f_n} + \alpha_{u_n} + \mu_e, R_e) \tag{3-10}$$

其中 α 表示各個影響因素的 AP，r_n 與 f_n 分別為第 n 個音節的能量韻律狀態與韻母類別，μ_e 與 R_e 分別表示 global mean 與音節能量殘餘值的共變異數矩陣。

接著對停頓聲學模型 $P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$ 進一步化簡：

$$P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L}) \approx P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{l}) \approx \prod_{n=1}^{N} P(pd_n, ed_n, pj_n, dl_n, df_n \mid B_n, \mathbf{l}_n) \tag{3-11}$$

其中 $P(pd_n, ed_n, pj_n, dl_n, df_n \mid B_n, \mathbf{l}_n)$ 是經由分類樹與決策樹(Classification and Regression Tree, CART)推導出來，其節點的分裂準則是依據最大概似函數增益(Maximum Likelihood Gain)，利用一個已經設計好的問題集，依據不同韻律邊界停頓對所有音節的 pd_n、ed_n、pj_n、dl_n 和 df_n 做分類。在此將 pd_n 以 gamma distribution 建構，而 ed_n、pj_n、dl_n 和 df_n 以 normal distribution 建構，因此 $p(pd_n, ed_n, pj_n, dl_n, df_n \mid B_n, \mathbf{l}_n)$ 會是一個 gamma distribution 和四個 normal distribution 的乘積。

$$\begin{aligned} P(pd_n, ed_n, pj_n, dl_n, df_n \mid B_n, \mathbf{l}_n) = {} & g(pd_n; \alpha_{B_n,\mathbf{l}_n}, \beta_{B_n,\mathbf{l}_n}) N(ed_n; \mu_{B_n,\mathbf{l}_n}, \sigma^2_{B_n,\mathbf{l}_n}) N(pj_n; \mu^{pj}_{B_n,\mathbf{l}_n}, \sigma^{2\,pj}_{B_n,\mathbf{l}_n}) \\ & N(dl_n; \mu^{dl}_{B_n,\mathbf{l}_n}, \sigma^{2\,dl}_{B_n,\mathbf{l}_n}) N(df_n; \mu^{df}_{B_n,\mathbf{l}_n}, \sigma^{2\,df}_{B_n,\mathbf{l}_n}) \end{aligned} \tag{3-12}$$

而韻律狀態模型可進一步針對三種韻律狀態拆解成三個子模型，表示為

$$P(\mathbf{PS}|\mathbf{B}) \approx P(\mathbf{p}|\mathbf{B}) P(\mathbf{q}|\mathbf{B}) P(\mathbf{r}|\mathbf{B}) \tag{3-13}$$

而 $P(\mathbf{p}|\mathbf{B})$、$P(\mathbf{q}|\mathbf{B})$ 和 $P(\mathbf{r}|\mathbf{B})$ 可以用雙連文模型(Bigram Models)分別表示為

$$P(\mathbf{p}|\mathbf{B}) \approx P(p_1) \left[\prod_{n=2}^{N} P(p_n \mid p_{n-1}, B_{n-1}) \right] \tag{3-14}$$

$$P(\mathbf{q}|\mathbf{B}) \approx P(q_1) \left[\prod_{n=2}^{N} P(q_n \mid q_{n-1}, B_{n-1}) \right] \tag{3-15}$$

和

$$P(\mathbf{r}|\mathbf{B}) \approx P(r_1)\left[\prod_{n=2}^{N} P(r_n | r_{n-1}, B_{n-1})\right] \tag{3-16}$$

其中 $P(p_1)$ 、 $P(q_1)$ 和 $P(r_1)$ 分別表示各個不同韻律狀態的起始機率(Initial probability)，$P(p_n | p_{n-1}, B_{n-1})$ 、 $P(q_n | q_{n-1}, B_{n-1})$ 和 $P(r_n | r_{n-1}, B_{n-1})$ 則分別表示三種韻律狀態，給定停頓標記 B_{n-1} 的情況下，從第 n-1 個音節的韻律狀態到轉移到第 n 個音節韻律狀態的轉移機率(Transition probability)。

最後簡化停頓語法模型 break-syntax 模型 $P(\mathbf{B}|\mathbf{l})$ ，假設能將每個音節邊界分開模擬，故可表示成

$$P(\mathbf{B}|\mathbf{l}) = \prod_{n=1}^{N-1} P(B_n | \mathbf{l}_n) \tag{3-17}$$

其中 $P(B_n | \mathbf{l}_n)$ 同樣由 CART 演算法依據最大概似函數增益為分裂準則訓練得到。

3.2.3 韻律標記及模型訓練方法

A-PLM 法依據最大似然法則(Maximum likelihood, ML)，同時預估 8 個韻律模型的參數並對所有語句做韻律標記，經一連串的最佳化程序直到收斂。整個演算過程可分為兩部份：初始化和疊代，初始化過程會對所有語句做初始的韻律標記，及預估前一節所討論的 8 個子模型韻律參數的初始值；疊代的過程先對所有語句定義一概似函數(Likelihood function)

$$\begin{aligned} Q = &\left(\prod_{n=1}^{N} P(\mathbf{sp}_n \mid p_n, B_{n-1}^{n}, t_{n-1}^{n+1})\, p(sd_n \mid q_n, t_n, s_n, u_n)\, p(se_n \mid r_n, t_n, f_n, u_n)\right) \\ &\left(P(p_1)P(q_1)P(r_1)\prod_{n=2}^{N} P(p_n \mid p_{n-1}, B_{n-1})P(q_n \mid q_{n-1}, B_{n-1})P(r_n \mid r_{n-1}, B_{n-1})\right) \\ &\left(\prod_{n=1}^{N-1}\left(p(pd_n, ed_n, pj_n, dl_n, df_n \mid B_n, \mathbf{l}_n)P(B_n \mid \mathbf{l}_n)\right)\right) \end{aligned} \tag{3-18}$$

接著利用一個多重步驟的疊代程序，反覆更新所有韻律標記和 8 個韻律子模型的參數，詳細的說明可參考【7】。

3.3 以韻律標記幫助訓練 HSMM

本研究利用 HSMM 來描述語音的頻譜特性變化，本節將詳細介紹訓練 HSMM 模型的過程，與如何利用韻律標記來幫助模型訓練。

3.3.1 特徵參數抽取

訓練模型前，必須擁有足以充分描述語音特性之特徵參數，而語音處理當中，梅爾頻率倒頻譜係數（Mel-Frequency Cepstrum Coeiffient, MFCC）是一最廣泛為人使用之頻譜特徵參數，本研究也將使用此特徵參數對音檔做預切割，以 32 毫秒之漢明窗（Hamming window）且每位移 10 毫秒為一筆資料，求取 12 維 MFCC 並加上一維能量係數，以及這 13 維係數之一階與二階變量（Delta and delta-delta）為特徵參數，但因能量此維較缺乏鑑別性，因此去除能量係數，得到 38 維向量作為本研究切割語音資料之聲學特徵參數；另一方面，合成則使用廣義梅爾倒頻譜參數(Mel-generalized Cepstrum, MGC)，其中設定 α=0.42、γ=0 (mcep)，以 25 毫秒之漢明窗(Hamming window)，每位移 5 毫秒為一筆資料，求取含能量共 25 維的參數，並使用此 25 維參數的一階與二階變量，總共得到 75 維的聲學參數，而因為合成時需要有能量大小的資訊，故在這裡能量一併拿來訓練模型。

3.3.2 音檔切割

本研究所用語料皆為長句，若採用平均切割(Flat-start)的方式取得初始模型並不適當，基於對模型訓練品質的要求，在訓練模型前先利用一可靠的語者獨立模型(Speaker Independent , SI)對音檔作強制切割(Forced alignment)得到一可靠的切割位置，再根據此切割位置求得初始模型進行訓練。下圖為音檔切割的流程圖，首先在前處理階段求得 39 維的 MFCC(模型為 38 維)，因語料已有人工修正過之音節切割位置，雖然此切割位置有尾音切割不準的問題，但為了使切割的音素(Phone)邊界更加準確，仍假設此音節邊界為正確切割位置，固定音節邊界(對參數切割成音節為單位之音段，得到音素切割位置後再做串接合併)，使用語者獨立的單音素模型(Monophone model)做強制切割，以得到音素切割位置。

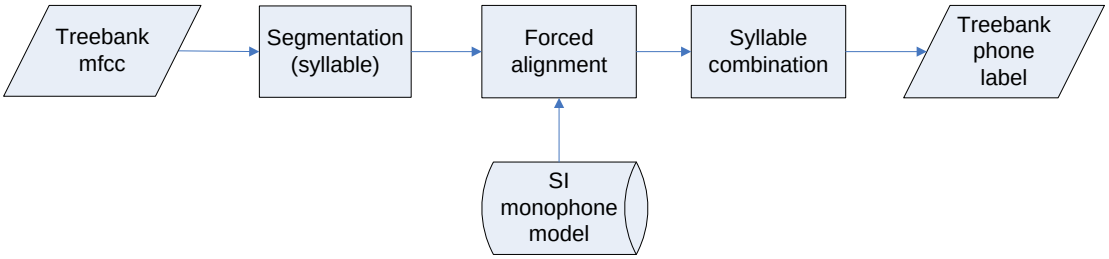

圖 3.4 音素切割位置求取流程圖

3.3.3 中文馬可夫模型設定

中文為一以音節為單位之聲調語言(Tonal language)，每個中文字對應一個音節。中文有 411 個基本音節(basic syllable)，搭配 5 種聲調，約可組成 1300 多個音節，而每個基本音節以聲母、韻母組成，韻母又可進一步拆解為介音、韻腹、韻尾三類，透過聲母、韻母的組合即可涵蓋大部分中文音節，本研究在建立中文音素 HMM 模型時，為探討模型單元對訓練結果的影響，分別以音素與聲、韻母為基本單元進行實作。相較我們選擇的單元，音節單元更大，每個單元特性應較一致，如此訓練之模型品質應該會較好，之所以不選擇音節當基本單元，主要是因音節的種類太多，若語料庫不夠大，可能有音節出現頻率過低，或甚至不包含部分音節的情形，如此會導致訓練得到之模型不夠強健，影響最後結果。

本研究使用之半隱藏式馬可夫模型，音素部分定義了 38 個音素(見附錄一)加句中短靜音停頓(Short pause , sp)與句首句尾長停頓(Silence)，共 40 個音素單元，其中頻譜模型(Spectrum model)的每一個音素單元之 HMM 由 3 個左至右（Left-to-right）的狀態（State）表示；聲母、韻母部分，聲母可分為 22 類，韻母可分為 40 類(見附錄一)，同樣加上句中短靜音停頓與句尾句首長停頓，每一個聲母與韻母單元皆以 5 個左至右的狀態表示；半隱藏式馬可夫模型定義狀態持續時間模型(State duration model)每一個基本單元之 HMM 只由一個狀態，即一個高斯（Gaussian）分佈表示，維度則與頻譜模型的狀態數相同，兩個模型的維度分別為 3 維與 5 維，所有頻譜與狀態持續時間模型的每個狀態皆以一個高斯分布描述其特徵參數的分布情形。

3.3.3 模型建立流程

本研究對兩種基本單元的模型訓練方式一致，因此本節中其餘小節僅對音素單元進行介紹。本研究利用 HTS2.1 的函式庫進行模型訓練，詳細流程圖參考圖 3-5，首先由預切割得到之音素切割位置，先經過 Segmental k-means 訓練得初始頻譜模型，接著利用 Baum-Welch(Forward-Backward)演算法進一步估測，並由此階段最後一次估測得到的網狀結構(Trellis)輸出音長模型(State duration model)的初始模型，合併所有獨立音素模型即為初始的半隱藏式馬可夫模型；接著根據此模型的狀態統計資訊(State statistics)與機率分布特性使用決策樹分群(Tree clustering)，對分佈近似的狀態(State)進行綑綁(Tying)，使共用同一組機率分佈，此即文脈相關模型(Context-dependent model)，而為使訓練之決策樹與模型更加強健(Robust)，本實驗利用此模型重新求取統計資訊，做第二次的決策樹分群，再使用最終得到之決策樹與模型進行合成。

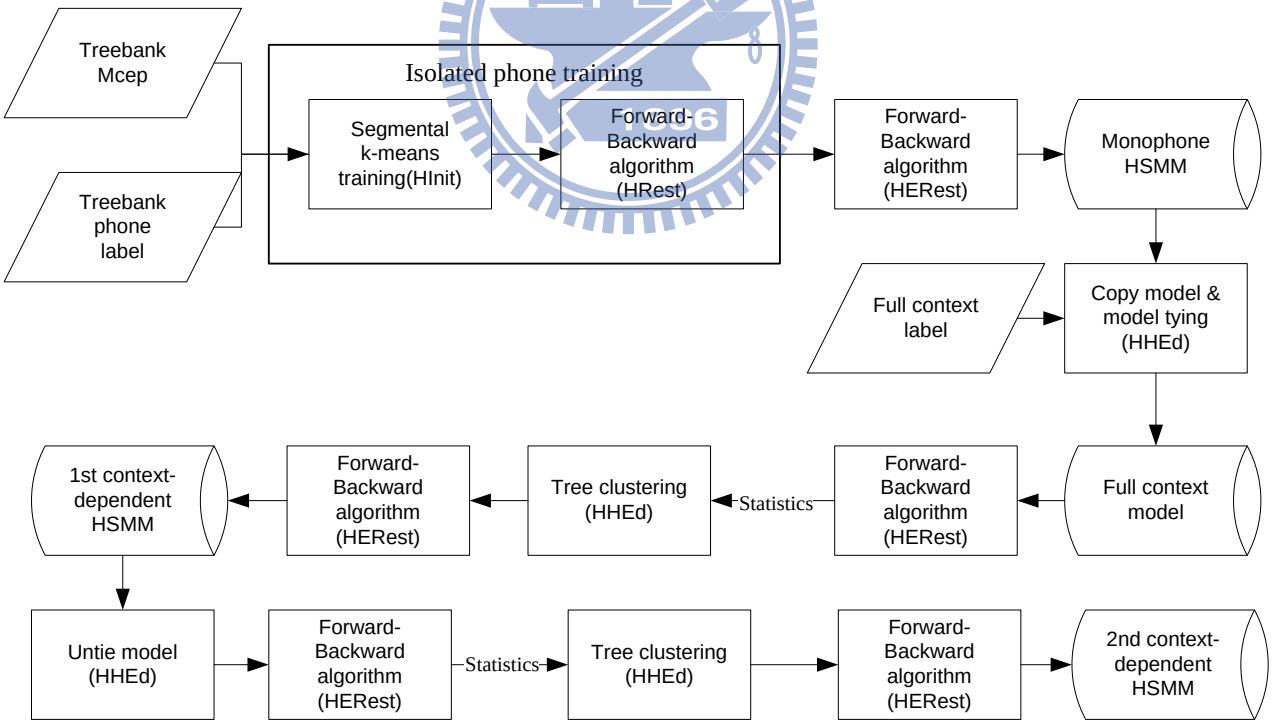

圖 3.5 HSMM 模型訓練流程

在此，文脈相關模型利用決策樹根據文脈相關標記做分群，頻譜與狀態持續時間兩個模型各有一顆決策樹，決策樹的頻譜模型由每個文脈相關(Context dependent)標記之中心音素(Central phone)的每個狀態為一根節點(Root node)開始分裂(往後稱此為 phone-based，而決策樹的狀態持續時間模型結構兩種單元並不相同，詳見 3.4 節)，使用的分裂準則為最小描述長度(Minimum Description Length , MDL)增益，化簡後如式 3-20，式 3-19 即為最大概似函數增益(Maximum Likelihood Gain)的式子，可發現式 3-20 為式 3-19 再加上一個項，一般稱為懲罰項(Penalty)，MDL 準則即利用此項來控制決策樹之分裂深度，當決策樹分裂越深，模型益加複雜，此項也會跟著變大，如此可自動使決策樹在最恰當的深度終止分裂，避免分裂過深造成過適(Overfitting)的情形，解決最大概似函數增益，若臨界值(Threshold)調整不適當，對結果有很大影響的缺點。

