國 立 交 通 大 學

電信工程研究所

碩士論文

中英夾雜語音之階層式韻律架構建立與語音合成之應用

Prosody Hierarchy Construction for Mixed Chinese-English Spelling Speech and its Application to TTS

研 究 生：蔡承燁

指導教授：陳信宏 博士

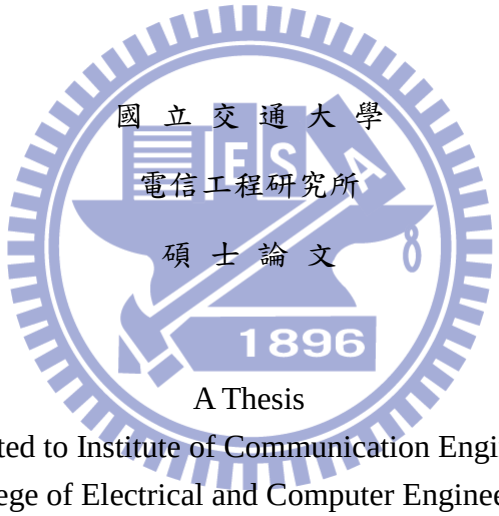
中華民國九十九年八月

中英夾雜語音之階層式韻律架構建立與語音合成之應用

Prosody Hierarchy Construction for Mixed Chinese-English Spelling Speech and its Application to TTS

研 究 生：蔡承燁　　Student：Cheng-Yeh Tsai

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



國 立 交 通 大 學

電信工程研究所

碩 士 論 文

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 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年八月

中英夾雜語音之階層式韻律架構建立與語音合成之應用

研 究 生：蔡承燁　　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

本論文針對以中文文句為主體但內含英文字母之中英夾雜文句，透過語言參數和聲學參數間的關係，建立一個中英夾雜的韻律模型，並完成自動化的韻律標記。本研究所標記的韻律標記為停頓標記及韻律狀態，其中停頓標記表示韻律單元的邊界，而韻律狀態的序列表示上層韻律單元的變化。透過分析訓練出的模型參數，探討停頓標記、聲學參數、語言參數和上層韻律狀態的關係。由實驗結果顯示英文字母之上層韻律狀態是隨著整體中文語句的韻律變化而起伏，而停頓標記則是在 code-switch 處會有較強的韻律斷點。此外也發現到名詞片語的韻律層次結構和其語法結構有很高關聯性。

最後利用此模型提出兩種韻律產生方法，第一種為藉由停頓標記的預估，產生韻律層次的文脈相關資訊，透過 HTS 產生韻律參數，第二種則是應用前述的韻律模型直接預估韻律參數。由客觀評估的實驗結果顯示，第一種方法的確能改善傳統 HTS 所產生之韻律參數，第二種方法則是在音節長度預測有顯著的效果。而主觀評估的結果也顯示第一種方法在聽覺上有最佳的自然度表現，代表透過本研究所預估的停頓標記能抓到更自然的韻律節奏變化。

Prosody Hierarchy Construction for Mixed Chinese-English Spelling Speech and its Application to TTS

Student：Cheng-Yeh Tsai　　Advisor：Dr. Sin-Horng Chen

Institute of Communication Engineering
National Chiao Tung University

Abstract

In this thesis, an unsupervised joint prosody labeling and modeling (PLM) method for mixed Chinese-English word spelling speech is proposed. It labels an unlabeled corpus with two types of prosodic tags (i.e., break type of inter-syllable juncture and prosodic state of syllable) and builds four prosodic models simultaneously. The break tags can be used to delimit prosodic constituents of a hierarchical prosody structure, and the prosodic state can be used to construct the prosodic feature patterns of prosodic constituents. The four prosodic models describe the relationships of acoustic prosodic features, prosodic tags of utterances, and the linguistic features of the associated texts. The experimental results showed that prosodic variation in English word spelling was influenced by both the prosodic state that describes underlying intonation and Chinese tone borrowing effect. Besides, the relationship between hierarchical noun phrase structure and corresponding break type was also analyzed. The analysis suggested that magnitude of the break type was highly correlated with syntactic hierarchy in a noun phrase.

Lastly, we propose two prosody generation methods for mixed Chinese-English word spelling

Text-to-Speech system (TTS) based on PLM. In the first method, a break predictor is constrcted by CART method. Then, the related linguistic features and the predicted break tags are used for HMM-based Text-to-Speech system (HTS) training. In the second method, PLM is directly used as a prosody generator. Experimental results confirmed that the proposed method one was superior to the conventional HTS that only use linguistic features both in objective and subjective tests. Besides, the proposed method two was significantly better than the conventional HTS method at syllable duration prediction. Therefore, we conclude that the proposed PLM method was successful in prosody labeling and modeling for constructing a mixed Chinese-English word spelling TTS.

致謝

首先感謝陳信宏老師當初毫不介意非本科系出身的我，讓我能順利進入這個實驗室，進入語音這個領域。非常感謝陳信宏老師和王逸如老師這兩年來在研究上的細心指導，感謝陳老師在百忙之中仍然心繫著我的研究，仍抽空提點我在研究上的盲點。感謝王老師教導我如何做一個真正的研究生，而不是交作業的大學生。

接著要感謝偉大的性獸學長，感謝性獸這兩年來不辭辛勞地教導我所有語音相關的大小知識，不論是在觀念上的指導或是程式撰寫上的技巧，真的讓我獲益良多，而在最後那段十萬火急的時刻，即便本身就很忙的性獸仍然義不容辭地親下火線幫忙，更讓我由衷地敬佩你，也多虧你才能讓陶小姐完整的回來，實驗室真的不能沒有你。也感謝阿德學長這兩年來不論在研究上和 8051 教學上的指導幫忙，或是一些其他生活上的分享。感謝總是很有耐心的讓我問一些小問題的智合學長，讓我養成隨手關檯燈和螢幕等好習慣的希群學長，提供我們一些面試上資訊的巴金學長，還有愛嚇人的輝哥，在你的恐嚇中也讓我變得更積極。也非常感謝普烏學長這兩年的照顧，感謝你即使當兵了也常常回來探望我們和載大家出去吃宵夜談談心，我會再幫你物色松山正妹的。也感謝效率超高超早寫完論文的 Q 哥、熱愛梁靜茹的小宋和深受文字獄所苦的小帥哥等學長的照顧。感謝非常優秀的實驗室一哥宥余這兩年來不論是在 8051 或是研究上一些幫忙和相助，感謝和我修課幾乎一樣的皓翔哥，讓我修課修的很安心，感謝最後一起睡實驗室奮鬥且和文良一樣認識很多學妹和人妻的普馬，趕快衝一個吧!感謝口試變得超謙虛但真的很有實力的財祿，我相信你會撐過三個月的，感謝交大正妹依玲，恭喜妳擺脫客語的糾纏，感謝似乎真得有點耳背且是小帥哥傳人的舒舒，也恭喜妳逃離文字獄。也在此感謝學弟們：人妻殺手文良、仙道全、大胖哥、舞林高手智障、應該念博班的銘傑、帥氣的小蝦、豆腐喵。也感謝松山高中 316 的友情支持與鼓勵，還有交大土木各位的支持，讓我能一路走過來。

最後感謝我爸媽在我求學路上的一路支持，無論我做出任何決定都尊重我，讓我無後顧之憂。也感謝妳一路的相伴與支持，最後僅以此論文獻給以上的各位。

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究動機

身在 21 世紀科技爆炸的時代裡，電腦、手機等電子資訊產品以飛快的速度推陳出新中，且幾乎已成為人類生活中不可或缺的必需品，因此近年來，人機介面的發展也備受注目。從以往透過按鍵、滑鼠等操作電子產品，為了追求更便利更人性化的方式，語音便成為人類與機器間溝通的最佳選擇。語音合成技術則在此扮演相當中重要的角色。

近年來，單一語言的文字轉語音系統(Text-to-Speech)已經相當成熟，但鮮少有人發展出品質極佳的多語言系統，然而在全球化的潮流裡，人們往往會接觸到不同國家的人，無論是在商場上，甚至大專院校中也可看見越來越多國際交換學生，因此人們俱備兩種以上語言能力已變得越來越普遍。人們也更加期待機器可以同時說出不同語言，使得生活更加便利。

在現今華人社會裡的使用環境中，如中英文夾雜，國台語夾雜等都是日常生活中常見的情形，又以中英文夾雜尤為實用，例如:常用的縮寫專有名詞(PDA、IBM)、地名(New York)、人名(Alice、John)等都常出現在新聞報章雜誌以及實際生活中的對話。因此本論文希望透過分析台灣人念的中英文夾雜文句之韻律，進一步完成中英文夾雜之文字轉語音系統。

1.2 文獻回顧

傳統上，研究多語言文字轉語音系統多半以實現整個系統為主，多以 Corpus-based【1-3】為基礎的語音合成系統。【1】先錄製大量單一語者的單一語言語料，採用音素對應(phone mapping)找出次要語言與主要語言間的轉換關係，在合成次要語言的音素時，找出所對應的主要語言的音素，並使用單元選取，從語料庫中選出適當的單元合成，但此方法面臨的問題是即便兩種語言有相近的音素可以共享，但卻沒有足夠的較長合成單元(longer chunk)，因此【1-2】又嘗試錄製單一語者的多語言語料庫，使用共同的單元選取模組，但仍然沒有討論到 code-switch 處可能存在的聲學現象，且要找到一個會說多國語言的語者，並錄製語料是