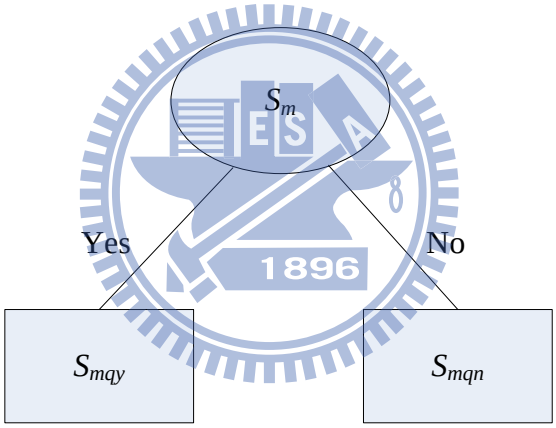

圖 3.6 MDL-based 決策樹分裂

$$\delta(D^m)_q^{ML} = L(S_{mqy}) + L(S_{mqn}) - L(S_m) \tag{3-19}$$

$$\delta(D^m)_q^{MDL} = \delta(D^m)_q^{ML} - \alpha K \log G \tag{3-20}$$

q 為母節點 S_m 問的問題，S_{mqy} 與 S_{mqn} 則分別代表經過此問題分裂後 yes 與 no 的子節點，D^m 表示落在 S_m 這個節點的資料，$L(\cdot)$代表此節點的對數概似值(log likelihood)， G 代表整體資料量，K 則代表分裂後參數的增加量。

3.3.4 引入韻律標記

本研究希望藉引入階層式韻律架構之韻律資訊幫助頻譜模型訓練，傳統文脈相關模型，相信頻譜模型的分佈與語言資訊密切相關，故決策樹根據所能得到的語言資訊訂立問題集做分群。然而上層語言資訊，雖與頻譜模型的分佈有一定的關係存在，但兩者相關性不高，故依此類上層語言資訊對頻譜模型分群並不能得到最好的效果，為了找到一與頻譜模型更直接相關的資訊幫助分群，本研究提出利用韻律資訊取代傳統語言資訊做決策樹分群，引入本研究使用之韻律模型定義的兩類韻律標記─音節邊界停頓與韻律狀態。

韻律標記為合併考量聲學資訊與語言資訊得到的標記結果，因標記過程同時考慮了聲學參數，所以應較單純語言資訊更有能力描述頻譜參數的分佈特性。其中音節邊界停頓可表示前後音素連音效應強弱，音高、音長與能量三種韻律狀態描述了韻律架構上層對音節韻律的影響，此類韻律資訊可視為介於上層語法資訊(Syntax)與音節層之間的一中層資訊，其描述較音節層面範圍更大的影響，且較上層語法資訊更加貼近模型基本單元的真實聲學分布情形。換言之，本研究目標是藉由階層式韻律資訊的引入使決策樹產生更好的分群結果，避免聲學特性相異的單元互相混淆，降低訓練得到的頻譜模型參數平滑、特性不明顯的影響。

3.3.5 文本標示資訊與問題集設計

欲利用決策樹以韻律資訊對頻譜模型做分群，文本標示資訊與問題集的設計就顯得格外重要，兩者的選擇會大大影響決策樹最後分裂的結果。首先選定與聲學特性相關的文本資訊，本論文與傳統方法不同，不考慮多層複雜的語言資訊，只選定最底層的音素資訊加上韻律資訊，完全不加入任何上層語言資訊，僅利用韻律資訊幫助模型分群，依據表 3.2 所示文脈相關資訊，對所有語料之文本建立如下文脈標示：

Pr_Phn-Cur_Phn+Fol_Phn/p:Pit_ps/q:Dur_ps/r:Ene_ps/pb:Pr_B/nb:Fol_B

表 3.2 文脈相關資訊

Class	ID	Description
Linguistic Feature	Pr_Phn	Previous phone(Initial/Final)
	-Cur_Phn	Current phone(Initial/Final)
	+Fol_Phn	Following phone(Initial/Final)
Prosodic Feature	/pb:Pr_B	Previous break type of current syllable
	/nb:Fol_B	Following break type of current syllable
	/p:Pit_ps	Pitch prosody state of current syllable
	/q:Dur_ps	Duration prosody state of current syllable
	/r:Ene_ps	Energy prosody state of current syllable

接著根據標示的資訊設計相關問題集，為了達到最佳狀態分裂合併結果，本研究一共考量三大類問題集，說明如下(詳細韻律相關問題集請見附錄二)：

1. 音素(聲、韻母)相關：因為決策樹是以當前的音素為根節點下去分裂，當前音素並不會有混淆的情形，所以只考慮前、後相接的音素，分兩類問題：

   i. 每個音素個別問。

   ii. 音素依據發音方式跟發音位置，將相似特性的音素歸成一群，如摩擦音、爆破音...。

2. 音節停頓相關：藉由韻律模型標記出{B0, B1, B2-1, B2-2, B2-3, B3, B4}七類音節邊界停頓， 因為停頓屬於音節邊界(Juncture)的影響因素，且認為每個音節前、後邊界的停頓類別，應對整個音節的音素皆有影響，所以考慮當前音素所屬音節之前、後邊界是屬於哪種停頓類別，分兩類問題：

   i. 個別考慮每個停頓類別。

   ii. 將相似特性的音節邊界停頓類別歸群來問，總共規出下列四群，

   - {B0, B1}皆為很短幾乎不可察覺的停頓
   - {B2-1, B2-2, B2-3}都為詞邊界的停頓
   - {B2-2, B3, B4}同為較明顯可察覺的停頓
   - {B3, B4}因 B3、B4 有許多是標點符號的長停頓，相較於 B2-2 是更明顯的停頓，故另外增加這個類別

3. 韻律狀態相關：總共有音高、音長與能量三類韻律狀態，因為此為音節上的資訊，所以只考慮當前音素所屬的音節為哪種韻律狀態，因最高與最低的幾個類別特性較為明顯，故將

每類 16 種韻律狀態皆分為三大類，分別是最高和最低與中間的狀態，最高與最低的問法相似，故可視為兩類問題：

i. 針對最高或最低的幾種狀態歸群，歸群方法以最高或最低的前 N 個為一群問(N≤5)，舉最低來說是{1}、{1,2}、…、{1,2,3,4,5}，因我們認定低的群必有特性明顯一致之處，但不確定框出哪幾個低的狀態為一群較準確，所以定義各種範圍大小由決策樹選擇

ii. 中間的 3~14 則分三種集合大小，使用涵蓋中間整個集合的不同分群大小來考量，

- 每四個一群，一次遞移一個，如{3,4,5,6}、{4,5,6,7}、…、{11,12,13,14}
- 考慮大一點的集合，六個一群，如{3…8}、{6…11}、{9…14}
- {3…14}整個一群

3.4 模型訓練結果分析

本研究目標為引入階層式韻律訊息幫助模型訓練，進一步提升聲學模型的品質，以求得到更好的合成聲音品質。而本研究使用的模型為 HSMM，其同時訓練頻譜與狀態持續時間模型，但模型分群是採用分開的兩棵決策樹，因此可以在同樣參數下選定兩組不同的問題集，針對兩組模型各自選定有意義的問題，限制問題集避免不相關的問題影響分群，甚至可同時採取不同的模型結構訓練模型。模型結構所指為決策樹欲歸群參數的類別，總共分為兩類，一類為音素相關決策樹(Phonetic decision tree)，對每個音素中屬於不同狀態的參數個別去歸群，決策樹的數量會等於音素單元個數乘上 HMM 的狀態數，往後以 phone-based 表示之；另一類則為共享單元之決策樹(Shared decision tree)，對屬於同一 HMM 狀態的所有音素單元對應參數歸群，決策樹數量會等於 HMM 的狀態數，往後以 state-based 稱之。

因本研究著重在韻律訊息對於頻譜模型的影響，因此本節的分析將著重在頻譜模型與其決策樹的生長情形，首先分析決策樹的結構，再以客觀實驗與主觀實驗探討韻律資訊的加入是否有效提升合成品質。

3.4.1 決策樹分析

文脈相關模型的好壞與決策樹的分裂情形密切相關，決策樹經過一個問題分裂後，分裂為子節點的兩群資料在空間中即被完全分開，若問題不恰當產生不好的分群，子節點得到的資料分佈易使接續的問題也變的不合理，造成的影響可能會延續到最後的葉節點，如此得到的決策樹偏壓嚴重，對不涵蓋在訓練資料中的外來資料分群效果必定不好。因此一棵決策樹中越上層被問出的問題，表示其對整體資料的分群影響越大，也代表這個問題相對其他問題，有能力對大群的資料分出差異明顯的兩類，若問到的問題確實能有效區別資料差異，更可避免差異極大的資料混成一群，提升最後訓練得到的模型效果。

經由觀察決策樹的分裂情形可判定分群結果是否有意義並符合預期，由於 state-based 決策樹深度過深不易分析，因此本節分析的頻譜決策樹結構為 phone-based，這也是本研究最後所採用的模型結構。首先表 3.3 與表 3.4 統計了兩種基本單元的決策樹韻律相關問題在每棵決策樹中第一次出現的位置，可發現近八成的決策樹都在前兩層就已經出現韻律相關問題，本研究使用韻律標記來代表語音階層式韻律架構造成的影響，也認定韻律上的變化與頻譜的變化有極大相關性，因此期許韻律相關問題可有效在前端被問出，統計結果也確實如預期，大部分的決策樹都在前端就出現韻律相關問題，證明韻律標記可在決策樹過程有效影響模型特性，如此若韻律標記確實與頻譜模型特性直接相關，將可使最終分類結果的每個節點，對應的模型分佈特性明顯。

另一方面，本研究也期望音節邊界停頓標記能有效區隔音節連音效應強、弱的兩類，當音節間停頓類別屬於長停頓的 B3 或 B4 時(或部分的 B2)，就不該跨過音節去考慮前、後音素彼此的影響，若音節間有長停頓卻仍出現過多跨音節考慮連音的問題有兩種可能，一個是代表停頓標記對分群無效或停頓標記不一致，但此標記在江振宇博士所發表之韻律模型的相關文獻中已證實是有效且一致的，因此若有問題即很有可能是問題集設計不當，造成決策樹的分類結果不合理，如此問到的問題將只是剛好符合訓練語料的分佈，當決策樹應用於測試語料，就很容易將資料判定到不合理的一群中。為了判定上述現象是否發生，我們對決策樹

統計，當出現長停頓類別的問題後，「是」或「否」兩群再問跨過音節的音素連音問題數量，結果「是」或「否」兩群再問跨過音節的音素連音問題數量比例約為 1:25，而聲、韻母單元的比例更達到約 1:50，經長停頓類別問題分類後的兩群資料，繼續考慮跨音節連音影響的比例差異極大，可間接證明停頓類別確實能有效區分出音節間的連音效應強弱，更證實決策樹有達到我們預期的分裂情形，由此可初步推斷決策樹的分裂結果應是合理的。(聲母與韻母加起來應有 62 類，但因 3 類韻母再訓練與料並沒有出現，所以表 3.4 中的個數總共只有 295 個。)

表 3.3 音素單元決策樹韻律相關問題第一次出現位置統計

層數	1	2	3	4	No
數量	41	50	18	2	3
累計	41	91	109	111	114

表 3.4 聲、韻母單元決策樹韻律相關問題第一次出現位置統計

層數	1	2	3	4	No
數量	189	64	24	4	14
累計	189	253	277	281	295

兩種單元中，聲、韻母單元在音節中的組成結構較為單純，方便分析，因此對其進一步觀察在決策樹根節點出現的韻律相關問題，可發現有以下幾個明顯的現象：一、當 HMM 狀態距離音節邊界越遠，音節停頓相關問題出現的比率就越低，停頓類別可對應連音效應的影響強弱，距離邊界越遠越不受前、後音節影響非常合乎常理。二、具明顯停頓特性的類別，可有效區分音節間連音關係是否存在，而連音效應對於聲、韻母的變化有極大影響，由表 3.5 統計結果確實可看出，在所有出現的停頓問題中，長停頓的問題(Bclass3 與 Bclass4)佔了絕大多數。三、相較於靠近音節邊界的 HMM 狀態，音節中間接近聲母、韻母交界處的 HMM 狀態，則容易出現韻律狀態相關問題，且以最高群組之音高韻律狀態為主，此與我們設計問題集時的預期相符合，極高或極低的韻律狀態特性明顯，對分群應會有較大影

響，且根據音高韻律狀態標記結果，可發現狀態高群組的數量多於低的群組，所以上層出現問題確實也多以高狀態(PitchPS16to12 和 PitchPS16to13)為主。

表 3.5 聲、韻母單元決策樹根節點所出現之韻律相關問題前的個數統計(前十名)

順序	問題	數量	順序	問題	數量
1	PitchPS16to12	76	6	PitchPS16to14	6
2	R=Bclass3	54	7	EnePS16to12	4
3	L=Bclass3	22	8	PitchPS3to8	3
4	PitchPS16to13	9	9	L=B0	2
5	R=Bclass4	9	10	R=Bclass1	2

而觀察傳統方法的標記與問題集，可發現傳統方法的文脈相關標記，將短靜音停頓(Short pause, sp)當做一個基本單元一併考慮進標記中，如此當音節間有 sp 時，因為模型只考慮前、後相連一個音素的影響，若有 sp 則表示必不跨過音節考慮音素連音效應，如此看來似乎可比我們所提出之考慮音節間停頓類別的方法更有意義，但由表 3.6 統計結果發現，此方法限定只要超過一定停頓時長(本論文定義>25ms)，即視為 sp，如此的標記並不可靠，根據韻律標記結果對 sp 的位置做統計，結果可發現 sp 所對應的停頓特性並無一致性，甚至在模型標記為 B1 人耳幾乎不可察覺之停頓位置，出現比例都是所有類別中最高的，代表直接給定一停頓時長標準，以此判定考不考慮連音效應影響似乎並不恰當，除停頓特性不一致有可能使問題不易出現外，若問題出現了也有可能因特性並不正確，造成資料差異極大卻混淆成一群的情形。

表 3.6 各類韻律標記停頓類別與短靜音停頓對應情形

停頓類別	B0	B1	B2-1	B2-2	B2-3	B3	B4
短停頓個數	0	4284	1511	3697	848	3778	3263

3.4.2 客觀結果分析

為得到一合理切割位置做參數還原，首先利用訓練得到之頻譜模型在固定音節長度的情況下對原始音檔做強制切割，因為訓練得到的 HSMM 中，頻譜模型並不包含轉移機率，為單純探討頻譜模型的優劣，人工給定每個狀態的轉移機率為 0.5，只允許由左到右一次前進一個相鄰狀態，此作法相當於完全相信頻譜模型所估出的分數對音檔進行切割，換言之即對應此頻譜 HMM 最佳的切割位置，由此切割位置衡量頻譜模型的優劣應是公平合理的；接著以切割所得到之狀態持續時間對模型做參數產生，便可還原得頻譜參數，以此與原始音檔抽取之頻譜參數計算梅爾倒頻譜距離(Mel-cepstral distance,MCD)，見式 3.21，此式計算的 MCD 值越小表示預估得到的參數軌跡(trajectory)越接近原始的參數軌跡。

$$\mathrm{MCD[dB]}=\frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}(c_d-\hat{c}_d)^2} \tag{3-21}$$

D is the analysis order of mel-cepstra

c_d and $\hat{c}_d$ are the d-th coefficients of the mel-cepstra of generated and natural speech

透過與傳統方法比較還原後的結果，計算何者較接近原始的參數軌跡來衡量方法是否有效，這邊所指的傳統方法，指的是給定同第二章中介紹的傳統方法參數與問題集，而以此章敘述之同樣方式訓練得到的模型。由於 HSMM 模型的狀態持續時間模型與頻譜模型的結構可不相同，故針對兩種基本單元對傳統方法與所提方法分別比較三種模型結構的合成結果，以(頻譜模型結構,音長模型結構)表示之，三種組合分別為(state-based, state-based)、(phone-based, phone-based)、(phone-based, state-based)，另一種組合(state-based, phone-based)則不納入考慮，因為本研究認定頻譜模型採 state-based 做法，有音素聲音互相混淆，導致結果變差的可能，故希望改採取 phone-based 做法取代傳統 state-based 的做法，故在此考慮(phone-based, state-based)與(phone-based, phone-based)兩組結構，而(state-based, state-based)之所以列入考慮是因為傳統 HTS 方法採用此模型結構。