相當困難且耗時的事情。【4】則提出以 RNN-MLP-based 為基礎下，目標合成出含有中英文夾雜的文句。以原有 RNN-based 中文韻律產生器【5】為基礎，將中英文夾雜文句中英文的語言參數，先以類似特性的中文之語言參數取代之，產生第一階段的中英文夾雜韻律，第二階段再使用 MLP 機制，針對第一階段 RNN 產生之英文韻律加以修正與學習，因而產生更精確的英文韻律。但此方法缺點在於要使用 RNN-MLP 訓練，需要大量的中英文夾雜文句，然而本研究採用的語料量略顯不足，且此方法不容易作錯誤分析。【6】則是少數專注在模擬混合語言中的韻律變化，觀察到中文夾雜少量英文單詞的文句中，英文單詞的韻律特性，如基頻和音節長度比起此單詞在全英文文句中來的更加變化劇烈與更長，因此採取了簡單的線性轉換，來修正原本使用全英文語料所預估出的英文單詞韻律，達到更符合在中英文夾雜文句中的英文韻律特性，但仍然沒有考慮 code-switch 時兩種語言的互相影響程度。【7】則是基於 HMM-based 合成器下，使用單一語者分別錄製中、英文的語料進行 HMM 訓練，並使用 Kullback-Leibler divergence【8,9】建立誇語言間的狀態對應(state mapping)關係，以供給只會說英文的語者，利用建立好的中英狀態對應關係，合成出中文或中英混合的語音。

1.3 研究方向

中英文夾雜文句中的英文部分又可分為兩種，一類為以字母為單位發音(spelling)的文句，如 NBA、MLB；另一類則是以音節組合而成，依照英文音標發音的文句，如 word、paper。本論文研究針對第一種類別進行韻律之分析，觀察此中文部分與一般中文語料有何不同之處，且著重在英文字母的韻律預估，並分析 code-switch 處帶給中文字與英文字母在韻律上各有何影響。因此本論文基於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)【10】為基礎，針對英文字母做特定修改，進行中英文夾雜的韻律標記(prosodic tags)與韻律模型訓練。

此韻律標記的方法是以中文語音的韻律階層式架構為基礎，透過語言參數，聲學韻律參數，包含音節基頻軌跡、長度以及能量，和音節間韻律參數，包括音節間停頓長度、音節間能量低點、音節間正規化基頻跳躍值及相鄰兩音節間正規化音節延長因子，針對每一個音節邊界(Syllable juncture)標記出其停頓標記與韻律狀態，而本研究語料主要以中文字為主，英

文字母鑲嵌在中文語句中，因此可將每個英文字母視為一個中文音節，同樣標記出每個英文字母的停頓標記與韻律狀態，並加以分析整個中英文夾雜語句的韻律標記分佈情形，code-switch 處是否有何不同等，進一步預估中英文字的韻律參數，最後結合 HMM-based Speech Synthesis System(HTS)完成整個中英文夾雜之文字轉語音系統。

1.4 中英文夾雜 TTS 系統架構簡介

本研究最終目標為改善單純以 HMM-based Speech Synthesis System 所預估產生的語音韻律及合成語音的自然品質。因此在此先介紹本研究所提出之中英夾雜語音合成系統架構圖，如圖 1.1 訓練階段(Training Phase)與圖 1.2 合成階段(Synthesis Phase)。

如圖 1.1 所示，訓練階段為藉由中英夾雜語料的語言參數與聲學參數，使用本研究所提出之 PLM 演算法，自動標記出每個音節所屬的韻律標記，並建立韻律模型，此部分將在第三章詳述，接著藉由傳統語言參數所產生的文本標示(Label)，並加入韻律標記，產生更多韻律層次的文脈相關資訊，幫助 HTS 之 context dependent model 訓練廣義梅爾倒頻譜參數(Mel-generalized Cepstrum, MGC)。

合成階段如圖 1.2 所示，輸入端為中英夾雜文字，經過文字分析器後得到其語言參數，結合訓練端的韻律模型，進行停頓標記與韻律狀態的預估，此部分將在第五章詳述，接著透過預估出的停頓標記，一方面產生中英文夾雜的韻律架構，並產生 HTS 所需之文本標示，進行頻譜參數的估計，另一方面也可以藉由訓練端之韻律模型和預估出的韻律狀態，直接預估出這些文字的韻律參數，包含其音高(Pitch)、音長(Duration)，最後將韻律參數和頻譜參數透過 MLSA filter(Mel-log Spectrum Approximation Filter)【11】，產生中英文夾雜語音。