下表 3.7 與 3.8 分別為音素與聲、韻母單元對兩種方法與三種結構，依據強制切割得到

之狀態持續時間對頻譜模型還原的結果，其中所有實驗對於訓練語料的 MCD 計算，因為對全部語料計算耗時過久，故僅挑選其中 50 句，訓練語料部分則對 43 句全部計算，分有聲與無聲根據手動修正過之聲、韻母切割位置當標準答案，對每個音節依據語言學家觀察出之特性，將所有韻母與聲母中的 m、n、l、r、INULL 判定為有聲，音素單元則以韻母對應之所有音素與聲母對應音素中的 m、n、l、r 判定為有聲，其餘聲母皆為無聲，音節內的有聲、無聲部分再分別以動態時軸校正(Dynamic Time Warping, DTW)演算法，尋找兩不同長度的音框序列間最為恰當的一對一對應關係，計算 MCD。

由結果可看出我們所提方法除了在聲、韻母單元的 state-based 不如傳統方法外，其餘不論何種單元、何種模型結構在訓練與測試資料的 MCD 計算皆勝過傳統方法，觀察影響發音結果較大的 voiced 部分，可發現我們方法在音素單元的模型中平均勝過傳統方法約 0.09dB，在聲、韻母單元中更平均勝過傳統方法約 0.11dB，更進一步，證實本研究提出以韻律標記取代傳統語言參數對頻譜模型做分群是有效的。

兩種單元互相比較可發現，對訓練語料的計算結果，聲、韻母單元略勝過音素單元，此為一合理結果，因單元越大，同一單元間的變異性應越小，而其缺點就反應在測試語料的結果上，因聲、韻母單元類別較多，考慮文脈相關模型時，可能出現的文本組合數相較於音素單元大幅增加，故訓練語料所能涵蓋的模型可能就越少，測試語料也更容易出現訓練未知的文本組合，故若單純考慮頻譜模型，根據 MCD 計算結果，可推斷音素單元可能為一較佳的選擇。在多種模型結構中，音素單元本論文所提方法以(phone-based, phone-based)的結構組合結果最好，但與(phone-based , state-based)相去不遠，傳統方法則以(phone-based , state-based)最好；聲、韻母單元則兩種方法皆以(phone-based , state-based)結果最好，其中本論文所提方法與(phone-based, phone-based)結果幾乎一樣，由此結果可推斷模型結構的好壞與問題集和選定的資訊有關聯，所以並無法定義何種結構一定可得到最好結果，但由結果可證實前面的推論，頻譜模型若採 state-based 訓練，確實會因模型較混淆導致結果較差，推測聲、韻母單元中本論文所提方法之所以會大幅下降，也是因為此原因；另外前面提過兩棵決策樹可使用不同的問題集，目前本論文所提方法的問題集主要是以對頻譜影響為考量所訂立，所以可能對

音長決策樹並不完全適合，但經初步嘗試發現並無法找到一更好的問題集組合，可進一步提升合成結果，因為本論文採用的 HSMM 模型訓練方法，兩個模型在訓練過程會互相影響，改變一模型另一個模型也會產生變化，對參數微調找出一合適問題集組合並不容易，所以在本研究中便同傳統 HTS 方法，只單純考慮一種問題集。

表 3.7 音素單元之 MCD 計算結果

(頻譜模型結構，狀態持續時間模型結構)	Training		Testing	
	Voiced	Unvoiced	Voiced	Unvoiced
Proposed method(state, state)	4.7152	4.5009	4.8556	4.6569
Conventional method(state, state)	4.8080	4.4813	4.9438	4.6536
Proposed method(phone, phone)	4.6571	4.4627	4.8356	4.6404
Conventional method(phone, phone)	4.7420	4.4472	4.9285	4.6549
Proposed method(phone, state)	4.6585	4.4639	4.8357	4.6396
Conventional method(phone, state)	4.7405	4.4506	4.9240	4.6540

表 3.8 聲、韻母單元之 MCD 計算結果

(頻譜模型結構，狀態持續時間模型結構)	Training		Testing	
	Voiced	Unvoiced	Voiced	Unvoiced
Proposed method(state, state)	4.8708	4.6443	5.2775	4.8244
Conventional method(state, state)	4.8045	4.4398	5.1544	4.6390
Proposed method(phone, phone)	4.6047	4.3964	5.0109	4.6254
Conventional method(phone, phone)	4.7192	4.4085	5.1337	4.6286
Proposed method(phone, state)	4.6027	4.3916	5.0106	4.6220
Conventional method(phone, state)	4.7157	4.7142	5.1278	4.6327

進一步針對頻譜模型分析所提方法與傳統方法的優缺點，以音節前、後所接韻律停頓類別去分別計算兩種方法中最好的模型結構組合，針對不同停頓類別的 MCD，討論以頻譜模型強制切割之切割位置還原之統計結果，音素單元與聲、韻母單元統計結果分別為表 3.9、

表 3.10 與表 3.11、表 3.12，表格代表的意思，舉例來說「停頓後」與 B0 交接那一欄代表音節停頓類別 B0 後的音節之 MCD；End 與 Start 則分別代表句尾與句首，這兩類在模型內仍歸類為 B4，僅此處統計時獨立出來。觀察兩組結果可發現所提方法不論何種單元皆僅在 B4、與 End 兩類停頓前的音節不如傳統方法，而音素單元在 B2-1、B4、Start 三類停頓後接的音節，聲、韻母單元除了上述三類外，在 B0 與 B2-3 兩類停頓前接的音節也大幅勝出，其中 B2-1 在韻律模型中定義為音節有明顯音高重置的停頓，可視為對停頓後音節後有較一致影響的聲學特徵，確實在停頓後之音節的 MCD 明顯勝出，同樣 B2-3 定義為音節延長效應明顯的停頓，可視為對停頓前音節有較一致影響的聲學特徵，確實也在停頓前音節的 MCD 明顯勝出，由此結果可證明本研究引入的韻律標記，確實有效改進了頻譜模型的分佈結果；至於 B4 之所以會在停頓前音節不如傳統方法，停頓後音節明顯勝出傳統方法，根據問題集可判定此結果應是受標點符號的影響，由表 3.13 的統計結果可發現 B4 絕大部分對應到標點符號的位置上，因為連接不同類型的標點符號，會讓句尾發音特性產生變化，傳統方法定義了各類標點符號的問題集，因此有機會分出不同標點符號所造成的影響，改進標點符號處的表現。

表 3.9 音素單元傳統方法(phone , state)對測試語料音節前後之停頓類別計算 MCD

	B0	B1	B2-1	B3	B4	B2-2	B2-3	End	Start
停頓後	4.9808	4.5850	5.2921	4.9152	5.9115	4.8011	4.6511		5.7133
停頓前	5.1384	4.9216	4.5637	4.3395	4.1607	4.6445	4.9287	4.2214	

表 3.10 音素單元我們方法(phone , phone)對測試語料音節前後之停頓類別計算 MCD

	B0	B1	B2-1	B3	B4	B2-2	B2-3	End	Start
停頓後	4.9156	4.5331	5.1013	4.8672	5.5995	4.7888	4.6072		5.5138
停頓前	5.0308	4.8294	4.4742	4.2844	4.2322	4.6242	4.8610	4.3291	

表 3.11 聲、韻母單元傳統方法(phone , state)對測試語料音節前後之停頓類別計算 MCD

	B0	B1	B2-1	B3	B4	B2-2	B2-3	End	Start
停頓後	5.1982	4.6887	5.3303	5.0378	6.0652	4.9095	4.7153		5.9371
停頓前	5.3026	5.0009	4.7041	4.5164	4.2285	4.7879	5.1409	4.2546	

表 3.12 聲、韻母單元我們方法(phone , state)對測試語料音節前後之停頓類別計算 MCD

	B0	B1	B2-1	B3	B4	B2-2	B2-3	End	Start
停頓後	5.0817	4.6284	5.1002	4.9316	5.6584	4.8511	4.6726		5.6711
停頓前	5.1533	4.8715	4.6491	4.4333	4.3303	4.7538	4.9994	4.4140	

表 3.13 標點符號對應各類停頓標記之個數統計

停頓標記 / 標點符號	B0	B1	B2-1	B2-2	B2-3	B3	B4	BE	總計
句號	0	0	0	8	1	45	603	326	983
問號	0	0	0	0	0	9	56	28	93
驚嘆號	0	0	0	1	0	1	38	17	57
分號	0	0	0	0	0	2	69	1	72
逗號	0	3	3	45	6	1594	1856	3	3510
冒號	0	0	0	0	0	1	1	0	2
頓號	1	5	6	65	2	156	29	0	264
其他	1	0	0	0	0	0	0	0	1
非標點符號	3232	8818	4713	3303	2993	1705	41	0	24805
總計	3234	8826	4722	3422	3002	3513	2693	375	

因為實際合成時，是利用狀態持續時間模型預估狀態時長，接著討論以狀態時間模型預估音長還原之結果，在固定音節長度的限制下，利用狀態持續時間模型以 Yoshimura【12】所提出之狀態音長預估方法預估狀態音長，見下式 3-22~24，其中 T 在本研究中代表音節長

度，K 表音節內 HMM 狀態數，d_k 表示狀態時長，$\xi(k)$ 為模型中第 k 個狀態的平均值(Mean)，$\sigma^2(k)$ 為模型中第 k 個狀態的變異數(Variance)，得到狀態時長後再搭配頻譜模型還原頻譜參數計算 MCD，衡量訓練之 HSMM 的好壞，由下表 3.14 與 3.15 可觀察到兩種單元的結果皆比單純考慮頻譜模型略差一點，但本論文所提方法仍是勝過傳統方法的，由此結果可推斷在此限定條件下(固定音節邊界)，所提方法訂立的問題集音素單元搭配(phone, phone)的模型結構，聲、韻母單元搭配(phone, state)模型結構可訓練出最好的 HSMM，此外觀察兩個表對照強制切割的結果可發現，原先音素單元 MCD 計算結果較聲、韻母單元小的情形，此處卻反了過來，由此推斷音素單元對應的狀態持續時間模型，預估出之狀態音長可能較不穩定，無法搭配頻譜模型產生一良好的序列，導致與強制切割的情況相比 MCD 大幅上升。

$$T = \sum_{k=1}^{K} d_k \tag{3-22}$$

$$d_k = \xi(k) + \rho \cdot \sigma^2(k) \tag{3-23}$$

$$\rho = (T - \sum_{k=1}^{K} \xi(k)) \Big/ \sum_{k=1}^{K} \sigma^2(k) \tag{3-24}$$

表 3.14 音素單元之 MCD 計算結果

(頻譜模型結構，狀態持續時間模型結構)	Training		Testing	
	Voiced	Unvoiced	Voiced	Unvoiced
Proposed method(state, state)	4.9819	4.5637	5.1335	4.6908
Conventional method(state, state)	5.0807	4.5668	5.2003	4.6910
Proposed method(phone, phone)	4.9182	4.5255	5.1119	4.6748
Conventional method(phone, phone)	5.0120	4.5287	5.1916	4.6868
Proposed method(phone, state)	4.9235	4.5299	5.1135	4.6755
Conventional method(phone, state)	5.0228	4.5336	5.1877	4.6818

表 3.15 聲、韻母單元之 MCD 計算結果

(頻譜模型結構，狀態持續時間模型結構)	Training		Testing	
	Voiced	Unvoiced	Voiced	Unvoiced
Proposed method(state, state)	5.1220	4.8424	5.2690	4.9743
Conventional method(state, state)	5.0669	4.5126	5.2013	4.6445
Proposed method(phone, phone)	4.8484	4.4520	5.0814	4.6186
Conventional method(phone, phone)	4.9841	4.4740	5.1840	4.6222
Proposed method(phone, state)	4.8644	4.4640	5.0757	4.6376
Conventional method(phone, state)	4.9925	4.4814	5.1823	4.6309

3.4.3 主觀結果分析

根據上一小節還原的頻譜參數給定正確基頻值實際合成聲音，因重點放在衡量 HSMM 的表現，此處傳統方法與所提方法皆給定正確的音節長度與音節基頻軌跡，音節中的有聲與無聲部分判定與計算 MCD 時相同，且根據原始音檔有停頓的位置(靜音停頓>25ms)給定停頓標記與正確的韻律標記，僅單純比較頻譜與音素狀態音長預估的好壞，也就是假定其他條件皆為理想，單純比較訓練得到之 HSMM 模型合成聲音的優劣。

根據非正式聽覺實驗發現兩種方法的聲音雖稍有差別，但是並無法明顯分辨出兩種方法的優劣，會有此結果推測是因此階段兩者給定的韻律條件太過類似，而聽感上的主要影響可能較直接反應在韻律上的差別，但頻譜上的差異並非不重要，要使合成器的品質有一定幅度的提升，頻譜與韻律預估絕對是需要同時變好的，更詳細的討論，將在第五章比較兩種不同方法合成器時進一步說明。

第四章 韻律產生器

在語音合成系統中，韻律產生器扮演了相當重要的角色，為了產生更自然的語音，準確的韻律預估是必須的，本研究提出利用兩類韻律標記描述整個韻律階層的影響，在訓練階段，同時給定聲學與語言參數預估韻律標記；而在合成階段，因輸入只有文字，只能根據輸入文字得到語言參數，因此得單純利用語言參數來預估韻律標記。

本研究提出利用條件式隨機域(Conditional Random Field , CRF)的方式，根據各類語言資訊預估兩類韻律標記，如圖 4.1，先利用語言資訊預估停頓類別，再結合語言資訊與預估的停頓類別預估韻律狀態，利用文字資訊搭配預估得到的韻律標記，從訓練模型中找尋對應的 APs，疊加合成出音節音高與音節音長。4.1 節簡介條件式隨機域，4.2 節介紹停頓類別標記的預估方法，4.3 節介紹韻律狀態的預估方法，4.4 節將介紹從韻律標記產生韻律參數的方法。

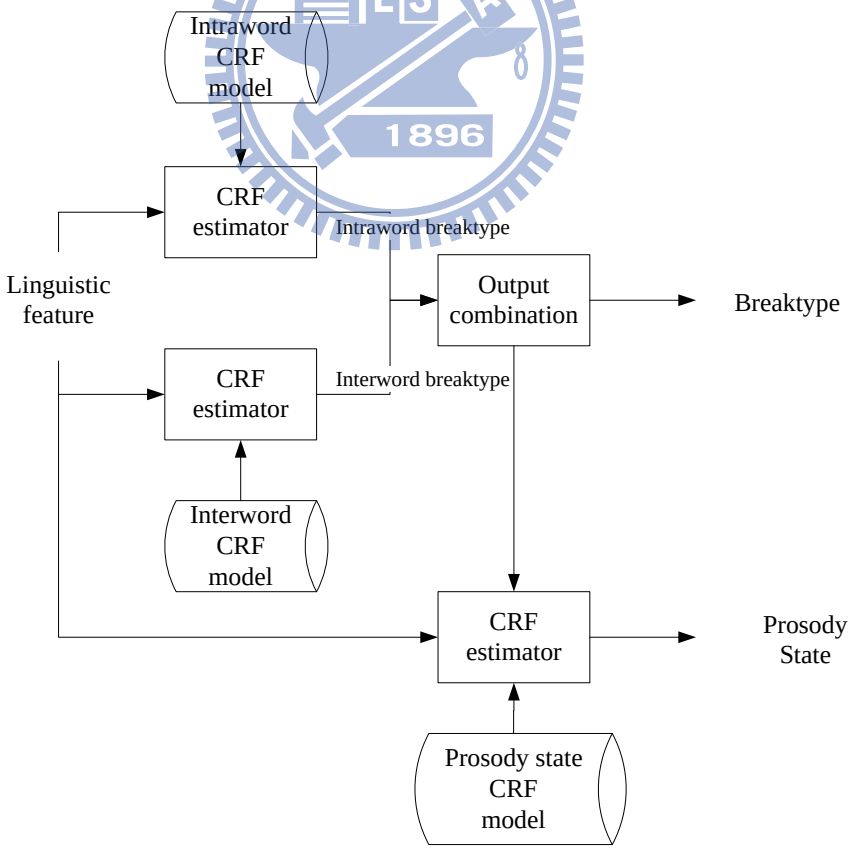

圖 4.1 CRF 預估韻律標記之系統架構圖

4.1 條件式隨機域

條件式隨機域可視為一個無向圖(Undirected graph)的模型，具序列型式架構且使用一個指數模型描述給予觀察序列條件下整個狀態序列的聯合機率，圖 4.1 即為 CRF 中最簡單的線性鏈結架構。根據 CRF 的定義，假設 *G(V,E)*為一圖型結構，而 **Y** 可被 *G* 的節點所表示 $\mathbf{Y} = (Y_v)_{v \in V}$ ，**(X,Y)**可表示為一以全域性觀察序列 **X** 為條件之條件隨機域，且Y_v 符從一階馬可夫條件，數學式可表示如下：

$$P(Y_V \mid \mathbf{X},\ Y_W, w \neq v) = p(Y_V \mid \mathbf{X},\ neighbor(Y_V)) \tag{4-1}$$

$$p(\mathbf{Y} \mid \mathbf{X}) = \frac{1}{Z(\mathbf{X})} \exp\left(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t) \right) \tag{4-2}$$

$$Z(\mathbf{X}) = \sum_{\mathbf{Y}} \exp\left(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t) \right) \tag{4-3}$$

其中$\{f_k(y_t, y_{t-1}, x_t)\}_{k=1}^{K}$為一特徵函數集合，每一特徵函數通常定義為{0,1}的雙值函數，λ_k 則為每個函數所對應的權重值，Z(**X**)則為一正規化的因子，目的為使聯合分佈機率的總合為 1。因此訓練 CRF 模型，目標即根據輸入的訓練資料估計出一組權重值，使其符合最大概似函數值或最大事後機率條件，且由於目標函數為一凸函數，可保證求得的收斂解必為一全域最佳解，此亦為 CRF 方法的一大優點。

預估階段，利用訓練得到之條件隨機域模型，根據最大事後機率準則，對已知輸入資料 **X** 使用維特比(Viterbi)演算法做動態規畫求出最佳標記序列，公式如下：

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} P(\mathbf{Y} \mid \mathbf{X}) \tag{4-4}$$

故建立 CRF 模型的兩大關鍵即為參數估計與特徵選擇，參數估計方法非本論文探討的重點，本論文目的為利用 CRF 模型做韻律標記的預估，因此後面的章節將著重在特徵選擇部分，介紹本論文對各類標記預估選定的特徵與樣板。

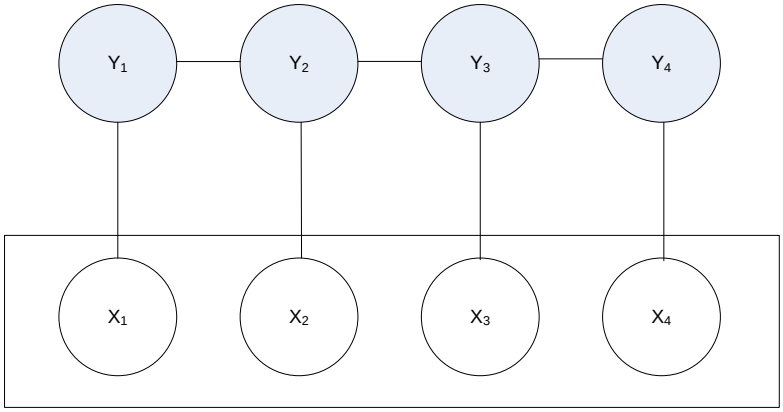

圖 4.2 線性鏈結 CRFs 圖型結構

4.2 音節邊界停頓類別預估

音節邊界依詞為單元區分，又可分詞內(Intraword)與詞間(Interword)邊界兩類，兩者特性不盡相同，詞內邊界一般發音習慣多不停頓，故音節間停頓類別多標記為 B0、B1 這類幾乎不可察覺的停頓；詞外邊界因中文構詞結構複雜，每個人發音習慣差異極大，因此本研究將詞內與詞間邊界分開預估，分別預估完成後再對結果進行合併，即可得到每個音節邊界的停頓類別。方法如下圖 4.2 所示，訓練階段藉由語言參數與八類停頓標記之間的關係，利用 CRF 方法分別訓練一詞內與詞間邊界的停頓類別預估模型；預估階段，輸入文本所對應語言參數，即可利用訓練得道之模型根據最大事後機率(Maximum Aposterior，MAP)準則，以維特比(Viterbi)演算法估得最佳標記結果序列。

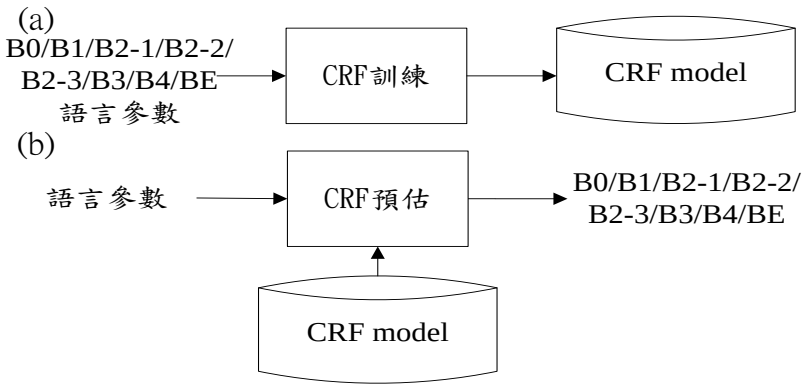

圖 4.3 CRF 停頓標記預估，(a)訓練階段，(b)預估階段

4.2.1 詞間邊界預估

詞間邊界類別為本研究韻律標記預估的重點，因為詞間邊界的停頓特性複雜不易預估，而停頓位置的正確與否卻為影響聲音聽覺感受的主要原因，停頓預估的好，對聽感韻律上的變化絕對有極大加分效果，且下一級韻律狀態的預估也須仰賴停頓標記，故如何正確預估詞間邊界，可視為本研究韻律標記預估的最重要部分。

參數與樣板的選擇，主要可分為三個階段，首先利用最直覺影響停頓的語言資訊，包括詞性(Part of speech, POS)、標點符號(PM)、詞長、與前或後 PM 的距離和詞邊界的聲、韻母這幾類資訊，並依照每個資訊的特性調整樣版；得到一基本預估正確率後，進一步加入上層語法樹資訊，希望解決標點符號位置與深度第一層邊界，容易出現長停頓之位置預估混淆情形，藉此降低模型各類停頓的混淆程度；最後再根據預估結果特別不好的部分加入針對性的問題，如連接詞(Cbb)特性、聲調特性，詳細的參數與樣板請參考表 4.1 與表 4.2。

以下對樣版的選擇稍加說明，首先是聲、韻母群組的部分，因為某些音的發音特性導致其必定無法與前一個字或後一個字連續發出，故對發音特性相近的聲、韻母規群當樣版；標點符號部分，因為標點符號大多對應到 B2 以上之停頓，且根據標點符號類別，如頓號、句號、問號...每種標點符號可能會對應到不同長短的停頓，以此為依據去規群，且標點符號前後通常不會連續發生停頓，故設計此類樣版；POS 樣版則因為部分詞性會與其他詞較容易分開來念或構成一個更大的詞甚至片語，而同時考慮前後詞時，因為若皆以 46 類考慮展開組合數會過多，所以此處離邊界較遠的位置視為影響較小以較少類別考慮；詞長樣版則因句中若有長停頓，考慮人的換氣特性，可能較容易發生在一個長詞之前或之後；考慮距離前或後最近 PM 的距離，因為標點符號一般已有長停頓，但若句子過長，因為一口氣無法講完句子，句中可能會有另外的長停頓，且應該較容易出現在距離兩側標點符號皆較遠的位置，相對的距離標點符號較近的位置也較不易出現長停頓；另外針對標點符號位置，考慮前、後句子的長度，因為標點符號位置的停頓長度，應也同樣符合人的換氣特性，會視句子的長短而不同；語法樹深度第一層位置的片語邊界，因為在一般人的發音習慣中，片語是較不易斷開的，斷

點較容易出現在片語之前或之後，因此特別針對語法樹深度第一層位置的片語邊界考慮；考慮聲調主要是因為 B2-1 會有音高重置情形，在某些連音情況下音高重置會造成聲調錯誤，故加入聲調樣版試圖解決此問題；另外也對連接詞類別特別進行考慮，因為兩個連接詞在 POS 標記屬於同一類，卻可能有兩種截然不同的發音特性，如「一方面」與「然而」兩個連接詞，前者後面往往會有停頓，後者則不會。

表 4.1 詞間邊界停頓模型的特徵參數

特徵	涵義
$PMtype_n$	詞後的標點符號類型(5 類){。！？；}、{， ：}、{、}、{others}、{non-PM}
POS_n	詞性(中研院標記之 46 類)
$POS3_n$	詞性(23 類)
$Cbbtype_n$	關聯連接詞(3 類)
WL_n	詞長(1,2,3,4,5(>4))
$WTone1st_n$	詞的第一個音節聲調(1,2,3,4,5)
$WTonelast_n$	詞的最後一個音節聲調(1,2,3,4,5)
$Wdist_n$	距離前或後 PM 位置最短幾個 word {1^{st}, 2^{nd}, 3^{rd}, [4^{th}, 5^{th}], [6^{th}, 7^{th}], [8^{th}, 9^{th}, 10^{th}, 11^{th}], 0last, 1^{st} last, 2^{nd} last, 3^{rd} last, [5^{th} last, 4^{th} last], [7^{th} last, 6^{th} last], [11^{th} last, 10^{th} last, 9^{th} last, 8^{th} last],others,single}
$Sdist_n$	距離前或後 PM 位置最短幾個 syllable {1^{st}, 2^{nd}, 3^{rd}, [4^{th}, 5^{th}], [6^{th}, 7^{th}], [8^{th}, 9^{th}, 10^{th}, 11^{th}], 0last, 1^{st} last, 2^{nd} last, 3^{rd} last, [5^{th} last, 4^{th} last], [7^{th} last, 6^{th} last], [11^{th} last, 10^{th} last, 9^{th} last, 8^{th} last], others, single}
$WS1sti_n$	詞的第一個音節之聲母發音類別 1. {b,d,g} 2. {p,t,k} 3. {m,n,l,r}

	4. {sh,s,h,f,x} 5. {ch,c,q} 6. {zh,z,j} 7. {INULL}
$WSlastf_n$	詞的最後一個音節之韻母發音類別 1. {a,ya,wa} 2. {eh,ye,yue} 3. {yi,ai,yai,wai,ei,wei} 4. {ao,yao,ou,you,o,wo,yo} 5. {an,yan,wan,yuan,en,yin,wen,yun} 6. {ang,wang,yang,eng,ying,weng,yung} 7. {FNULL1,FNULL2} 8. {e,er} 9. {yu} 10. {wu}
$PrePhSize_n$	PM 前句子的長度{0,1,[2,3],[4,5,6,7],[8,9,10,11,12],others} (詞邊界非 PM 位置給 0)
$FolPhSize_n$	PM 後句子的長度 {-2,0,[1,2],[3,4],[5,6],[7,8,9,10,11],others} (詞邊界非 PM 位置給 0，整句結尾給-2)
$PreDepth1Type_n$	語法樹結構中第一層深度(depth-1)之交界處前的片語類型(非交界處前位置或非片語標 None)
$PreDepth1Size_n$	語法樹結構中第一層深度(depth-1)之交界處前的詞或片語長度{ 0,1,2,3,4,5,[6,7,8],[9,10,11,12,13,14],orthers} (非交界處前位置給 0)
$FolDepth1Type_n$	語法樹結構中第一層深度(depth-1)之交界處後的類型(非交界處後或非片語位置標 None，此外交界處為 PM(頓號或非停頓類標點符號除外)時，標為 PM)
$FolDepth1Size_n$	語法樹結構中第一層深度(depth-1)之交界處後的詞或片語長度{ -2,-1,0,1,2,[3,4,5,6],orthers} (非交界處後位置給 0，整句結尾給-2，交界處為 PM 給-1)

表 4.2 詞間邊界停頓模型的特徵樣版

樣版
$WS1sti_{n+1}$, $WSlastf_n$
$PMtype_{n-1}$, $PMtype_n$, $PMtype_{n+1}$
POS_n, POS_{n+1}, ($POS3_{n-1}POS_n$), (POS_nPOS3_{n+1}), ($POS_{n+1}POS3_{n+2}$),
WL_n ,WL_{n+1}, ($WL_{n-1}WL_n$), (WL_nWL_{n+1}), ($WL_{n+1}WL_{n+2}$)
(POS_nWL_n)
$Wdist_n$
$Sdist_n$
$PrePhSize_n$, $FolPhSize_n$
$PreDepth1Type_n$, ($PreDepth1Type_n$, $FolDepth1Type_n$)
$PreDepth1Size_n$, $FolDepth1Size_n$
$Tonelast_n$, $Tone1st_{n+1}$
$Cbbtype_n$
($Cbbtype_n$, $PMtype_{n-1}$)

4.2.2 詞內邊界預估

詞內邊界因為停頓類別多為 B0、B1，故預估較為單純，參數及樣版的選擇可簡單分兩階段，第一階段同樣先選擇相關性應最大的語言資訊，包括詞長、音節在詞中位置與聲、韻母類別，先根據這些參數調整出一合適樣版；接著再根據預估情形，針對上述參數考慮不到的特殊情形加入新資訊幫助預估，如利用聲調解決特殊聲調連接的連音情形，並針對一些特殊長詞(多為數量複合詞)，加入一次詞彙(sub-lexical word)邊界的標籤，還有 POS 類別，考慮詞性的影響。

此處樣版的考慮大致上與詞間邊界一致，差別僅在於詞內邊界考慮的範圍限制為詞，而詞間邊界則為一整個句子，所以需考慮的樣版內容較少，此處比較特別的是額外定義了一考

慮是否為次詞彙邊界的樣版，此樣版特別為數量複合詞所定義，根據斷詞結果所有數量複合詞皆斷為一個詞彙詞，然而較長的數量複合詞，一般在發音時會有停頓不會一次念完，而這些停頓應該較可能發生在次詞彙的邊界，所以特別提出此類樣版。

表 4.3 詞內邊界停頓模型的特徵參數

特徵	涵義
WL_n	詞長(1,2,3,4,5(>=5))
Sforward	距離前面的詞邊界幾個 syllable(0,1,2,3,4,5(>=5))
Sbackward	距離後面的詞邊界幾個 syllable(0,1,2,3,4,5(>=5))
Si_n	音節的聲母發音類別 1. {b,d,g} 2. {p,t,k} 3. {m,n,l,r} 4. {sh,s,h,f,x} 5. {ch,c,q} 6. {zh,z,j} 7. {INULL}
Sf_n	音節的韻母發音類別 1. {a,ya,wa} 2. {eh,ye,yue} 3. {yi,ai,yai,wai,ei,wei} 4. {ao,yao,ou,you,o,wo,yo} 5. {an,yan,wan,yuan,en,yin,wen,yun} 6. {ang,wang,yang,eng,ying,weng,yung} 7. {FNULL1,FNULL2} 8. {e,er} 9. {yu} 10. {wu}
$POS3_n$	詞性(23 類)
$Tone_n$	當前音節的聲調(1,2,3,4,5)
$SubWbound_n$	音節後邊界是否為 sub-lexical word 的邊界(0,1,2) (最後一個音節給 2)

表 4.4 詞內邊界停頓模型的特徵樣版

樣版
Si_{n+1}, Sf_n
WL_n
$Sforward_n$
$Sbackward_n$
$POS3_n$
$Tone_{n+1}$
$SubWbound_n$

4.2.3 預估結果

本研究使用 CRF++ toolkit 做 CRF 的模型訓練與標記預估，首先依據選定之參數，將文本轉為樣板對應格式，即可利用 CRF++進行模型訓練與預估的工作，接下來將分別說明詞間邊界與詞內邊界的預估情形。