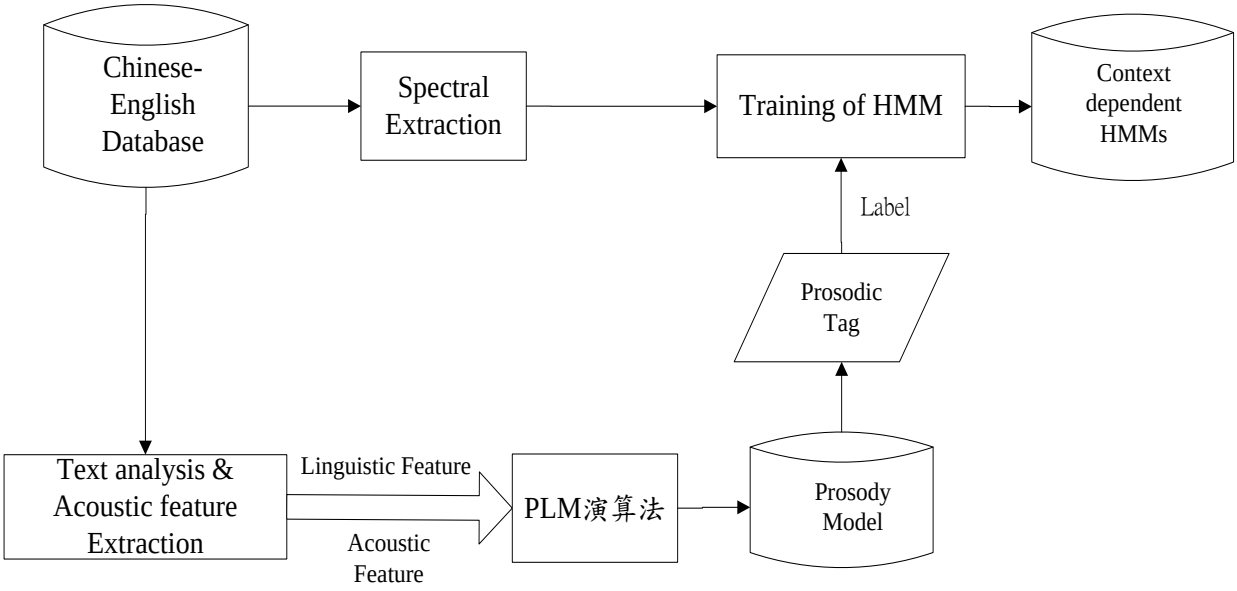


圖 1.1：訓練階段的系統架構圖

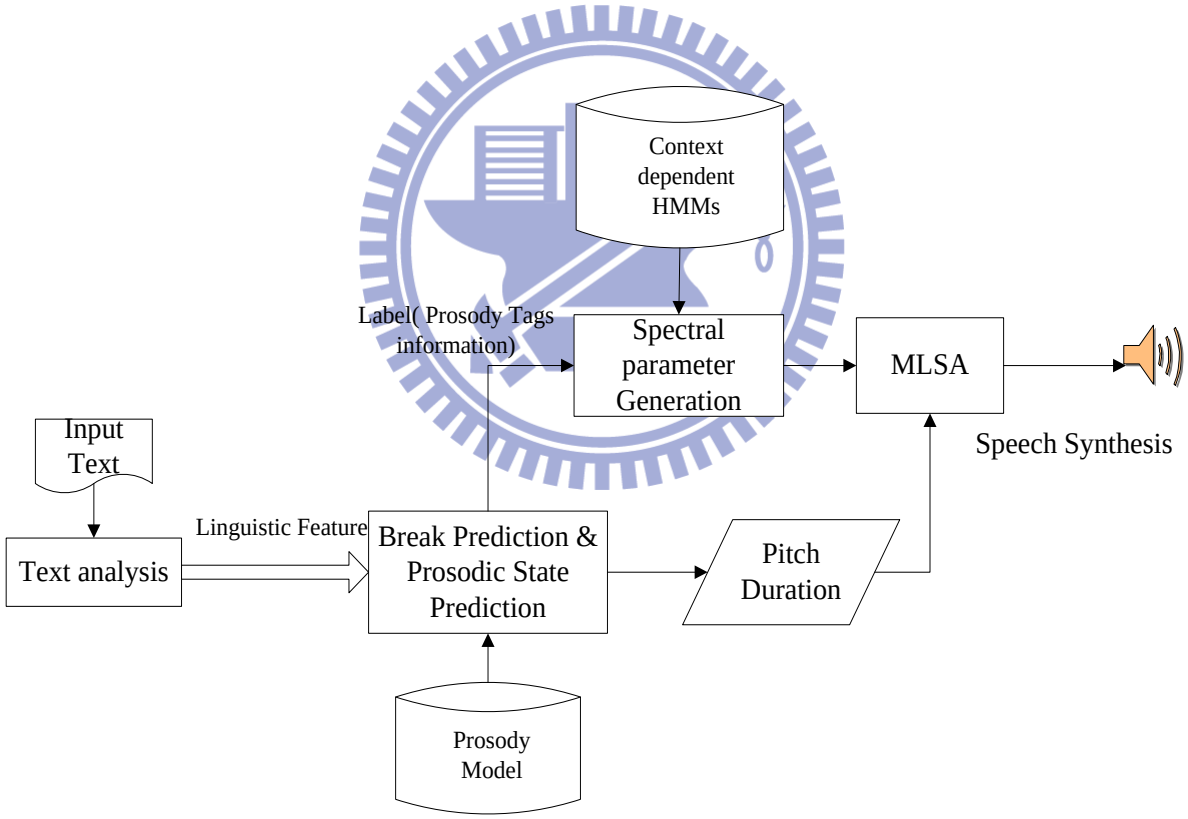


圖 1.2：合成階段的系統架構圖

1.5 語料庫簡介

本論文所使用之中英夾雜語料是由一個專業女性播音員所錄製而成，以中文為主體並穿插英文字母於中文語句中，共 539 個語句，總音節數為 13540 個音節，包含 11688 個中文音節與 1872 個英文字母。音檔為取樣頻率 20000 赫茲(Hertz)及 16 位元數之 PCM 格式，平均語速為一秒 3.5 個音節(3.5 音節/秒)。將此語料庫每 10 句中編號尾數為 7 之句子當作測試語料，其餘為訓練語料。訓練語料共 12185 個音節，包含 10504 個中文音節與 1681 個英文字母，測試語料則共 1355 個音節，包含 1164 個中文音節與 191 個英文字母，詳細訓練語料與測試語料之各英文字母出現次數如表 1.1 與表 1.2。至於所有音節的切割標記、基頻軌跡及能量的偵測均先自動由 HTK(Hidden Markov Model Toolkit)【12】和 WaveSurfer【13】完成，再經由人工修正。如此一來便可得到所有韻律參數，包含音節基頻軌跡、音節長度、音節能量及音節間停頓長度，其中本研究以一組四維正交化參數【14】來描述音節基頻軌跡。此外也透過交通大學語音處理實驗室之斷詞器得到每一個語句的斷詞情形與詞性，詞類分法則是依據中研院詞庫小組所制定的 46 類為準，其中英文詞的詞性多半為專有名詞(NB)，且詞長是以整個詞為單位，如『NBA』為一個三字詞。

表 1.1：訓練語料中個英文字母出現次數

A	B	C	D	E	F	G	H	I	J	K	L	M
122	91	147	147	53	34	56	33	93	9	34	43	102
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
74	51	128	10	62	126	104	37	66	29	16	11	3

表 1.2：測試語料中個英文字母出現次數

A	B	C	D	E	F	G	H	I	J	K	L	M
15	10	18	16	7	5	3	1	14	1	1	3	11
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
11	8	15	0	7	14	13	3	6	5	2	2	0

1.6 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、語音合成系統架構及語料庫說明。

第二章：HMM-based 中英夾雜語音合成器。

第三章：中英文夾雜韻律模型(PLM)：建構中英文夾雜韻律模型以及模型訓練之演算法。

第四章：韻律模型訓練結果分析。

第五章：基於 PLM 演算法之韻律產生器及實驗結果

第六章：結論與未來展望

第二章 HMM-based 中英夾雜語音合成器

本章將描述本論文之所使用之 HMM-based 語音合成系統。2.1 節介紹本研究所採用的基於隱藏式馬可夫模型的語音合系統(HMM-based Speech Synthesis System, HTS) ，2.2 節介紹 HMM-based 中英文夾雜語音合成之基礎系統的建立。

2.1 HMM-based 語音合成系統

隱藏式馬可夫模型(Hidden Markov Model, HMM)早期大量應用在語音辨識系統中，它成功以機率模型描述發音的現象。近年來則被應用到語音合成上，可說是目前語音合成系統中，合成品質相當好的系統。此系統為統計參數式語音合成，由於是統計式參數合成方法，比起傳統以大語料為基礎(Corpus-based)之合成方法，來得更佳彈性且不需耗費大量時間錄製語料與大量空間儲存語料。並透過參數的轉換與調適【15-16】，可輕易產生出不同語者特性的語音。

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(HMM-based Speech Synthesis System, version 2.1)【17】，此系統為基於 HTK 技術，所發展出針對使用隱藏式馬可夫模型建構的語音合成系統。基於隱藏式馬可夫模型的語音合成系統如圖 2.1 所示：

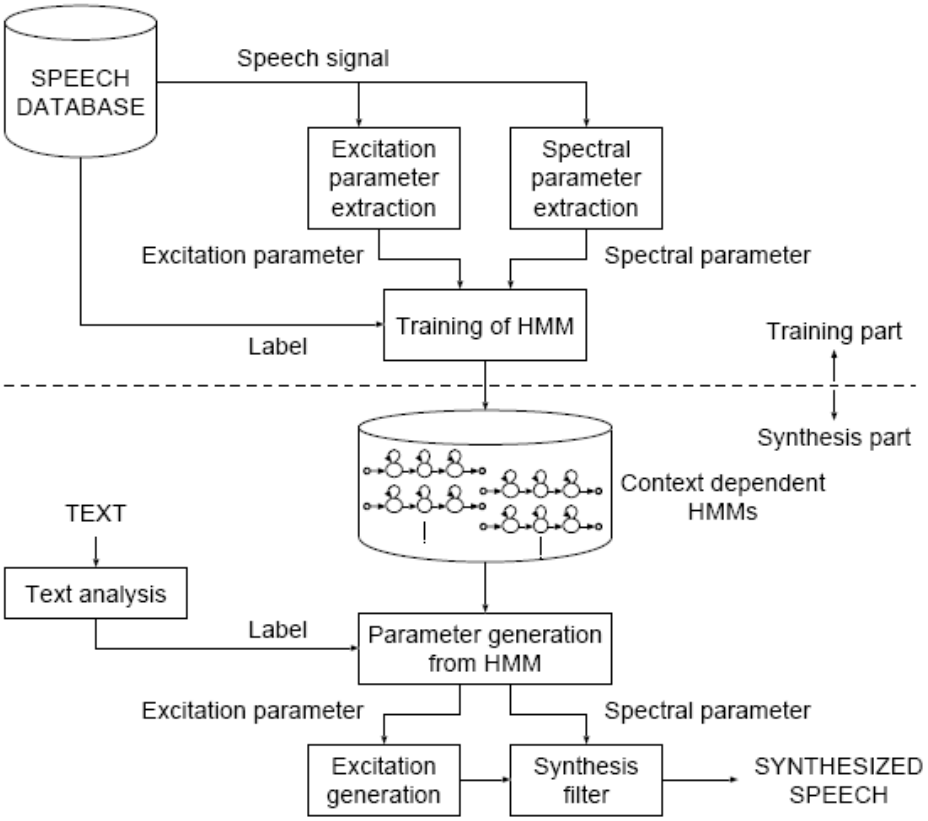


圖 2.1：HMM-based 語音合成系統架構圖，本圖摘錄自 Ref.【18】

如圖 2.1 所示，HTS 分為訓練部分與合成部分，在訓練部分，由語料中抽取其 MGC 參數(phonetic information)，與激發訊號參數(log F0, prosodic information)，搭配相對應的文字分析產生文本標示，再配合適當的文脈相關問題集，訓練狀態合併分裂樹，產生與文脈相對應的 HMM 模型，包含音高模型、頻譜模型及音長模型。合成部分則是輸入文字，透過文字分析器產生與前後文相關的文本標示，再藉由分類與回歸樹(CART)演算法，挑選對應的 HMM 模型序列，經由生成參數演算法，產生頻譜參數與激發訊號參數，再透過 MLSA filter 產生語音信號。

2.2 HMM-based 中英文夾雜語音合成基礎系統的建立

本節將對基於隱藏式馬可夫模型之中英夾雜語音合成基礎系統的建立做詳細說明，包含了中英夾雜音素模型的挑選與建立、文本標示及狀態合併分裂樹所需之問題集設計等。

2.2.1 中文與英文字母音素模型

中文為一聲調語言(Tonal language)，以音節(syllable)為單位，每個中文字對應到一個音節(syllable)。中文共有 411 個基本音節(basic syllable)，加上 5 種聲調，組成約 1300 多個音節，涵蓋了幾乎所有常見的中文發音，而每個音節以聲母、韻母及聲調組成，聲母可分為 22 類，韻母則為 40 類。透過聲母、韻母的組合即可涵蓋大部分中文音節，因此本研究在建立中文音素 HMM 模型時，採取以聲母、韻母為基本音素單元。如此一來可使得語料庫中不包含的音節，或是較少出現的音節，藉由其他相同聲母或是韻母來幫忙訓練，得到較可靠的 HMM 模型。

至於英文字母之 HMM 模型，則是考量由於本語料庫之英文字母數量相對太少，因此若是將每個英文字母視為一種音素，建立其 HMM 模型，恐怕會因資料量過少而產生不可靠的參數，進而影響聲音品質，為了避免上述情形，本研究採取將每個英文字母視為一種類似中文音節結構，拆解成與中文相似的聲母、韻母作為音素單元，例如：B 拆解成：B_1 類似中文聲母ㄅ(b)，B_2 類似中文韻母ㄧ(i)。此外原本英文字母共只有 26 個，但由於 M 這個字母在語料庫中明顯有兩種不同之基頻軌跡，其英文字母特性將在第三章中詳述，因此將 M 再細分為 M_r(rising)、M_f(falling)，得到共 27 類英文字母。採取以上方式建立英文字母之 HMM 模型的優點在於透過中英文相近的發音，可共享其 HMM 模型，避免英文字母出現次數太少而產生的不可靠模型，缺點則在於即便英文字母有發音相似的中文聲母、韻母，但仍有些差異較大的字母，如 F、H、S、X 則是剛好和中文音節結構相反，即韻母在前，聲母在後。又如 W、L 其實為雙音節字母，因此與中文聲母、韻母共享模型時可能會發生不匹配的狀況。

2.2.2 文本標示資訊與問題集設計

文本標示資訊為 HTS 相當重要的一環，採用哪些語言參數會直接影響到 context dependent model 的狀態分裂合併結果。根據 2.2.1 小節所定義的中英文音素模型，加上利用前後文相關的語言參數，輸出文本標示。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)，

最後特別加上 code-switch 的資訊，詳細所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 initial/final
	-Cur_Phn	Current initial/final
	+Fol_Phn	Following initial/final
	^Phn_in_Syl	Initial/final 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_Wrd	Syllable position in a lexical word (forward)
	|B_Syl_in_Wrd	Syllable position in a lexical word (backward)
Word level	/A:Pr_PM	PM type preceding current syllable
	/B:Fol_PM	PM type following current syllable
	/C:Pre_2_POS	46-type POS (Academia Sinica) of word preceding the previous word
	/D:Pre_1_POS	46-type POS of previous word
	/E:Cur_POS	46-type POS of current word
	/F:Fol_1_POS	46-type POS of following word
	/G:Fol_2_POS	46-type POS of the next word following the following word
	/H:Pre_2_WL	Word length of word preceding the previous word
	/I:Pre_1_WL	Word length of previous word
	/J:Cur_WL	Word length of current word
	/K:Fol_1_WL	Word length of following word
	/L:Fol_2_WL	Word length of the next word following the following word
Sentence level	/M:F_Syl_in_Snt	Syllable position in a sentence (forward)
	/N:B_Syl_in_Snt	Syllable position in a sentence (backward)
	/O:Snt_L_in_Syl	Sentence length in syllable
	/P:Snt_pre1_L_Syl	Previous sentence length(sp)
	/Q:Snt_fol1_L_Syl	Following sentence length(sp)
Code-Switch	/R:Code_Switch	Code-switch(Chinese to English or English to Chinese Inter-word or Intra-ward)
Phrase level	/S:F_Syl_in_Phr	Syllable position in a Phrase (forward)
	/T:B_Syl_in_Phr	Syllable position in a Phrase (backward)
	/U:Pre_1_POS	46-type POS of previous Phrase
	/V:Cur_POS	46-type POS of current Phrase
	/W:Fol_1_POS	46-type POS of following Phrase
	/X:Pre_1_WL	Word length of previous Phrase
	/Y:Cur_WL	Word length of current Phrase
	/Z:Fol_1_WL	Word length of following Phrase

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，為了達到最佳狀態分裂合併結果，考量五大類問題集，說明如下：

1. 音節層次(syllable level)：

   i. 考慮當前音素與前後音素(initial、final):

   - 聲母發音類別：爆破音、摩擦音、鼻音、邊音、塞擦音。
   - 韻母發音類別:單元音韻母、複合元音韻母、鼻尾音韻母。

   其中遇到英文字母時，採取 2.2.1 節中所提出之方式，將英文字母拆解成相似中文聲母、韻母結構，因此在問題集設計上可將中英文發音方式相近的音素共用，如 B 可以與爆破音中的ㄅ(b)共用。

   ii. 考慮當前聲調與前後聲調：

   - 考慮下一個音節聲調：可將聲調簡化考慮成基頻起始較高(H)，如中文一聲和四聲及基頻起始較低(L)，如中文二聲和三聲。
   - 考慮前一個音節聲調：可將聲調簡化考慮成基頻結尾較低及節尾較高。
   - 英文字母部分，以每一類字母獨立作為一種聲調，並可與相似基頻高度之中文聲調共同設定問題集。

   iii. 考慮音素在字中位置：由前面數來，由後面數來。

   iv. 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。在不同詞中位置都是有可能影響最後聲音的韻律特性，如在詞首或是詞尾。

2. 詞層次(word level)：

   i. 考慮當前詞(±0)與前後兩個詞(±1、±2)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)與前後兩個詞(±1、±2)的詞長。

   iii. 考慮前後音節是否有標點符號。

3. 片語層次(phrase level)：

   i. 考慮音節在片語中位置：由前面數來第幾個字，由後面數來第幾個字。

ii. 考慮當前片語與前後一個片語的詞類。

iii. 考慮當前片語與前後一個片語的詞長。

4. 句子層次(sentence level)：

i. 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。

ii. 考慮當前句子與前後句子長度。

5. Code-Switch：

i. 考慮當前音節是否在中英文交界處，且分詞內邊界(Inter-word)和詞外邊界(Intra-word)：如中文詞轉英文詞且位於詞外邊界(C-E-Inter-word)、英文詞轉中文詞且位於詞內邊界(E-C-Intra-word)等。

綜合以上類別的考量，本論文使用約 1700 個問題集及 2.2.1 節中的音素模型，並依照圖 2.1 之 HTS 方塊圖，實作出一套 HMM-based 中英夾雜語音合成基礎系統。

第三章 中英文夾雜韻律模型

本論文所使用之韻律模型是基於【10】中所提出之中文韻律模型，針對中英文夾雜特性作適度修改，完成中英文夾雜之韻律標記與韻律模型訓練。本研究利用聲學參數、語言參數和待預估的韻律標記，設計出四個子模型並基於最大概似度準則(maximum likelihood criterion)採用逐項最佳化程序(sequencial optimization procedure)求得最佳化參數，完成韻律標記以及韻律模型。3.1 節將介紹中英文語音韻律之特性。3.2 節將介紹階層式韻律架構。3.3 節將簡介中文 PLM 演算法。3.4 節將介紹中英文夾雜 PLM 演算法。3.5 節則介紹模型訓練及參數更新演算法。

3.1 中英夾雜語音韻律之特性

有別於語調語言(intonation language)，中文屬於一種聲調語言，其特點為一個中文字由一基本音節搭配不同聲調所組成，其基本音節共有 411 個，搭配五種聲調總音節數共 1300 多個。不同聲調則會產生不同語義，因此聲調為中文語音韻律一個重要的特徵，其影響韻律層面最大之處在基頻軌跡的變化，圖 3.1 為中文一到四聲之基頻軌跡，可知中文一聲的基頻軌跡通常為一水平線，且依照音高範圍分類為高至高(High-to-High)，即音節起始為高音頻，音節結尾也為高音頻；二聲的基頻軌跡呈現低至高(Low-to-High)的走勢；三聲的基頻軌跡為一勺狀曲線，即音節前後音高比中間來的較高；四聲的基頻軌跡則由高至低(High-to-Low)，至於中文五聲(輕聲)的基頻軌跡通常沒有固定形狀，因為其往往受到前後音節聲調影響。本研究所屬的英文字母本身屬於語調語言，但由於中文本身為聲調語言，因此由華人所念出之中文帶有少數夾雜英文字母之語句時，往往會不自覺地將英文字母發音成帶有聲調的韻律，我們稱此現象為 Tone Borrowing。為了驗證此說法，我們將 26 類英文字母依其平均基頻軌跡和相似地中文五種聲調在畫在同一張圖上，如圖 3.2 至圖 3.5。由圖 3.2 可看出英文字母{A,B,C,D,E,G,I,J,K,N,O,P,Q,T,U,V,Y}之基頻軌跡和中文一聲相似；圖 3.3 則可看出英文字母{F,H,R,S,X}之基頻軌跡和中文二聲相似;此外由圖 3.4 我們也發現 M 之基頻軌跡有兩種完全

不同的形狀，一種為類似中文二聲的低至高(Low-to-High)，另一種為類似中文四聲的高至低(High-to-Low)。至於剩餘英文字母 W 和 Z 之基頻軌跡曲線則較為凌亂，如圖 3.5 所示。