下表 4.5 為測試語料詞間邊界的預估情形，第一列的 B0~B4 表示預估結果，第一行的 B0~B4 表示標記答案，扣除 43 句的最後一個詞，總預估詞數為 1979 個，其中預估正確個數為 1152 個，正確率為 58.21%，若依表 4-6 將相近特性的停頓類別歸群統計，預估正確個數為 1445 個，正確率 73.02%，觀察結果可發現 B2 群組的三個類別預估正確率都偏低，少數邊界確實屬於可斷可不斷的地方，例如根本不該出現的 B0 預估成 B3 的兩個，基本上都屬於這一類，舉例來說{因應景氣低迷的方式為(B0)檢討企業經營發展策略}預估成{因應景氣低迷的方式為(B3)檢討企業經營發展策略}，此處即屬於可斷可不斷，不過 B3 的停頓還是有可能太長，但這類預估錯誤可能較不影響整體聽感的流暢度與語意的判定；然而其他變成長停頓類別的錯誤，可能就會使句子產生不自然停頓，如{日(B2-1)、夜間部}預估成{日(B4)、夜間部}、{中華藍(B2-1)、白}預估成{中華藍(B3)、白}、{現金(B1)增資}預估成{現金(B3)增資}等即為此類，若要進一步提升合成品質，這類問題勢必得解決。

表 4.5 測試語料詞間邊界停頓類別預估結果

	B0	B1	B2-1	B2-2	B2-3	B3	B4	Total
B0	97	27	16	8	7	2	0	157
B1	38	447	73	31	22	9	0	620
B2-1	21	79	158	29	12	6	1	306
B2-2	9	41	66	123	16	44	0	299
B2-3	16	77	28	20	32	9	0	182
B3	1	7	12	39	4	158	42	263
B4	0	0	0	0	0	15	137	152

表 4.6 測試語料詞間邊界停頓類別預估分群統計結果

	{B0,B1}	{B2-1,B2-2,B2-3}	{B3,B4}	Total
{B0,B1}	609	157	11	777
{B2-1,B2-2,B2-3}	243	484	60	787
{B3,B4}	8	55	352	415

下表 4.7 為測試語料詞內邊界的預估結果，總預估個數 1795 個，正確個數 1645 個，正確率為 91.64%，若依 4.8 相似歸群統計，總預估正確個數可達到 1785 個，正確率為 99.44%，如預期詞內邊界預估不大會有問題，此正確率已相當高，因訓練語料 B0、B1 以外的類別個數本身就少，所以 B2 那幾個錯誤想要預估正確，應該已經不是 CRF 可解決的，因為在模型訓練過程中為避免過適(overfitting)情形發生，通常會設定一個基準值，若滿足樣版的個數太少就忽略不考慮。

表 4.7 測試語料詞內邊界停頓類別預估結果

	B0	B1	B2-1	B2-2	B2-3	B3	B4	Total
B0	380	49	1	0	0	0	0	430
B1	87	1259	4	0	0	0	0	1350
B2-1	0	1	6	0	0	0	0	7
B2-2	0	0	3	0	0	0	0	3
B2-3	2	2	1	0	0	0	0	5
B3	0	0	0	0	0	0	0	0
B4	0	0	0	0	0	0	0	0

表 4.8 測試語料詞內邊界停頓類別預估分群統計結果

	{B0,B1}	{B2-1,B2-2,B2-3}	{B3,B4}	Total
{B0,B1}	1775	5	0	1780
{B2-1,B2-2,B2-3}	5	10	0	15
{B3,B4}	0	0	0	0

最後合併統計音節邊界的總預估結果，總預估個數為 3774 個，其中正確個數 2797 個，正確率為 74.11%，因為詞內與詞外邊界的停頓類別基本上不重複，所以合併一起看可發現主要問題還是 B2 類別的預估正確率過低，這部分有待未來繼續努力。

4.3 音節韻律狀態預估

由於韻律狀態的變化與音節間停頓類別緊密相關，所以針對韻律狀態的預估，本論文以韻律架構相關參數與樣版為主，因停頓類別可能預估錯誤，加上韻律狀態與部分語言參數也有些許關連，此處為求預估模型的強健性，除韻律架構外又搭配少許相關語言參數訓練 CRF 模型。

4.3.1 韻律狀態預估

本論文需對三類韻律狀態皆進行預估，其中音高與音長韻律狀態，因為與韻律產生直接相關，所以格外重要，本論文只對這兩類樣板與參數選擇特別去調適，能量韻律狀態使用的參數與樣版則僅選用這兩類的交集中應與能量相關的。此處先利用正確的音節停頓標記進行模型訓練，起初嘗試使用三類相同樣版，但增加特徵後，測試結果的變化趨勢並不相同，所以改對三類韻律狀態採用不同樣版與參數。特徵參數的選擇與樣版的調整順序，以相關性應最大的韻律架構相關參數優先考量，接序步驟與停頓類別預估時考慮順序不同，在確定一初版結果後，跟著考慮停頓類別相關的雙連相關樣版(Bigram)，因為由韻律模型標記結果可觀察到，部分停頓類別與韻律狀態的轉移情形明顯相關，故此處優先考慮此類樣板，嘗試增加現在音節後與-1 跟-2 位置的停頓類別，但增加此類樣板後，結果並沒有變好反而有過適的情形發生，推測應是樣板展開後的組合太多，而資料量卻不夠大造成，為使這個樣板能有效使用，我們根據三類韻律狀態的狀態轉移分佈情況，對轉移特性相近的停頓類別歸群，分群後雙連相關樣版即可有效幫助預估，最後為求模型強健性，視情況增加了一些韻律架構相關資訊無法涵蓋到的語言參數，下方所列即最後選擇的樣版與特徵參數。

表 4.9 韻律狀態預估模型的特徵參數

特徵	涵義
Syl_in_PW_n	現在這個音節在 PW(prosodic word)中的位置，分 7 類(1:詞首(第一個音節)，2:五字詞以上非詞首、詞尾的中間部分，3:詞尾(最後一個音節)，4:

	一字詞，32:三字詞的第二音節，42:四字詞的第二音節，43:四字詞的第三音節)
$PWlength_n$	現在這個音節所屬 PW 共有幾個音節，分八類(1,2…,7,8(>7))
$Syl\_in\_PPh\_f_n$	現在這個音節距離最近的前一 PPh(prosodic phrase) boundary 幾個音節 {1^{st}, 2^{nd}, 3^{rd}, [4^{th}, 5^{th}], [6^{th}, 7^{th}], [8^{th}, 11^{th}], others }
$Syl\_in\_PPh\_b_n$	現在這個音節距離最近的後一 PPh boundary 幾個音節 {1^{st}, 2^{nd}, 3^{rd}, [4^{th}, 5^{th}], [6^{th}, 7^{th}], [8^{th}, 11^{th}], others }
$PPhlength_n$	現在這個音節所屬 PPh 共有幾個音節，分五類 {[1,2],[3,4,5,6,7],[8,9,10,11,12],[13,14,15,16,17],others}
$Syl\_in\_BG\_f_n$	現在這個音節距離最近的前一 BG(breath group) boundary 幾個音節 {1^{st}, 2^{nd}, 3^{rd}, [4^{th}, 5^{th}], [6^{th}, 7^{th}], [8^{th}, 11^{th}], others }
$Syl\_in\_BG\_b_n$	現在這個音節距離最近的後一 BG boundary 幾個音節 {1^{st}, 2^{nd}, 3^{rd}, [4^{th}, 5^{th}], [6^{th}, 7^{th}], [8^{th}, 11^{th}], others }
$BGlength_n$	現在這個音節所屬 BG 共有幾個音節，分七類 {[1,2…,7],[8,9…,12],[13,14…,17],[18,19…22],[23,24…,27],[28,29…33],others}
$POS2_n$	詞性(8 類)
$POS3_n$	詞性(23 類)
$Tone_n$	當前音節的聲調(1,2,3,4,5)
$Syl\_in\_LW_n$	現在這個音節在 LW(Lexical word)中的位置，分 7 類(1:詞首(第一個音節)，2:五字詞以上非詞首、詞尾的中間部分，3:詞尾(最後一個音節)，4:一字詞，32:三字詞的第二音節，42:四字詞的第二音節，43:四字詞的第三音節)
$Breaktype_n$	現在這個音節後的 breaktype，分八類(B0,B1,B2-1,B2-2,B2-3,B3,B4, BE)
$Breakclass_n$	對 breaktype 依停頓長短分三類{[B0,B1]、[B2-1,B2-2,B2-3]、[B3,B4]}
$Breakclass\_pps_n$	對 breaktype 依對應 pitch prosody transition 的情形分四類{[B0,B1, B2-3]、

	B2-1、B2-2、[B3,B4]}
Breakclass_qps$_n$	對 breaktype 依對應 duration prosody transition 的情形分四類 {[B0,B1,B2-1]、[B2-2,B3]、B2-3、B4}

表 4.10 音高韻律狀態預估模型的特徵樣版

Template(Unigram)
Breakclass_pps$_n$, Breakclass_pps$_{n-1}$
Syl_in_PW$_n$
PWlength$_n$
Syl_in_PPh_f$_n$, Syl_in_PPh_b$_n$
PPhlength$_n$
Syl_in_BG_f$_n$, Syl_in_BG_b$_n$
BGlength$_n$
Syl_in_LW$_n$
POS2$_n$
Template(Bigram)
Breakclass_pps$_n$

表 4.11 音長韻律狀態預估模型的特徵樣版

Template(Unigram)
Breaktype$_n$, Breaktype$_{n-1}$
Syl_in_PW$_n$
PWlength$_n$
Syl_in_PPh_f$_n$, Syl_in_PPh_b$_n$
PPhlength$_n$
Syl_in_BG_f$_n$, Syl_in_BG_b$_n$

$BGlength_n$
$Syl\_in\_LW_n$
$POS3_n$
$(Tone_n, Tone_{n+1})$
Template(Bigram)
$Breakclass\_qps_n$, $Breakclass\_qps_{n-1}$
$Syl\_in\_PW_n$

表 4.12 能量韻律狀態預估模型的特徵樣版

Template(Unigram)
$Breaktype_n$, $Breaktype_{n-1}$
$Syl\_in\_PW_n$
$PWlength_n$
$Syl\_in\_PPh\_f_n$, $Syl\_in\_PPh\_b_n$
$PPhlength_n$
$Syl\_in\_BG\_f_n$, $Syl\_in\_BG\_b_n$
$BGlength_n$
$Syl\_in\_LW_n$
$POS3_n$
Template(Bigram)
$Breakclass_n$, $Breakclass_{n-1}$

4.3.2 預估結果

根據上方定義之樣版，討論當輸入資料為正確音節停頓類別或預估停頓類別對結果有何影響，所指預估音節停頓為 4.2 節中之預估結果，這邊分別以音節時長與音節音高均方根誤

差(Root Mean Square Error, RMSE)衡量預估之韻律狀態結果。式 4-3 為音節時長均方根誤差的計算公式，d_t 表第 t 個音節原始長度，$\bar{d}_t$ 代表預估得到之第 t 個音節長度，N 為總音節數。式 4-4 則為音節音高之均方根誤差計算公式，a_{ti} 與 $\bar{a}_{ti}$ 分別代表原始與預估第 t 個音節中第 i 維的四維正交係數，此處每個音節以一組勒讓德四維正交係數表示，音節基頻軌跡投影到勒讓德基底的公式如下：

$$a_j = \frac{1}{N+1}\sum_{i=0}^{N} f(i)\cdot\phi_j(\frac{i}{N}) \quad , \quad \text{for } j=0,1,2,3 \tag{4-1}$$

其中，$f(i)$ 為原始的基頻軌跡，$0 \le j \le N$ ；N+1 為基頻軌跡的長度；a_0–a_3 為四維正交係數；$\phi_j(\frac{i}{N})$ 為四維正交的勒讓德多項式基底，分別如下：

$$\phi_0(\frac{i}{N}) = 1$$

$$\phi_1(\frac{i}{N}) = [\frac{12\cdot N}{(N+2)}]^{\frac{1}{2}}\cdot[(\frac{i}{N})-\frac{1}{2}]$$

$$\phi_2(\frac{i}{N}) = [\frac{180\cdot N^3}{(N-1)(N+2)(N+3)}]^{\frac{1}{2}}\cdot[(\frac{i}{N})^2-(\frac{i}{N})+\frac{N-1}{6\cdot N}] \tag{4-2}$$

$$\phi_3(\frac{i}{N}) = [\frac{2800\cdot N^5}{(N-1)(N-2)(N+2)(N+3)(N+4)}]^{\frac{1}{2}}\cdot[(\frac{i}{N})^3-\frac{3}{2}(\frac{i}{N})^2+\frac{6N^2-3N+2}{10N^2}(\frac{i}{N})-\frac{(N-1)(N-2)}{20N^2}]$$

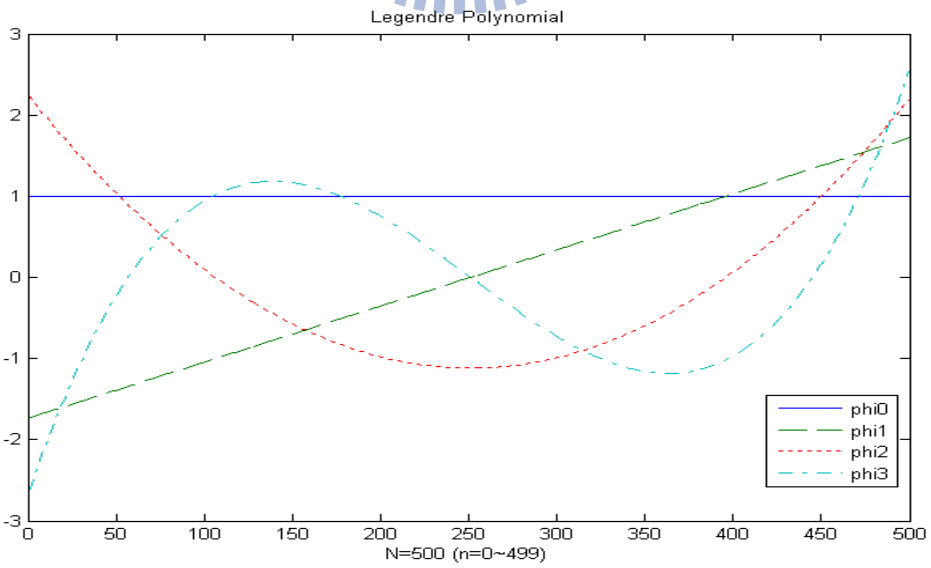

圖 4.4 勒讓德多項式四維正交基底

$$RMSE_{dur} = \sqrt{\frac{\sum_{t}^{N} \left(d_t - \bar{d}_t\right)^2}{N}} \tag{4-3}$$

$$RMSE_{lf0} = \sqrt{\frac{\sum_{t=1}^{N} \sum_{i=1}^{4} \left(a_{ti} - \bar{a}_{ti}\right)^2}{N}} \tag{4-4}$$

表 4.13 音節音長之均方根誤差

	正確韻律標記		正確音節停頓		預估音節停頓	
Duration RMSE (ms)	訓練	測試	訓練	測試	訓練	測試
	10.8667	12.6225	42.1137	40.6476	45.3018	43.3257

表 4.14 音節音高之均方根誤差

	正確韻律標記		正確音節停頓		預估音節停頓	
lF0 RMSE	訓練	測試	訓練	測試	訓練	測試
	0.0718	0.0693	0.1697	0.1584	0.186	0.1726