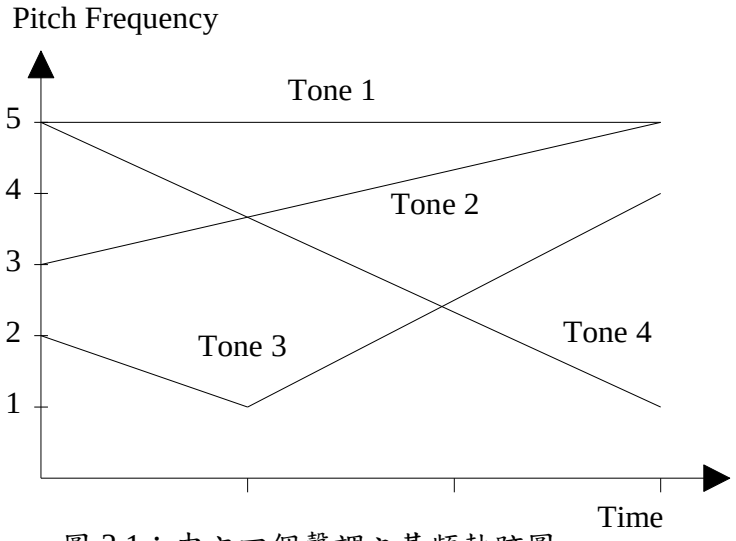


圖 3.1：中文四個聲調之基頻軌跡圖

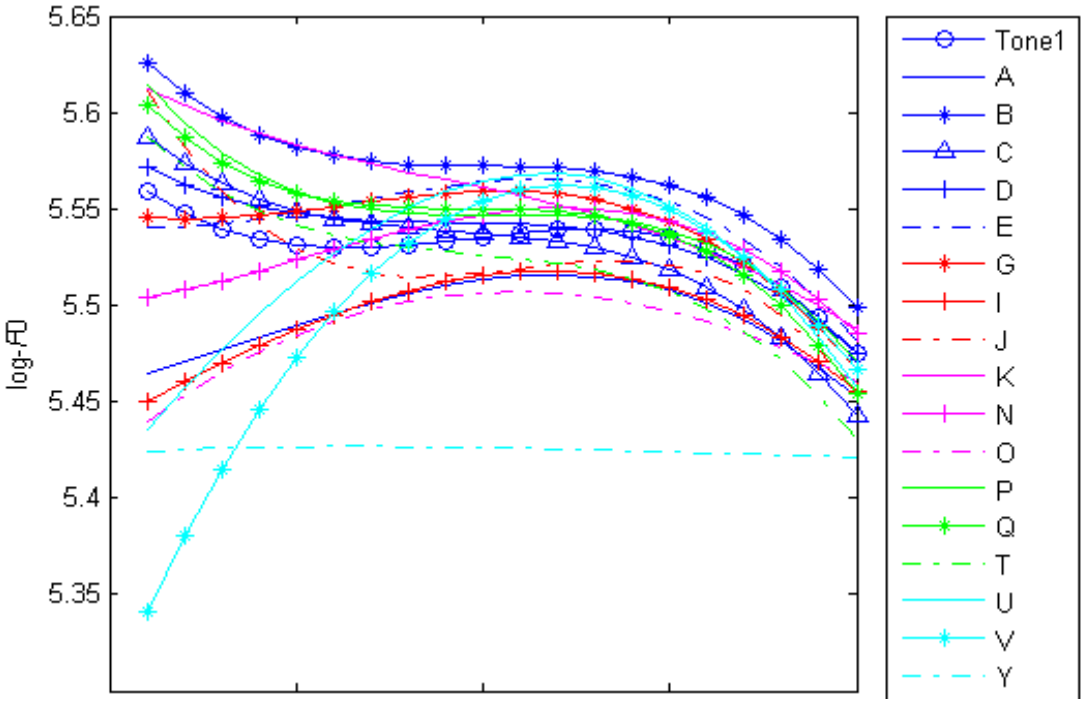


圖 3.2：中文一聲和英文字母{A,B,C,D,E,G,I,J,K,N,O,P,Q,T,U,V,Y}之基頻軌跡

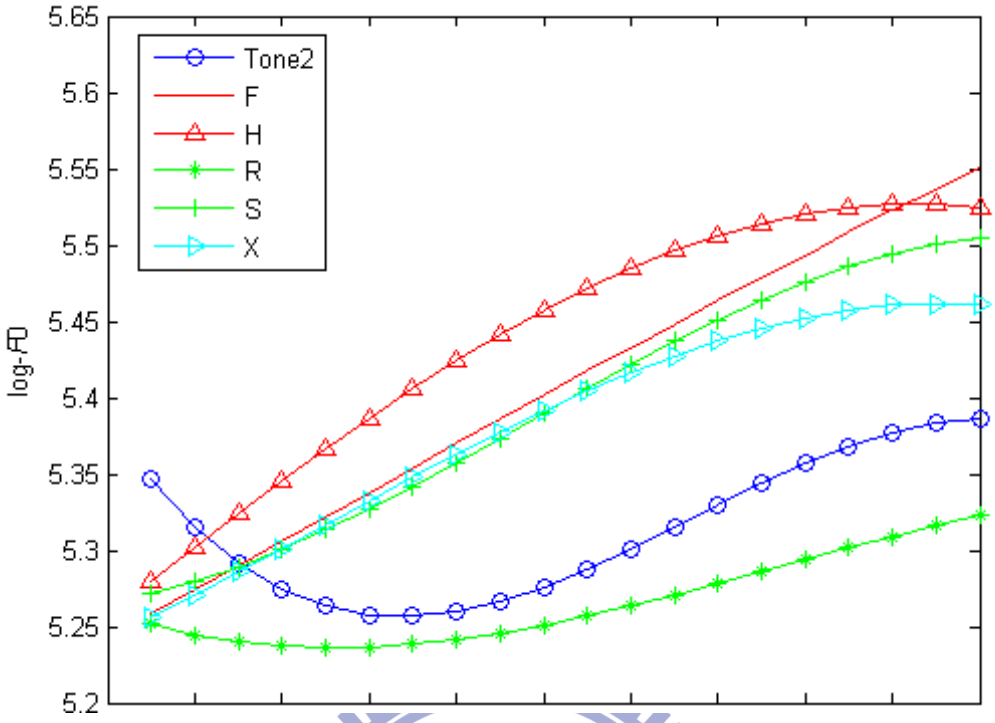


圖 3.3：中文二聲和英文字母{F,H,R,S,X}之基頻軌跡

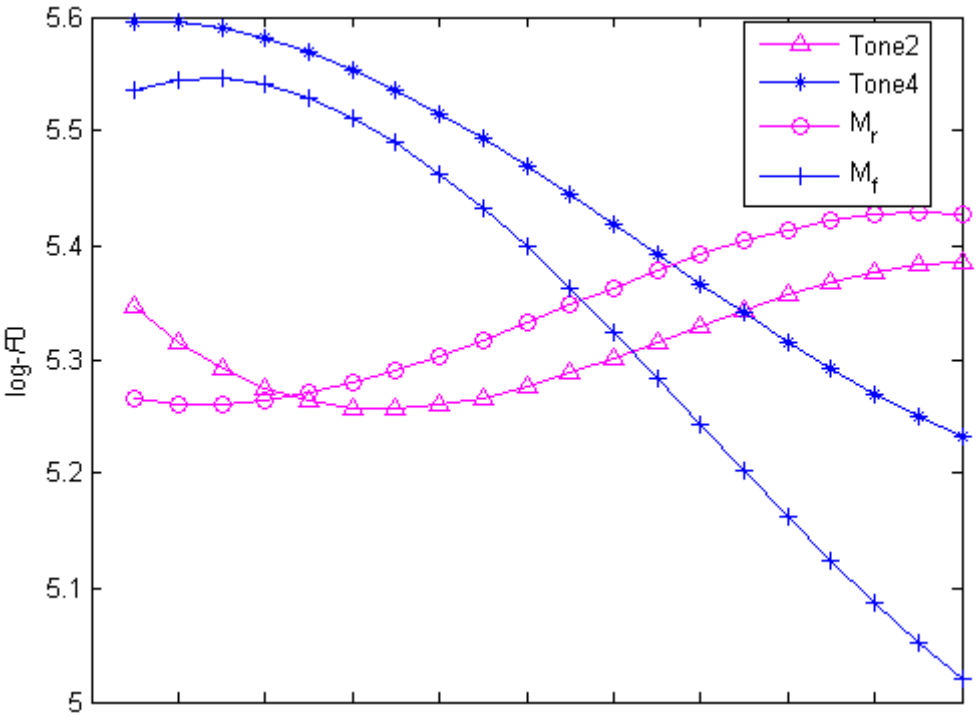


圖 3.4：中文二聲和四聲與英文字母 M 之基頻軌跡

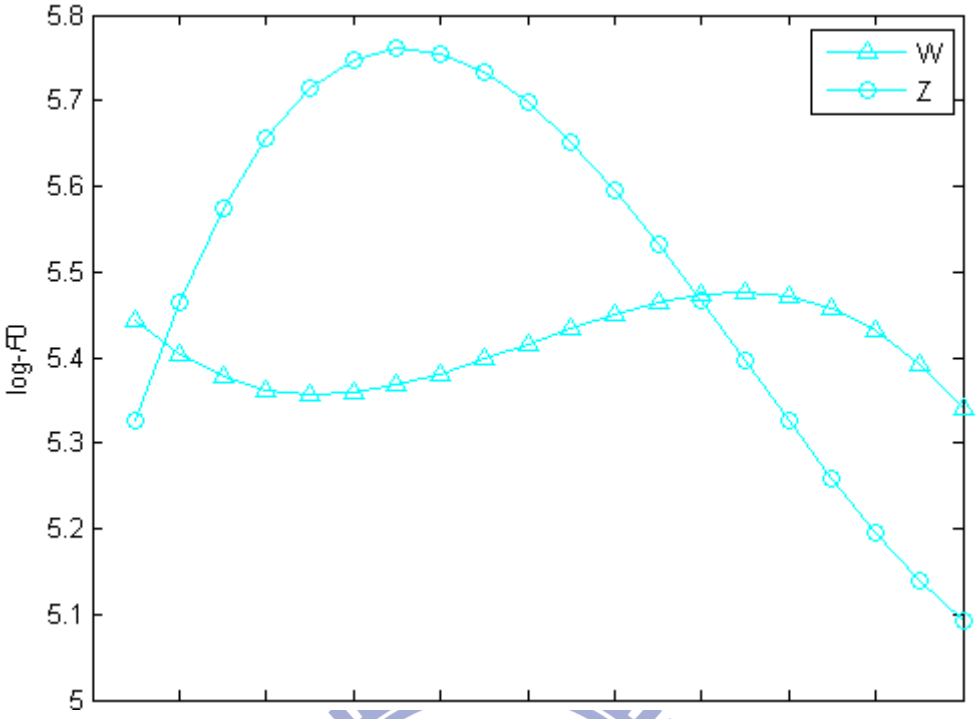


圖 3.5：英文字母 W 和 Z 之基頻軌跡

除了 Tone Borrowing 現象之外，我們也發現本研究語料往往在 code-switch 處有較長的停頓，表 3.1 中顯示在詞外邊界(Inter-wod)且為 code-switch 邊界時，其停頓長度明顯比 Non-code switch 要來的長很多，而在詞內邊界(Intra-word)且為 code-switch 邊界時，也有同樣的結果。我們認為會有此現象的原因在於英文字母通常為一句話的重點，為了強調英文字母通常會念的比較慢，因此 code-switch 處的平均停頓長度會來的比一般 Non-code switch 處要來的長。

表 3.1：code-switch 和 Non-code switch 處的平均停頓時長，其中括號中數值為其數量，且此統計量不跨標點符號。

	Chinese-Chinese	English-English	Chinese-English	English-Chinese
Inter-word	0.05 sec (4416)	0.06 sec (1)	**0.16** sec (441)	**0.11** sec (435)
Intra-word	0.014 sec (4206)	0.037 sec (993)	**0.06** sec (20)	**0.0408** sec (47)

3.2 階層式韻律架構

根據許多研究中文韻律的文獻【19】顯示，中文的韻律結構呈現階層式韻律架構(hierarchy structure)，一般來說分成四層結構，如圖 3.6 所示，由底層至上層分別為音節層次(Syllable, SYL)、韻律詞層次(Prosodic Word, PW)、韻律短語層次(Prosodic Phrase, PPh)以及語調短語層次(Intonation Phrase, IP)。由於一個中文字為一個音節的特性，因此最底層的韻律單元為音節層次，而不同聲調所帶來的不同語義，也使得聲調成為音節層次中最重要的韻律影響因素，聲調不只影響音節音高甚深，也進而影響音節長度與音節能量。第二層的韻律詞層次則是由雙音節或多音節的詞組所組成，這些詞組通常在句法或是語意上緊密相關，因此往往會將這些詞組發音成一個單元。第三層的韻律短語層次則是由一個或多個韻律詞所組成，其結尾通常有帶有可察覺但不明顯的停頓。第四層的語調層次則是中文韻律架構的最上層，通常限制了一個句子或是數個韻律短語所組成的句子音高，其結尾則會有明顯的停頓。基本上，四層的韻律架構詮釋了一個句子中每個音節的音高和音長變化。

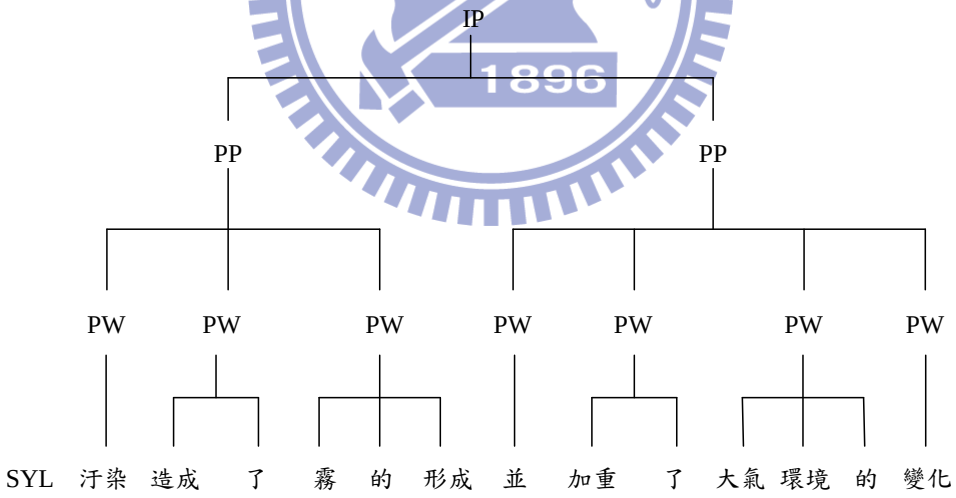


圖 3.6：中文常見的階層式韻律架構 ，本圖摘錄自【19】

此外鄭秋豫博士【20】提出另一個韻律架構並提出韻律標記的概念，如圖 3.7 所示。此架構將中文韻律結構分成五層，前三層和第一種韻律架構一樣，分別為音節層次(Syllable, SYL)、韻律詞層次(Prosodic Word, PW)以及韻律短語層次(Prosodic Phrase, PPh)。第四層則是將連續的 PPh 組合成一個呼吸群(Breath Group, BG)來代表大範圍且有基頻及音長變化的篇

章或是段落，藉此表示韻律更上層的貢獻，同理定義了第五層為連續的 BG 所組成的韻律群(Prosody Group, PG)。而上述所說的五層架構則採用六種標記來區分，B0 和 B1 代表 SYL 的邊界，差別在於 B0 表示 reduced syllable boundary，B1 則是 normal syllable boundary，且通常在 B0 及 B1 的邊界聽不出停頓。B2 及 B3 分別區分 PW 和 PPh 的邊界，B4 和 B5 則是區分 BG 和 PG 的邊界，B4 代表一個呼吸的停頓，B5 則為一個完整語音段落的結束，並且可以明顯感受到句尾的音節長度拉長(final lengthening)以及能量的減弱。

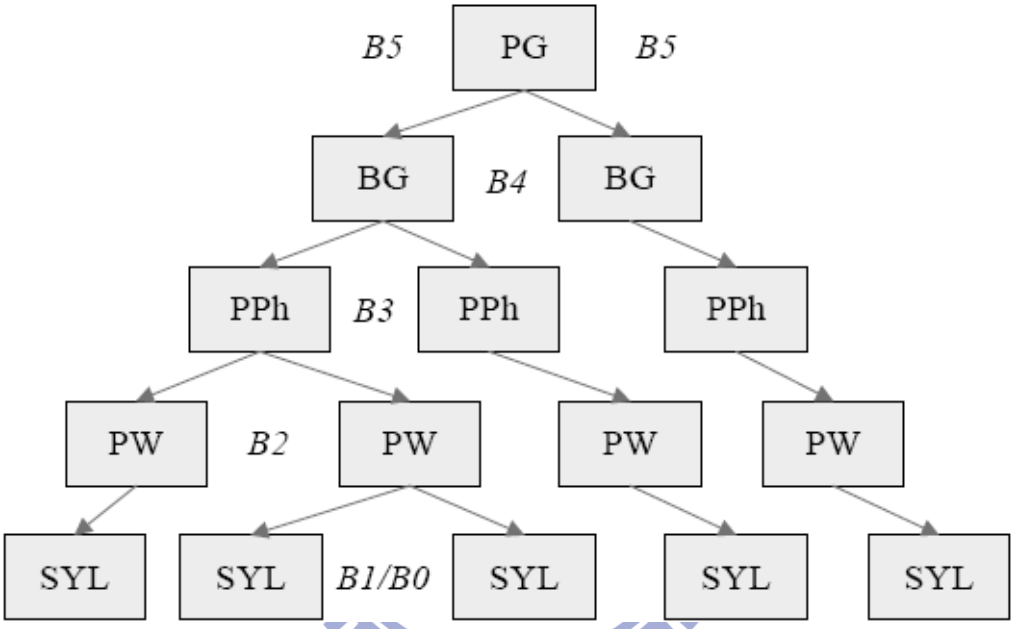


圖 3.7：階層式多短語韻律句群(Hierarchical Prosodic Phrase Gruoping，HPG)架構。【20】

本研究主題雖然為中英文夾雜的語句，但文句還是以中文為主體，英文字母少量穿插在中文句子中，且英文字母大多數都為單音節結構，和中文字結構相仿，因此我們假設在正常語流的情況下，英文字母本身的上層韻律變化(High-Level Prosodic information)與韻律斷點會受到中文整體韻律變化的限制與影響。此觀點也可與 3.1 節中所提到的 Tone Borrowing 之現象呼應，因為當前後中文字都屬於聲調語言的情況下，英文字母的發音也很容易變成帶有聲調的發音，如 A 類似中文一聲。基於上述理由，本研究仍以 HPG 的中文韻律架構為基礎，進一步對其做修改，利用修改後的架構作為中英文夾雜文句的韻律模型架構。

首先我們將 B2 再細分為 B2-1、B2-2 及 B2-3，分別代表明顯音高重置(pitch reset)之韻律詞邊界、短停頓(short pause)之韻律詞邊界以及含有音節延長效應(duration lengthening)之韻律