根據上述結果可發現當音節停頓正確時，確實可有效提升韻律狀態的預估正確率，但即便音節停頓正確預估得之韻律狀態仍與正確的標記結果有一段差距，為了更進一步觀察韻律狀態預估結果，圖 4.4(qps 音長韻律狀態)與圖 4.5(pps 音高韻律狀態)則是模型標記之韻律狀態和給定正確音節停頓(B-est-qps, B-est-pps)預估與使用預估之音節停頓(est-qps, est-pps)預估韻律狀態的比較圖， 圖中標記的停頓類別 correct-Btype 為正確標記結果，est-Btype 為我們方法所預估出 B2 類以上的停頓標記，由兩張圖可看出整體韻律狀態的走勢，基本上符合停頓類別所對應之預期變化，具停頓類別對應之音高重置與音節延長現象，音高的韻律狀態在呼吸群組也呈現一個由高到低的走勢，然而當停頓類別預估錯誤時也會連帶影響韻律狀態的結果，如{放寬的(B2-2)兩岸}預估成{放寬的(B2-3)兩岸}，在圖 4.5 即可觀察到，原先 B2-2 對應音長韻律狀態並不會有明顯音節拉長，當停頓類別預估成 B2-3 時整個走勢就改變了。

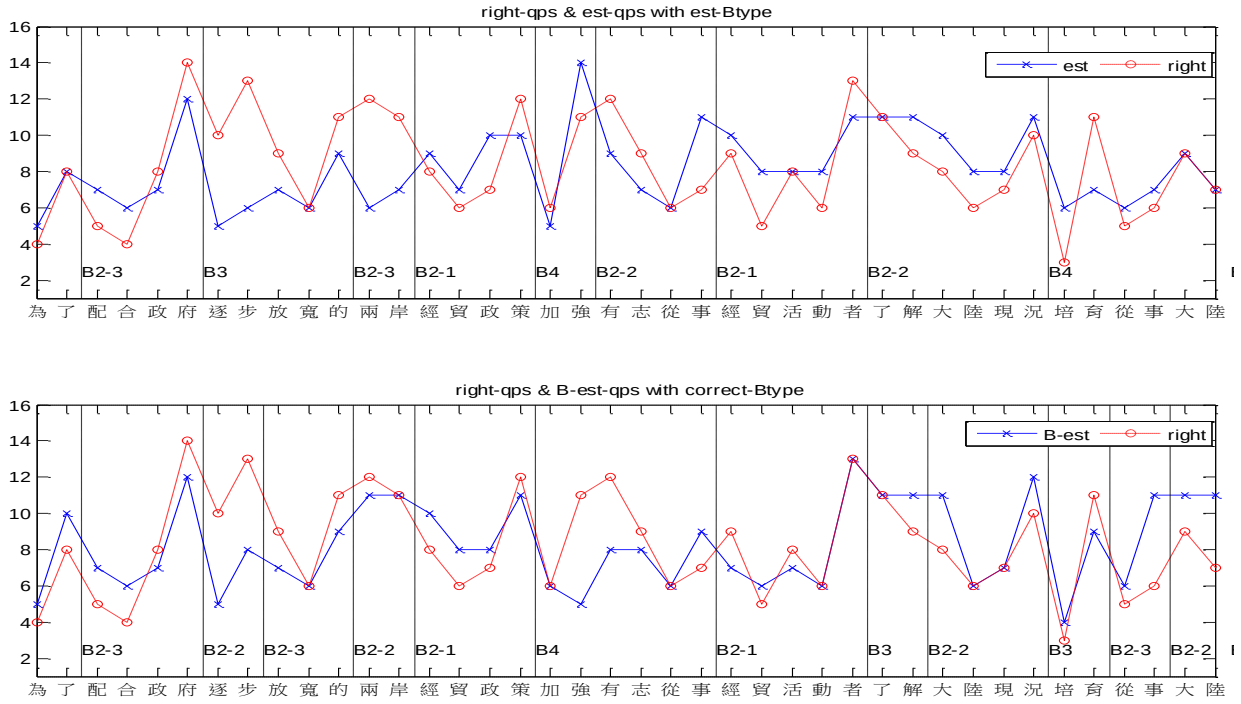

圖 4.5 音長韻律狀態預估結果

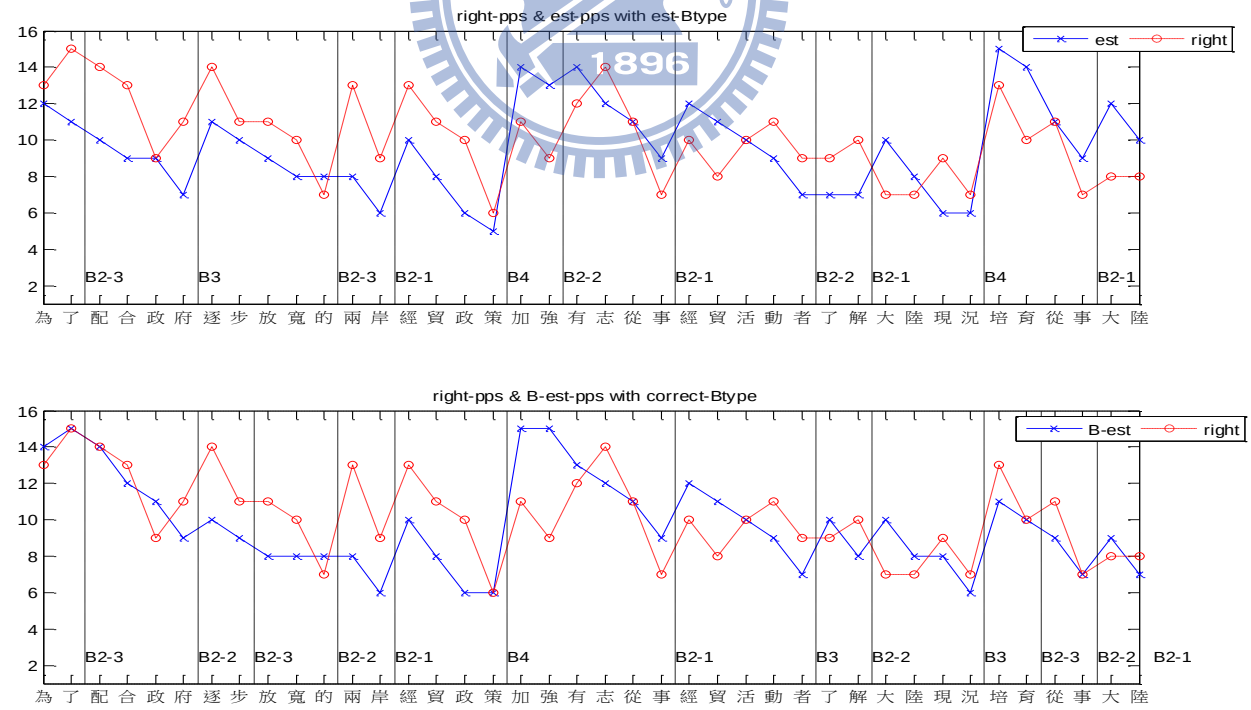

圖 4.6 音高韻律狀態預估結果

4.4 韻律參數產生

參考第三章介紹之韻律模型中的音節韻律模型 $P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})$，此模型可分解出三個子模型，分別模擬音節基頻軌跡序列 **sp**、音長序列 **sd** 和音節能量序列 **se**，本研究便利用音節基頻與音長模型的平均值來代表音節的基頻與音長值，故可得到下面兩個式子：

$$\mathbf{sp}_n^* = \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n^*} + \boldsymbol{\beta}^{f}_{B_{n-1}^*, tp_{n-1}} + \boldsymbol{\beta}^{b}_{B_n^*, tp_n} + \boldsymbol{\mu} \tag{4-5}$$

$$sd_n^* = \gamma_{t_n} + \gamma_{s_n} + \gamma_{q_n^*} + \mu_d \tag{4-6}$$

其中 $\mathbf{sp}_n^*$ 與 sd_n^* 分別代表預估的音節基頻與音長，其中式 4-6 與式 3-9 相比，可發現少了一句子的影響因素，這邊省略句子的影響因素，因為訓練時此項目的為正規化不同句子的影響，合成的句子通常與訓練時不相同，故無法直接從模型中得到這個 AP。

根據 4.2 與 4.3 兩節介紹的方法，可分別預估得到音節間的停頓類別與音節的韻律狀態，再加上輸入的文字資訊，便可得到式 4-5 與式 4-6 還原需要的所有影響因素，藉此可從訓練得到的韻律模型中找到對應的各個 APs，根據上面兩個式子，對相關的 APs 做疊加，即可得到每個音節的基頻軌跡(四維正交係數)與音節音長。

得到音節音長後，可利用式 3-22~24，搭配狀態持續時間模型預估出狀態音長；而音節四維正交的基頻軌跡，可利用式 4-7 的方法對其還原回音節基頻值，然而此公式需要音節中有聲部分的長度 *N+1*，因為本論文所提方法並沒有以 MSD 方法訓練音高模型，故得要使用其他方式判定音節中的有聲部分，第一版的判定是根據語言學知識，如同第三章中計算 MCD 時的判定方法，但以此版本還原的音節基頻軌跡做聲音合成，會發現在有聲與無聲交界處附近，部分音節會頻繁出現與頻譜不一致導致聲音破掉或啞掉的情形。為了解決這個問題，我們利用頻譜模型強制切割得到之訓練語料切割位置，蒐集每個 HMM 對應的 F0 序列，接著依據頻譜決策樹對每個 HMM 所對應的 F0 分群，最後統計決策樹每個節點中 F0 非零的音框數，定義非零值的比例超過一基準值(本論文設定為 0.5)，此節點的模型狀態即視為有聲，藉此判定方法找到與頻譜模型一致的基頻軌跡值，根據合成結果可發現，此方法確實有效解決

了不少頻譜與基頻軌跡有聲、無聲部分不一致的問題。

$$f(i)=\sum_{j=0}^{3} a_j \cdot \phi_j\left(\frac{i}{N}\right) \quad \text{，for} \quad 0 \le i \le N \tag{4-7}$$

本研究提出之合成方法，另一需要的韻律資訊為音節間停頓時長，表 4.15 為音節間靜音停頓(Pause duration)模型，各類靜音停頓對應的的平均停頓時長，可發現其中除了 B2-2、B3、B4 三類的平均停頓時長明顯較長外，其他停頓類別對應的平均停頓時長，皆為一極短，聽覺上幾乎無法察覺之長度，因此在本研究中僅針對停頓類別預估為 B2-2、B3、B4 三類的音節邊界給定靜音停頓，韻律模型中定義靜音停頓模型為一 Gamma 分佈，但本論文為實作上的方便，直接以此類停頓的平均時長代表所有這三個停頓類別對應位置的停頓時長。

表 4.15 音節間靜音停頓模型的各類停頓平均時長

停頓類別	B0	B1	B2-1	B2-2	B2-3	B3	B4
長度(ms)	3	11	18	109	16	287	543

第五章 合成系統實作與評估

綜合第三章以韻律資訊幫助訓練得到之文脈相關 HSMM，與第四章介紹的韻律資訊產生器，搭配前級的文字分析器即一完整合成系統，本論文便以此系統與 2.2 節中介紹的傳統 HSMM-based 合成系統比較。

5.1 實驗介紹

此章節結合系統訓練與合成部分，對完整合成器的成果進行評估，其中合成系統中前級的文字分析器，因不在本研究範疇內，故此處不作探討，並假設所有需要之文本資訊皆可透過文字分析器取得。

本論文使用的傳統 HSMM-based 合成系統，分別考慮了音素與聲、韻母兩種單元，HMM 的狀態數分別設定為 3 個與 5 個狀態；至於模型結構，因模型訓練不只考慮第三章中實作的頻譜部分，為避免影響整體系統表現，所以此處參考過去作法，三棵決策樹皆採用相同問題集，模型結構也依照原始設定，三個模型皆為狀態結構(state-based)。另外，傳統 HTS 雖然在模型訓練階段，可以根據音檔實際停頓時長(本論文設定>25ms)給定靜音停頓標記，一併訓練靜音停頓模型，但合成階段卻無法以同樣方式獲得此類標記，尤其在長句的應用，合成器必定得要有合適的斷句停頓，最簡單的做法即直接對文本標記之標點符號位置安插靜音停頓，本論文也採取此種方法標記停頓位置，進行系統實作評估，對此兩種單元之合成系統分別以 HTS-Phone 與 HTS-IF 稱之。

本研究提出的合成方法，同樣考慮上述兩種單元，此合成系統基於韻律模型發展而成，合成階段同樣面臨沒有音檔無法得到韻律標記的情形，故在合成器前級加入了第四章中介紹之以文字預估韻律標記的方法，由此結合韻律模型與第三章所介紹之文脈相關模型，即可完成圖 2.5 所示之完整合成系統，兩種單元對應系統分別稱為 PLM-Phone 與 PLM-IF，由於韻律標記正確與否可能會影響整體結果，故另外實作一給定正確韻律標記的版本(僅考慮聲、

韻母單元)，此方法稱 PLM-Correct，藉此衡量韻律標記的正確與否對合成結果影響程度，也可知道若韻律標記預估結果完全正確，合成器所能得到之最好結果。

上述所有方法使用的訓練語料與測試語料皆如同前面章節設定，本研究同時採用客觀實驗評估與主觀實驗評估評量系統優劣，接下來將分別對兩種評估方法做詳細介紹與結果分析。

5.2 客觀實驗評估

客觀評估對頻譜計算 MCD，音長與音高則利用 RMSE 統計，公式皆如同前面章節，除頻譜 MCD 因對整體訓練語料運算時間過久，僅隨機選取其中 50 句計算外，其餘皆對整體訓練與測試語料計算結果。

首先是頻譜 MCD 的計算，與第三章相同根據手動修正過之聲、韻母切割位置當標準答案，所有韻母與聲母中的 m、n、l、r、INULL 判定為有聲，音素單元則以韻母對應之所有音素與聲母對應音素中的 m、n、l、r 判定為有聲。加入考慮全域變異數(Global Variance, GV)的方法【4】後，為使 variance 變大，新的軌跡(trajectory)可能會因此偏離原始軌跡，導致 MCD 變大許多，但聲音卻反而會更加清晰，GV 屬於一種後期加強解決軌跡過度平滑(Over-smoothing)問題的方法，做法是對初始生成之參數軌跡拉大動態範圍(Dynamic range)，GV 對高維度參數的影響較大，其並不會改變原始軌跡走勢，只是拉大軌跡變化的幅度，但如此可能會使原本小錯的地方變成大錯，而原先大錯的地方，有可能並沒有受到 GV 放大影響，導致前者計算結果與後者相近，但此並非頻譜模型造成的錯誤，所以若目標為衡量頻譜模型參數生成結果與原始軌跡差距，以初始生成之軌跡計算 MCD 應較為合理。

下方表 5.1 為加 GV 前的 MCD 計算結果，比較 PLM-Correct 與 PLM-IF 可看出當使用預估之韻律標記作參數還原，會使結果大幅變差，但 PLM-IF 仍勝過 PLM-Phone，此結果與第三章一致，然而我們所提方法卻都不如傳統方法，推論是因為韻律標記預估錯誤導致單元選擇與原先不一致造成，因我們方法每一群資料較集中，相當於群與群間差異變大，故選錯單元就容易造成 MCD 變大，但如此與 GV 原理可能相同，因為參數的動態範圍本身變大，所

以才造成 MCD 大的現象，因此誤差大不代表聲音就會比較差。圖 5.1 即為此處 5 種比較方法的頻譜 GV 計算結果，包括能量共 25 維，可看出我們的方法確實變異較大，雖然不如 GV 方法的效應明顯，但確實稍微降低了 Over-smoothing 的情形，而其中 PLM-Correct 不只 GV 最大且 MCD 也最小，表示其軌跡最像原始軌跡且參數動態變化範圍又大，有很大機會可以合成出與原始音檔相似品質又好的音檔。

表 5.1 整體語料之 MCD 值

	Training		Testing	
	Voiced	Unvoiced	Voiced	Unvoiced
HTS-Phone	5.0519	4.5856	5.1730	4.7308
HTS-IF	5.0364	4.5495	5.1947	4.6977
PLM-Phone	5.1631	4.5607	5.2473	4.6819
PLM-IF	5.1020	4.5117	5.2196	4.6651
PLM-Correct	4.8336	4.4545	5.0487	4.6469