詞邊界。接著我們將 B4、B5 合併為 B4，整個韻律架構由 5 層變回 4 層，如圖 3.8 所示。綜合上述，本研究採用了 7 種停頓標記(Break Type) **B** ={B0,B1,B2-1,B2-2,B2-3,B3,B4}，來標記這四種韻律單元：音節(SYL)、韻律詞(PW)、韻律短語(PPh)及呼吸群/韻律群(BG/PG)。值得注意的是由於上述中英文夾雜語句的特性，本研究並不會因為是英文字母所對應到的停頓標記而給定特別不同於中文的標記。

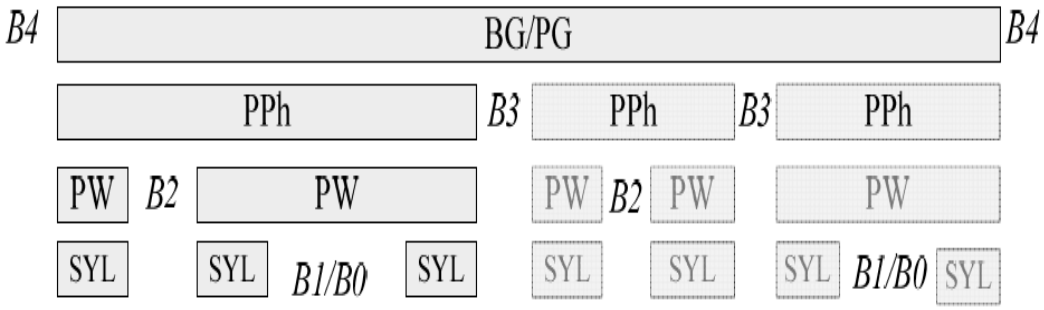


圖 3.8：本研究所用之階層式韻律架構

在此將 B2 分成 3 類是因為雖然同屬於韻律詞的邊界，但其對應的聲學特性仍然有所不同，原先的單一類別不足以將其差異性描述出來；而將 B3 及 B4 合併則是因為它們所對應到的聲學特性相近，故不需要再用額外的韻律邊界停頓來表示之。

為了要更進一步描述這四層的階層式韻律架構，除了描述韻律單元邊界的停頓標記外，還需要描述韻律單元變化的韻律標記或參數。而本研究利用一些帶有韻律組成份子資訊的標記來間接的表示這些韻律組成份子，此標記即為韻律狀態，其意義代表圖 3. 8 架構中最上面三層之韻律組成份子個別的貢獻。在本研究中會採用三種不同的韻律狀態，分別量化正規化後的音高、音長和音節能量。正規化後的音高為扣除掉音節層次對音高的貢獻，此時音高的韻律狀態代表的是韻律詞、韻律短語、呼吸群/韻律群對音高的貢獻。至於音長或音節能量則同理扣除音節層次影響因素，使其分別表示最上面三層之韻律詞、韻律短語、呼吸群/韻律群對音長和音節能量的貢獻。簡而言之，音高、音長和音節能量的韻律狀態分別表示每個音節在韻律詞層次以上貢獻的音節音高平均、音節音長和音節能量。這樣做的好處在於，我們能將音高、音長和音節能量在低層次和高層次的影響因素分開，將複雜的高層次影響因素通通由韻律狀態來表示。此外韻律狀態標記一樣不會因為英文字母而給予特定標記，原因如

上述中英文夾雜語句特性，其上層韻律變化會受到中文整體韻律變化的限制與影響，因此韻律狀態的標記重點在於前後音節上層韻律變化而不是底層的音節所屬類別。

3.3 中文 PLM 演算法

由於本研究所採用的中英文夾雜韻律模型是基於江振宇博士所提出之中文韻律模型(PLM)演化而來，因此在此先介紹何謂中文 PLM，待 3.4 節再介紹中英文夾雜韻律模型。

中文 PLM 演算法為江振宇博士依據中文階層式韻律架構，如圖 3.8，利用語言參數和聲學參數間的相互關係，針對一個未經人工事先標記好的語料，自動標記出其停頓標記及上層韻律狀態，此演算法優點：1.可自動標記，傳統上的韻律標記多為人工標記，既耗時又耗力，且有不一致性的問題。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。

自動韻律標記的問題可視為，在給定聲學參數集合**A**及相對應的語言參數集合**L**之下，目標找到最佳韻律標記之集合$\mathbf{T}^*$，因此可視為一種求取最佳參數解的過程，即:

$$\mathbf{T}^* = \arg\max_{\mathbf{T}} P(\mathbf{T}|\mathbf{A},\mathbf{L}) = \arg\max_{\mathbf{T}} P(\mathbf{T},\mathbf{A}|\mathbf{L}) \tag{3.1}$$

韻律標記集合包含了兩類很重要的中文語音韻律資訊，第一類是參考階層式韻律架構所定的音節邊界停頓標記，在此演算法中定義音節邊界停頓標記集合**B**={B0,B1,B2-1,B2-2,B2-3,B3,B4}，其意義如 3.2 節所述，主要區分每一層韻律架構的邊界。另一類的韻律標記為音節的韻律狀態，此演算法定義三種韻律狀態，分別代表經過量化和正規化音節的音高韻律狀態**p**、音長韻律狀態**q**和音節能量韻律狀態**r**。音高韻律狀態代表扣除音節層次對音高的貢獻，即扣除聲調和連音影響參數，此時**p**則為韻律詞、韻律短語、呼吸群/韻律群對音高的貢獻。同理音長韻律狀態和音節能量韻律狀態則分別扣除聲調、基本音節類型或韻母類型的影響參數，使其分別代表韻律詞層以上對音長和音高的貢獻。此演算法定義韻律標記之集合為**T**={**B**,**PS**}，其中，**PS**={**p**,**q**,**r**}為韻律狀態標記集合。

此演算法所使用之聲學參數可分為三類，第一類聲學參數為音節韻律參數(syllable prosodic feature)，假設第一類參數和韻律狀態標記有很高的關連性，但與音節邊界停頓標記相關性則很低或是獨立，屬於這類的聲學參數有音節基頻軌跡、音長和音節能量。二、三類

聲學參數則用來說明音節邊界停頓標記，我們假設這兩類的聲學參數和音節邊界停頓標記有很高的關聯性，但與韻律狀態標記的相關性很低或是獨立，第二類聲學參數為音節間韻律參數(inter-syllable prosodic feature)，包含音節邊界的停頓時長、音節邊界的 energy-dip level。第三類聲學參數為音節差韻律參數(differential prosodic feature)，包含正規化基頻跳躍值(normalized pitch jump)和正規化的音節長度延長因子(normalized duration lengthening factor)。

綜合上述討論，定義聲學參數 **A** 包含了音節音高軌跡序列 **sp**、音節長度序列 **sd**、音節能量序列 **se**、停頓時長序列 **pd**、音節能量低點(energy-dip level)序列 **ed**、正規化的音節內基頻跳躍值 **pj** 及正規化的音節長度延長因子序列 **dl**、**df**。

其中 **pj**、**dl**、**df** 分別定義為：

$$pj_n = \left(\mathbf{sp}_{n+1}(1) - \boldsymbol{\beta}_{t_{n+1}}(1)\right) - \left(\mathbf{sp}_n(1) - \boldsymbol{\beta}_{t_n}(1)\right) \tag{3.2}$$

在此 **x**(1) 定義為向量 **x** 的第一維度，下標 n 表示為第 n 個音節，$\boldsymbol{\beta}_{t_n}$ 為聲調影響因素 t_n 的 affecting patterns (APs)。

$$dl_n = \left(sd_n - \gamma_{t_n} - \gamma_{s_n}\right) - \left(sd_{n\text{-}1} - \gamma_{t_{n-1}} - \gamma_{s_{n-1}}\right) \tag{3.3}$$

$$df_n = \left(sd_n - \gamma_{t_n} - \gamma_{s_n}\right) - \left(sd_{n+1} - \gamma_{t_{n+1}} - \gamma_{s_{n+1}}\right) \tag{3.4}$$

在此 γ_t 和 γ_s 分別表示聲調與基本音節類型影響因素在音長的 APs。

因此聲學參數集合為 **A**={**sp**,**sd**,**se**,**pd**,**ed**,**pj**,**dl**,**df**}，再因參數特性細分為音節韻律參數 **X**={**sp**,**sd**,**se**}，音節內韻律參數 **Y**={**pd**,**ed**} 以及音節差韻律參數 **Z**={**pj**,**dl**,**df**}。

至於語言參數方面，考慮音節層次中的聲調、基本音節類型及韻母類型，詞層次中的音節邊界型態(Intra-word、Inter-word)、詞長及 POS 和對應到的標點符號類型等。將整個語言參數集合以 **L** 表示，由於聲調、基本音節類型及韻母類型分別對音節音高、音長及音節能量有顯著影響，因此將這些語言參數獨立出來表示，最後因考慮到不同語句時，說話速度上的變動會造成音長變化及說話音量變動會造成能量的變化，因此再將語句層次的正規化因子獨

立出來，而剩下之語言參數則定義為 reduced linguistic feature **l**。

為了清楚了解這些符號定義，將上述說明列在表 3.2。

表 3.2：韻律標記、韻律特徵和語言特徵的表示法

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	

綜合上述討論，可將(3.1)改寫為：

$$\begin{aligned} P(\mathbf{T},\mathbf{A}|\mathbf{L}) &= P(\mathbf{A}|\mathbf{T},\mathbf{L})P(\mathbf{T}|\mathbf{L}) = P(\mathbf{X},\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{PS},\mathbf{L})P(\mathbf{B},\mathbf{PS}|\mathbf{L}) \\ &\approx P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})P(\mathbf{PS}|\mathbf{B})P(\mathbf{B}|\mathbf{L}) \end{aligned} \tag{3.5}$$