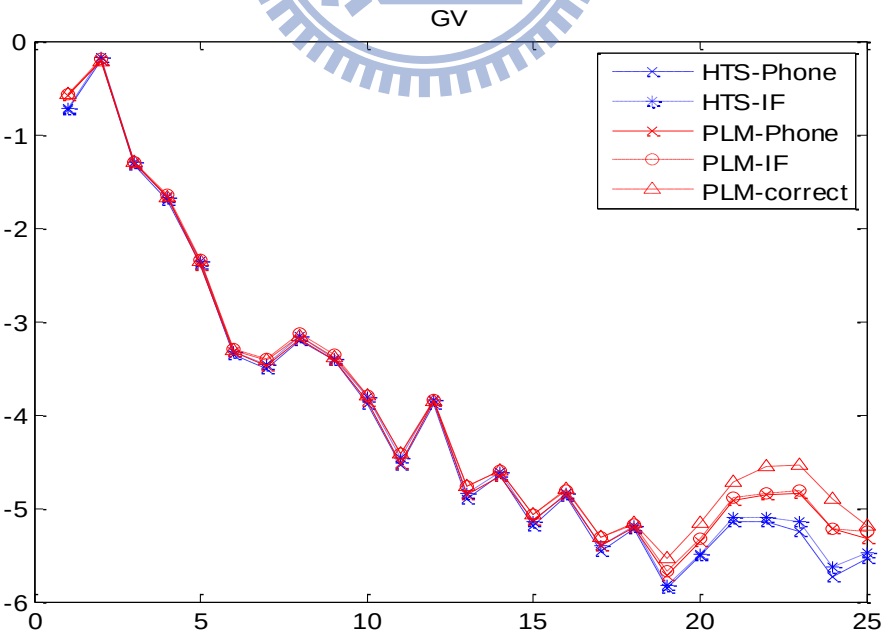

圖 5.1 測試語料五種方法之 log global variance 比較圖

接著計算 lF0 的 RMSE，因為每個音節中有聲部分的長度不同，此處將每個音節中有聲部分的對數基頻值(logF0)投影到四維正交勒讓德多項式基底上，對每個音節的基頻軌跡以一組四維正交係數表示，透過計算此四維正交係數的 RMSE 比較兩種單元與兩種方法的優劣，因我們所提兩種方法，當韻律標記完全一致的情況下結果會是相同的，所以這邊僅列出一組。比較測試語料的結果可發現我們方法小輸給 HTS-Phone 然而贏過 HTS-IF，而 HTS-Phone 雖然 RMSE 最小，但是它卻有少數音節並沒有預估出 F0，這個問題是我們方法所不會有的，若將沒預估到的音節也一併考慮進去計算，其 RMSE 絕對會大幅上升，故可證明我們方法仍是有優勢在的。

表 5.2 整體語料之 lF0 RMSE

lF0 RMSE(dB)	Training	Testing
HTS-Phone	0.1651	0.1726
HTS-IF	0.17	0.1815
PLM-Phone & PLM-IF	0.1862	0.1733

最後比較預估音節長短，觀察結果可發現對訓練語料計算我們方法最好，而測試語料我們則排第二，小輸 HTS-IF，目前我們方法合成的語速偏快，猜測可能是因為我們在合成階段，以韻律模型還原音節音長時，少考慮句子的正規化 AP 造成，若能針對此做改善，音節音長的預估即有機會勝過傳統方法。

表 5.3 整體語料之音節音長 RMSE

Duration RMSE (ms)	Training	Testing
HTS-Phone	47.3255	45.1842
HTS-IF	45.3939	42.712
PLM-Phone & PLM-IF	45.3018	43.3257

5.3 主觀實驗評估

主觀實驗總共進行兩個實驗，一為偏好測定(Preference test)，一為平均主觀值分數(Mean Option Source, MOS)。

5.3.1 偏好測定

偏好側定依據合成方法與合成單元產生四組配對，分別為(HTS-Phone, HTS-IF)、(HTS-Phone, PLM-Phone)、(PLM-Phone, PLM-IF)、(HTS-IF, PLM-IF)，並從測試語料中隨機選取 10 個相同句子進行四組測試，由 15 位受試者進行偏好測定實驗，每組編號相同音檔代表內容相同，採隨機排序方式撥放，由受試者從相同內容的兩句中挑選比較喜歡的一句，並設定一個選項為結果相近比較不出好壞，測試結果如下：

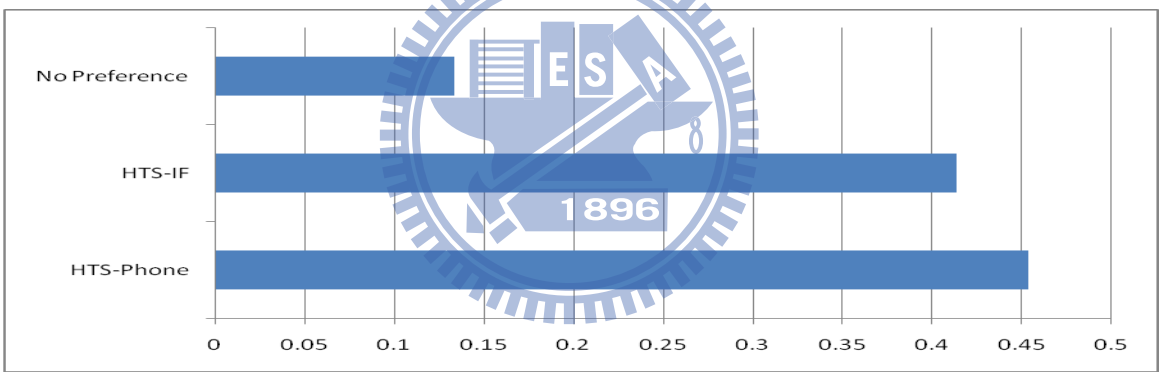

圖 5.2 (HTS-Phone, HTS-IF)偏好測定評估結果

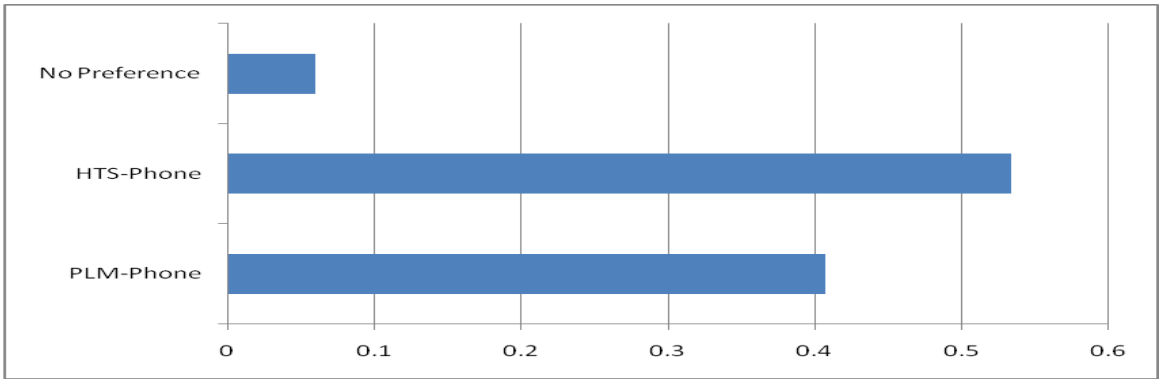

圖 5.3 (HTS-Phone, PLM-Phone)偏好測定評估結果

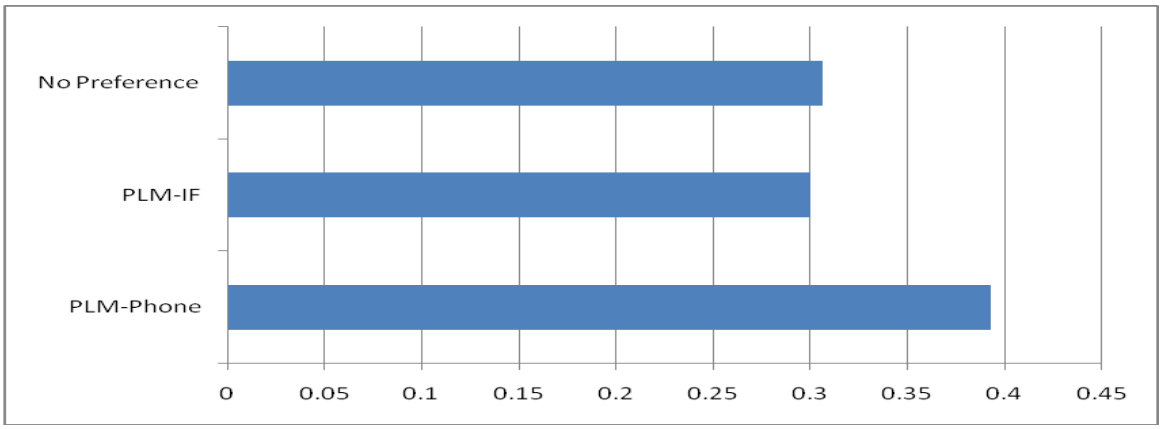

圖 5.4 (PLM-Phone, PLM-IF)偏好測定評估結果

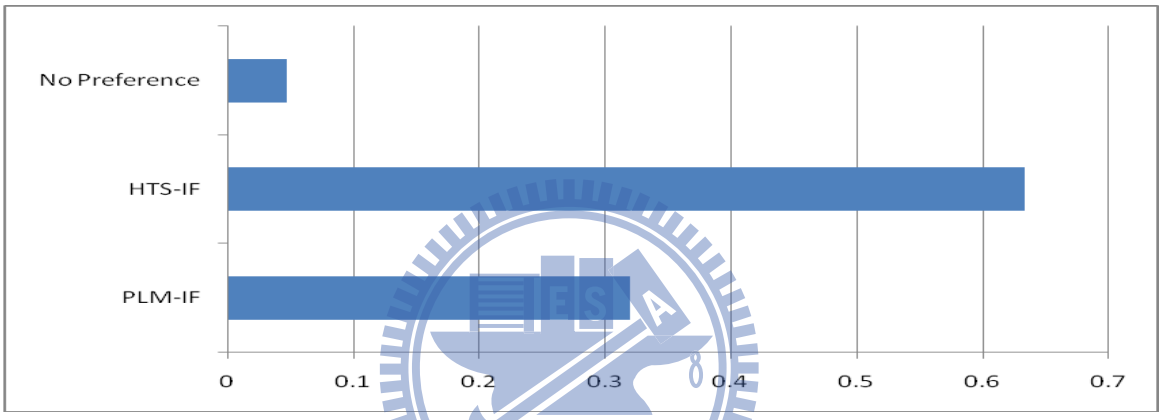

圖 5.5 (HTS-IF, PLM-IF)偏好測定評估結果

此結果與預期大不相同，本論文所提方法不論何種單元，結果均不如傳統方法。傳統方法只在標點符號處標記靜音停頓，原本預期部分句子會因標點符號間距過遠，造成傳統方法韻律變化過度平滑的特性更加明顯，使聲音聽起來呆版，然而本論文所提方法，因能提供豐富韻律變化，在合理的韻律預估結果下，應可使合成語音聽起來更為自然，即便長句也不會讓受試者感覺煩悶，然而結果卻完全相反，數據呈現受試者似乎對傳統方法合成的句子較為喜歡，分析每個語者的回答情況，可發現有五個受試者，相當一致的對本論文所提方法幾乎完全不喜歡，10 句中至少有 8 句選擇 HTS 的方法，詢問過這幾個受試者，發現此群受試者判定音檔偏好相似，多以停頓點正確與否當作衡量的主要準則，且對停頓點的容錯度相當低，只要一發現停頓不合理，必定選擇另一方法，如此因為傳統方法皆斷在標點符號位置，故幾乎不會出現不合理的情況，而我們提出的方法，停頓位置是依據預估的停頓類別來判定，以

目前 CRF 停頓模型的預估能力，仍無法避免此類問題發生，不恰當的靜音停頓會造成語流上的中斷，進而影響受試者對音檔好壞的評斷，另外原本靜音停頓時間是以 Gamma 分佈對它模型化，本論文中僅以靜音停頓模型各類停頓對應的期望值表示同一停頓類別的靜音停頓時長，此也可能會使部分句中停頓變成過長或過短，且因句子太長的緣故，一些音檔上的細節易被忽略，受試者往往只會記得停頓過長這類明顯錯誤。除了停頓不當問題外，有些人也提出部分語句速度過快問題，此部分問題發生原因還有待探討，不過初步推斷應該是音長韻律狀態預估出來偏低造成的。

除了上述缺點外，受試者大多也覺得，本論文所提方法在語句中的韻律變化確實較為豐富，若扣除停頓明顯錯誤的地方外，韻律變化基本上算是相當不錯的；且若更細微的去注意聲音變化應可發現，PLM-IF 方法在聲調的表現上是相當不錯的，因為我們方法有考慮 tone 的影響因素，因此在聲調上較不易出錯，且 PLM-IF 有效解決了 PLM-Phone 音素長短不穩定的缺點，大部分的音檔聽起來品質都相當不錯，不易發生字好像只念一半或者一個字中間聲音啞掉的情形。

此外比較圖 5.4 的結果，可發現當兩個方法韻律條件相同時，確實多數人是無法分辨出頻譜差別的，驗證了我們第三章主觀實驗中的論點。而根據研究結果發現 PLM-IF 與 PLM-Phone 相比，主要的明顯差別應該只在音節聲音完整度的部分，PLM-Phone 的狀態時長預估相對來說較不穩定，因此常會有韻母過短導致聲音急促消失的情形，但此現象在長句中容易被忽略，這組實驗如果重新排序重做一次結果應該又會不同。

5.3.2 平均主觀值分數

下圖為 MOS 的結果，評分標準可參考表 5.4，MOS 的結果一樣我們方法並不如傳統方法，其中 PLM-Phone 與 PLM-IF 更是明顯輸給傳統方法，主要原因與上一小節偏好實驗相同，所以就不重複討論，至於給定正確韻律狀態與停頓類型的 PLM-Correct 卻仍不如傳統方法，推測是由於我們合成的聲音能量變化較大，因此即便速度合理、停頓恰當，聽起來相較傳統方法，仍易覺得聲音較不穩定，因為能量在整個訓練過程並沒有特別考慮，而韻律模型中其

實有音節能量這項可以運用，未來或許可以考慮改以韻律模型產生，取代 MGC 中第一維之能量訓練方式，如此應有機會可以解決目前能量不穩定的問題。

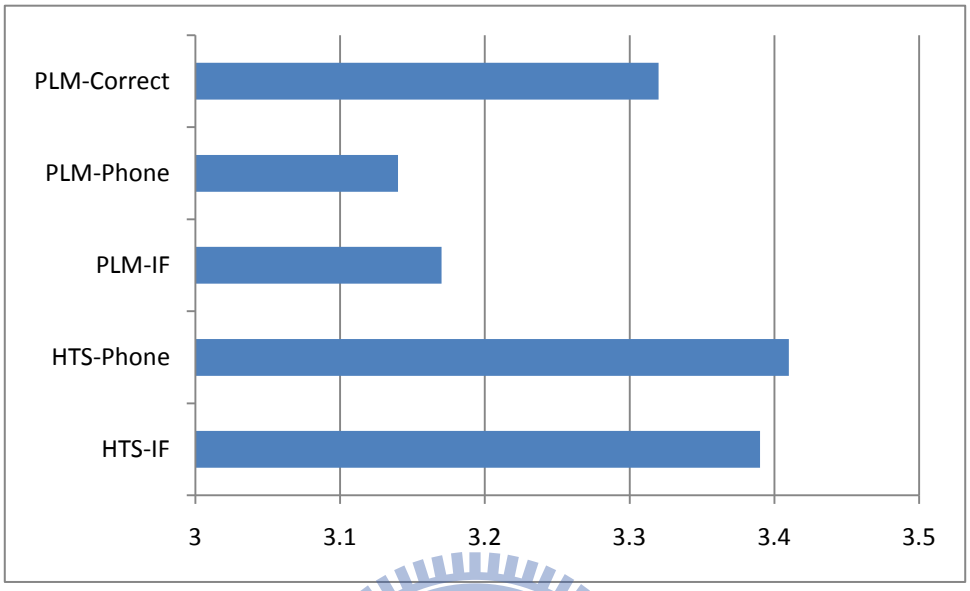

圖 5.6 五種方法之 MOS 結果

表 5.4 MOS 評分標準

評等	分數	說明
優	5	合成語音非常自然
良	4	合成語音自然
可	3	合成語音自然度表現尚可
差	2	合成語音不太自然
劣	1	合成語音非常不自然

第六章 結論與未來展望

本論文結合階層式韻律模型提出之合成架構，希望可以進一步改善傳統 HMM-based 合成器的缺點，藉由韻律標記引入幫助頻譜模型訓練，由決策樹分群部分著手解決參數軌跡過度平滑的缺點，由此提升頻譜方面的表現；並搭配韻律預估，直接以韻律模型產生豐富韻律特性的韻律參數，一次改善傳統方法頻譜與韻律上的問題，提升合成器的聲音品質

頻譜的部分，由客觀實驗數據上看來，我們所提方法應是有效的，雖不如 GV 方法的明顯增益，但確實改進了頻譜模型，但搭配 CRF 韻律預估與韻律模型完成之合成器結果確不如預期，由最後主觀實驗的結果可發現，多數人對於合成語音的第一要求是聲音穩定，本論文所提之方法雖然在韻律的變化上大大勝過傳統方法，但過於明顯的錯誤卻會把優點給掩蓋，HTS 具有聲音缺乏變化的缺點，卻也因此聲音較為穩定不易出現明顯大錯，而受到測試者的喜愛，綜觀各項客觀實驗結果可發現，若能解決穩定性不足的問題，我們所提之方法仍然是相當具有競爭力的。

目前發現我們方法主要有能量不穩、速度偏快與停頓錯誤這三類問題，其中停頓錯誤最直接與人類聽感相關，應是未來最迫切得要解決的，若能有效解決，不僅可以排除目前聽覺實驗中我們方法最大的缺點，更有機會進一步提升韻律狀態預估的結果，或許沒有辦法做到百分之百正確，但把明顯的錯誤給解決掉應該是辦得到的，若能做到這點，相信此方法所合成之聲音自然度絕對是勝過傳統方法的。

參考文獻

【1】 T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, T. Kitamura, “Speaker interpolation for HMM-based speech synthesis system,” *J. Acoust. Soc. Jpn*. (E), vol.21, no.4, pp.199-206, 2000

【2】 M. Tamura, T. Masuko, K. Tokuda, T. Kobayashi, “Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR,” *Proc of ICASSP*, pp.805-808, May 2001

【3】 K. Tokuda, T. Mausko, N. Miyazaki, and T. Kobayashi, “Multispace probability distribution HMM,” *IEICE Trans. Inf. &Syst.*, vol. E85-D, no. 3, pp. 455–464, 2002.

【4】 Tomoki Toda, Keiichi Tokuda, “A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis,” *IEICE Trans. Inf. &Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

【5】 Yi-Jian Wu, Ren-Hua Wang, “MINIMUM GENERATION ERROR TRAINING FOR HMM-BASED SPEECH SYNTHESIS,” in *Proc. ICASSP*, Toulouse, May 2006

【6】 Heiga Zen, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, Tadashi Kitamura, “Hidden Semi-Markov Model Based Speech Synthesis,” in *Proc. ICSLP*, 2004

【7】 C. Y. Chiang,“Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech,” Department of Communication Engineering, NCTU, Dissertation for Doctor of Philosophy, March 2009.

【8】 Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. C., The HTK Book, version 3.4. Cambridge University Engineering Department, Cambridge, UK. 2006.

【9】 K. Sjlander and J. Beskow, “Wavesurfer - an open source speech tool,” in *Proceeding of the ICSLP* 2000, Vol. 4, pp. 464-467.

【10】 C. Bennett, "*Large scale evaluation of corpus-based synthesizers: Results and lessons*

from the Blizzard Challenge 2005," in *Interspeech*, 2005, pp. 105-108

【11】 Zen, H., Nose, T., Yamagishi, J., Sako, S. and Tokuda, K., The HMM-based Speech System(HTS) Version 2.1,2007,http://hts.sp.nitech.ac.jp/

【12】 T. Yoshimura, "Simulations Modeling of Phonetic and Prosodic Parameters, and Characteristic Conversion for HMM-based Text-to-Speech Systems," Department of Electrical and Computer Engineering Nagoya Institute of Technology, 2002

【13】 Z. Sheng, J.-H. Tao, and D.-L. Jiang, "Chinese prosodic phrasing with extended features," *Proceedings of the IEEE ICASSP* ,Vol. 1, pp. 492–495. 2003

【14】 C.-Y. Tseng, S.-H. Pin, Y.-L. Lee, H.-M. Wang, and Y.-C. Chen, "Fluent speech prosody: Framework and modeling,"*Speech Commun*. special issue on quantitative prosody modeling for natural speech description and generation, 46, 284–309 (2005).

附錄一

子音編號(22類)	子音拼音(22類)	注音	母音編號	母音符號(40類)	注音	38類音素
1	INULL		1	FNULL1	Φ1	FNULL1
2	zh	ㄓ	2	a	ㄚ	FNULL2
3	ch	ㄔ	3	o	ㄛ	a
4	sh	ㄕ	4	e	ㄜ	b
5	r	ㄖ	5	eh	ㄝ	ch
6	z	ㄗ	6	ai	ㄞ	c
7	c	ㄘ	7	ei	ㄟ	d
8	s	ㄙ	8	ao	ㄠ	en
9	g	ㄍ	9	ou	ㄡ	er
10	k	ㄎ	10	an	ㄢ	e
11	h	ㄏ	11	en	ㄣ	eh
12	j	ㄐ	12	ang	ㄤ	f
13	q	ㄑ	13	eng	ㄥ	g
14	x	ㄒ	14	yi	ㄧ	h
15	d	ㄉ	15	wu	ㄨ	j
16	t	ㄊ	16	yu	ㄩ	k
17	n	ㄋ	17	ya	ㄧㄚ	l
18	l	ㄌ	18	ye	ㄧㄝ	m
19	b	ㄅ	19	yai	ㄧㄞ	ng
20	p	ㄆ	20	yao	ㄧㄠ	n
21	m	ㄇ	21	you	ㄧㄡ	o
22	f	ㄈ	22	yan	ㄧㄢ	p
			23	yin	ㄧㄣ	q
			24	yang	ㄧㄤ	r
			25	ying	ㄧㄥ	s
			26	wa	ㄨㄚ	sh
			27	wo	ㄨㄛ	t
			28	wai	ㄨㄞ	wu1
			29	wei	ㄨㄟ	wu2

30	wan	ㄨㄢ	wu3
31	wen	ㄨㄣ	x
32	wang	ㄨㄤ	yi1
33	weng	ㄨㄥ	yi2
34	yue	ㄩㄝ	yi3
35	yuan	ㄩㄢ	yu1
36	yun	ㄩㄣ	yu2
37	yung	ㄩㄥ	zh
38	er	ㄦ	z
39	yo	ㄧㄛ	
40	FNULL2	Φ2	

附錄二

決策樹韻律相關問題集

QS "L=B0" {*/pb:B0/*}

QS "L=B1" {*/pb:B1/*}

QS "L=B21" {*/pb:B21/*}

QS "L=B22" {*/pb:B22/*}

QS "L=B23" {*/pb:B23/*}

QS "L=B3" {*/pb:B3/*}

QS "L=B4" {*/pb:B4/*}

QS "R=B0" {*/nb:B0/*}

QS "R=B1" {*/nb:B1/*}

QS "R=B21" {*/nb:B21/*}

QS "R=B22" {*/nb:B22/*}

QS "R=B23" {*/nb:B23/*}

QS "R=B3" {*/nb:B3/*}

QS "R=B4" {*/nb:B4/*}

QS "L=Bclass1" {*/pb:B0/*,*/pb:B1/*}

QS "L=Bclass2" {*/pb:B21/*,*/pb:B22/*,*/pb:B23/*}

QS "L=Bclass3" {*/pb:B22/*,*/pb:B3/*,*/pb:B4/*}

QS "L=Bclass4" {*/pb:B3/*,*/pb:B4/*}

QS "R=Bclass1" {*/nb:B0,*/nb:B1}

QS "R=Bclass2" {*/nb:B21,*/nb:B22,*/nb:B23}

QS "R=Bclass3" {*/nb:B22,*/nb:B3,*/nb:B4}

QS "R=Bclass4" {*/nb:B3,*/nb:B4}

QS "PitchPS1" {*/p:1/*}

QS "PitchPS1to2" {*/p:1/*,*/p:2/*}

QS "PitchPS1to3" {*/p:1/*,*/p:2/*,*/p:3/*}

QS "PitchPS1to4" {*/p:1/*,*/p:2/*,*/p:3/*,*/p:4/*}

QS "PitchPS1to5" {*/p:1/*,*/p:2/*,*/p:3/*,*/p:4/*,*/p:5/*}

QS "PitchPS16" {*/p:16/*}

QS "PitchPS16to15" {*/p:16/*,*/p:15/*}

QS "PitchPS16to14" {*/p:16/*,*/p:15/*,*/p:14/*}

QS "PitchPS16to13" {*/p:16/*,*/p:15/*,*/p:14/*,*/p:13/*}

QS "PitchPS16to12" {*/p:16/*,*/p:15/*,*/p:14/*,*/p:13/*,*/p:12/*}

QS "PitchPS3to6" {*/p:3/*,*/p:4/*,*/p:5/*,*/p:6/*}

QS "PitchPS4to7" {*/p:4/*,*/p:5/*,*/p:6/*,*/p:7/*}

QS "PitchPS5to8" {*/p:5/*,*/p:6/*,*/p:7/*,*/p:8/*}

QS "PitchPS6to9" {*/p:6/*,*/p:7/*,*/p:8/*,*/p:9/*}

QS "PitchPS7to10" {*/p:7/*,*/p:8/*,*/p:9/*,*/p:10/*}

QS "PitchPS8to11" {*/p:8/*,*/p:9/*,*/p:10/*,*/p:11/*}

QS "PitchPS9to12" {*/p:9/*,*/p:10/*,*/p:11/*,*/p:12/*}

QS "PitchPS10to13" {*/p:10/*,*/p:11/*,*/p:12/*,*/p:13/*}

QS "PitchPS11to14" {*/p:11/*,*/p:12/*,*/p:13/*,*/p:14/*}

QS "PitchPS3to8" {*/p:3/*,*/p:4/*,*/p:5/*,*/p:6/*,*/p:7/*,*/p:8/*}

QS "PitchPS6to11" {*/p:6/*,*/p:7/*,*/p:8/*,*/p:9/*,*/p:10/*,*/p:11/*}

QS "PitchPS9to14" {*/p:9/*,*/p:10/*,*/p:11/*,*/p:12/*,*/p:13/*,*/p:14/*}

QS"PitchPS3to14"{*/p:3/*,*/p:4/*,*/p:5/*,*/p:6/*,*/p:7/*,*/p:8/*,*/p:9/*,*/p:10/*,*/p:11/*,*/p:12/*,*/p:13/*,*/p:14/*}

QS "DurPS1" {*/q:1/*}

QS "DurPS1to2" {*/q:1/*,*/q:2/*}

QS "DurPS1to3" {*/q:1/*,*/q:2/*,*/q:3/*}

QS "DurPS1to4" {*/q:1/*,*/q:2/*,*/q:3/*,*/q:4/*}

QS "DurPS1to5" {*/q:1/*,*/q:2/*,*/q:3/*,*/q:4/*,*/q:5/*}

QS "DurPS16" {*/q:16/*}

QS "DurPS16to15" {*/q:16/*,*/q:15/*}

QS "DurPS16to14" {*/q:16/*,*/q:15/*,*/q:14/*}

QS "DurPS16to13" {*/q:16/*,*/q:15/*,*/q:14/*,*/q:13/*}

QS "DurPS16to12" {*/q:16/*,*/q:15/*,*/q:14/*,*/q:13/*,*/q:12/*}

QS "DurPS3to6" {*/q:3/*,*/q:4/*,*/q:5/*,*/q:6/*}

QS "DurPS4to7" {*/q:4/*,*/q:5/*,*/q:6/*,*/q:7/*}

QS "DurPS5to8" {*/q:5/*,*/q:6/*,*/q:7/*,*/q:8/*}

QS "DurPS6to9" {*/q:6/*,*/q:7/*,*/q:8/*,*/q:9/*}

QS "DurPS7to10" {*/q:7/*,*/q:8/*,*/q:9/*,*/q:10/*}

QS "DurPS8to11" {*/q:8/*,*/q:9/*,*/q:10/*,*/q:11/*}

QS "DurPS9to12" {*/q:9/*,*/q:10/*,*/q:11/*,*/q:12/*}

QS "DurPS10to13" {*/q:10/*,*/q:11/*,*/q:12/*,*/q:13/*}

QS "DurPS11to14" {*/q:11/*,*/q:12/*,*/q:13/*,*/q:14/*}

QS "DurPS3to8" {*/q:3/*,*/q:4/*,*/q:5/*,*/q:6/*,*/q:7/*,*/q:8/*}

QS "DurPS6to11" {*/q:6/*,*/q:7/*,*/q:8/*,*/q:9/*,*/q:10/*,*/q:11/*}

QS "DurPS9to14" {*/q:9/*,*/q:10/*,*/q:11/*,*/q:12/*,*/q:13/*,*/q:14/*}

QS"DurPS3to14"{*/q:3/*,*/q:4/*,*/q:5/*,*/q:6/*,*/q:7/*,*/q:8/*,*/q:9/*,*/q:10/*,*/q:11/*,*/q:12/*
,*/q:13/*,*/q:14/*}

QS "EnePS1" {*/r:1/*}

QS "EnePS1to2" {*/r:1/*,*/r:2/*}

QS "EnePS1to3" {*/r:1/*,*/r:2/*,*/r:3/*}

QS "EnePS1to4" {*/r:1/*,*/r:2/*,*/r:3/*,*/r:4/*}

QS "EnePS1to5" {*/r:1/*,*/r:2/*,*/r:3/*,*/r:4/*,*/r:5/*}

QS "EnePS16" {*/r:16/*}

QS "EnePS16to15" {*/r:16/*,*/r:15/*}

QS "EnePS16to14" {*/r:16/*,*/r:15/*,*/r:14/*}

QS "EnePS16to13" {*/r:16/*,*/r:15/*,*/r:14/*,*/r:13/*}

QS "EnePS16to12" {*/r:16/*,*/r:15/*,*/r:14/*,*/r:13/*,*/r:12/*}

QS "EnePS3to6" {*/r:3/*,*/r:4/*,*/r:5/*,*/r:6/*}

QS "EnePS4to7" {*/r:4/*,*/r:5/*,*/r:6/*,*/r:7/*}

QS "EnePS5to8" {*/r:5/*,*/r:6/*,*/r:7/*,*/r:8/*}

QS "EnePS6to9" {*/r:6/*,*/r:7/*,*/r:8/*,*/r:9/*}

QS "EnePS7to10" {*/r:7/*,*/r:8/*,*/r:9/*,*/r:10/*}

QS "EnePS8to11" {*/r:8/*,*/r:9/*,*/r:10/*,*/r:11/*}

QS "EnePS9to12" {*/r:9/*,*/r:10/*,*/r:11/*,*/r:12/*}

QS "EnePS10to13" {*/r:10/*,*/r:11/*,*/r:12/*,*/r:13/*}

QS "EnePS11to14" {*/r:11/*,*/r:12/*,*/r:13/*,*/r:14/*}

QS "EnePS3to8" {*/r:3/*,*/r:4/*,*/r:5/*,*/r:6/*,*/r:7/*,*/r:8/*}

QS "EnePS6to11" {*/r:6/*,*/r:7/*,*/r:8/*,*/r:9/*,*/r:10/*,*/r:11/*}

QS "EnePS9to14" {*/r:9/*,*/r:10/*,*/r:11/*,*/r:12/*,*/r:13/*,*/r:14/*}

QS"EnePS3to14"{*/r:3/*,*/r:4/*,*/r:5/*,*/r:6/*,*/r:7/*,*/r:8/*,*/r:9/*,*/r:10/*,*/r:11/*,*/r:12/*,*/r:13/*,*/r:14/*}