由(3.5)式拆解出四個子模型，分別為音節韻律模型 $P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})$，用來說明音節韻律參數受到 **B**,**PS**,**L** 影響所產生的變化；停頓標記聲學模型 $P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$，用來說明在不同停頓標記 **B** 和語言參數 **L** 之下，音節間韻律參數 **Y** 及音節差韻律參數 **Z** 的分佈情形；韻律狀態轉移模型 $P(\mathbf{PS}|\mathbf{B})$，用來說明韻律狀態 **PS** 受到停頓標記 **B** 不同影響之變化；停頓標記語言模型

$P(\mathbf{B}|\mathbf{L})$，用來說明停頓標記 **B** 和語言特徵 **L** 之間的關係。

以下將進一步說明四個子模型，首先，音節韻律模型 $P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})$ 可再分解為三個模型，分別模擬音節音高軌跡序列 **sp**、音長序列 **sd** 和音節能量序列 **se**，並假設 **sp**、**sd** 和 **se** 的變化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 **PS**={**p**,**q**,**r**} 和停頓標記 **B**，因此可得到

$$\begin{aligned} P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L}) &\approx P(\mathbf{sp}|\mathbf{B},\mathbf{p},\mathbf{t})P(\mathbf{sd}|\mathbf{q},\mathbf{t},\mathbf{s},\mathbf{u})P(\mathbf{se}|\mathbf{r},\mathbf{t},\mathbf{f},\mathbf{u}) \\ &\approx \prod_{n=1}^{N} P(\mathbf{sp}_n|B_{n-1}^{n}, p_n, t_{n-1}^{n+1}) \prod_{n=1}^{N} P(sd_n|q_n,t_n,s_n,u_n) \prod_{n=1}^{N} P(se_n|r_n,t_n,f_n,u_n) \end{aligned} \tag{3.6}$$

(3.6)式中第一個子模型 $\prod_{n=1}^{N} P(\mathbf{sp}_n|B_{n-1}^{n}, p_n, t_{n-1}^{n+1})$ 為音節基頻軌跡模型，假設所觀察到的第 n 個音節基頻軌跡 $\mathbf{sp}_n$ 受到目前音高韻律狀態 p_n、目前聲調 t_n 以及在給定停頓標記 B_{n-1} 和 B_n 時，前後各一個音節聲調 t_{n-1} 和 t_{n+1} 造成的連音影響，因此 $B_{n-1}^{n}=(B_{n-1},B_n)$，$t_{n-1}^{n+1}=(t_{n-1},t_n,t_{n+1})$。如此一來可將音節基頻軌跡模型改寫成(3.7)式。

$$\mathbf{sp}_n = \mathbf{sp}_n^r + \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\beta}^{f}_{B_{n-1},tp_{n-1}} + \boldsymbol{\beta}^{b}_{B_n,tp_n} + \boldsymbol{\mu} \qquad \text{for } 1 \le n \le N \tag{3.7}$$

$\mathbf{sp}_n$ 為觀察到的第 n 個音節音高軌跡(observed)，此演算法是將音節音高軌跡使用正交展開(Orthogonal expansion)，投影到四個 Legendre 多項式基底得到四維正交化參數。(3.7)式中的 $\boldsymbol{\beta}_x$，表示音節音高軌跡影響因素 x 時的 AP，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 個音節所貢獻的前後音節影響效應的 AP。此外，每個語句的韻律邊界都有兩個特例，即為語句開始與結束，分別以 B_b 和 B_e 表示之，因此 $\boldsymbol{\beta}^{f}_{B_b,t_1}=\boldsymbol{\beta}^{f}_{B_0,tp_0}$，$\boldsymbol{\beta}^{b}_{B_e,t_N}=\boldsymbol{\beta}^{b}_{B_e,tp_N}$ 為兩個特例的連音效應 AP。$\mathbf{sp}_n^r$ 則為正規化後的 $\mathbf{sp}_n$，亦可稱為 $\mathbf{sp}_n$ 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 3.9 顯示出 $\mathbf{sp}_n$ 與這些影響因素之間的關係圖，藉由假設 $\mathbf{sp}_n^r$ 是一 zero-mean 的 normal distribution，即 $N(\mathbf{sp}_n^r;\mathbf{0},\mathbf{R})$，我們可以得到

$$P(\mathbf{sp}_n \mid p_n, B_{n-1}^{n}, t_{n-1}^{n+1}) = N(\mathbf{sp}_n; \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\beta}^{f}_{B_{n-1}, tp_{n-1}} + \boldsymbol{\beta}^{b}_{B_n, tp_n} + \boldsymbol{\mu}, \mathbf{R}) \quad \text{for } 1 \le n \le N \tag{3.8}$$

其中 **R** 定義為 $\mathbf{sp}_n^r$ 的共變數矩陣(covariance matrix)。

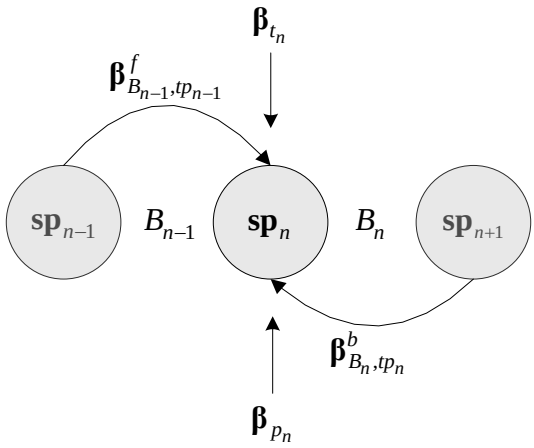


圖 3.9：觀察到的音節音高軌跡與其影響因素的關係圖

同理(3.6)式中第二、第三個子模型經過推導，可以得到:

$$P(sd_n \mid q_n, t_n, s_n, u_n) = N(sd_n; \gamma_{t_n} + \gamma_{q_n} + \gamma_{s_n} + \gamma_{u_n} + \mu_d, R_d) \tag{3.9}$$

$$P(se_n \mid r_n, t_n, f_n, u_n) = N(se_n; \alpha_{t_n} + \alpha_{r_n} + \alpha_{f_n} + \alpha_{u_n} + \mu_e, R_e) \tag{3.10}$$

(3.9)式模擬了音長 sd_n，其中 $\gamma's$ 定義各種不同的 APs，μ_d 與 R_d 分別表示 global mean 與音長殘餘值的共變數矩陣；(3.10)式模擬了音節能量 se_n，其中 $\alpha's$ 定義各種不同的 APs，μ_e 與 R_e 分別表示 global mean 與音節能量殘餘值的共變數矩陣。

接著停頓標記聲學模型 $P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$ 可進一步簡化，得到

$$P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L}) \approx P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{l}) \approx \prod_{n=1}^{N} P(pd_n, ed_n, pj_n, dl_n, df_n \mid B_n, \mathbf{l}_n) \tag{3.11}$$

其中 $P(pd_n, ed_n, pj_n, dl_n, df_n \mid B_n, \mathbf{l}_n)$ 是由分類樹與回歸樹(Classification and Regression Tree, CART)推導出來，其節點的分類準則是採用最大概似函數增益(Maximum Likelihood Gain)，將音節間停頓長度(pd_n)、音節間能量低點(ed_n)、相鄰兩音節之正規化基頻跳躍值(pj_n)、相鄰兩音節之正規化延長因子(dl_n 、df)，對於不同停頓標記，根據不同語言參數 **l** 進行分類，意即每

種停頓標記建立一顆決策樹，將每個節點裡的不同參數分別用不同 *pdfs* 來模擬，用一個 gamma distribution 來描述 pd_n 分佈情形，用四個 normal distribution 分別描述 ed_n、pj_n，及 dl_n、df 的分佈情形，因此(3.11)可改寫成五個機率分佈的乘積：

$$P(pd_n, ed_n, pj_n, dl_n, df_n \mid B_n, \mathbf{l}_n) = g(pd_n; \alpha_{B_n,\mathbf{l}_n}, \beta_{B_n,\mathbf{l}_n}) N(ed_n; \mu_{B_n,\mathbf{l}_n}, \sigma^2_{B_n,\mathbf{l}_n}) N(pj_n; \mu^{pj}_{B_n,\mathbf{l}_n}, \sigma^{2\,pj}_{B_n,\mathbf{l}_n}) \\ N(dl_n; \mu^{dl}_{B_n,\mathbf{l}_n}, \sigma^{2dl}_{B_n,\mathbf{l}_n}) N(df_n; \mu^{df}_{B_n,\mathbf{l}_n}, \sigma^{2df}_{B_n,\mathbf{l}_n}) \tag{3.12}$$

此外韻律狀態轉移模型可進一步針對三種韻律狀態分解成三個子模型，表示為：

$$P(\mathbf{PS}|\mathbf{B}) \approx P(\mathbf{p}|\mathbf{B}) P(\mathbf{q}|\mathbf{B}) P(\mathbf{r}|\mathbf{B}) \tag{3.13}$$

而每個子模型，可用雙連文模型(bigram models)分別表示為：

$$P(\mathbf{p}|\mathbf{B}) \approx P(p_1) \left[\prod_{n=2}^{N} P(p_n \mid p_{n-1}, B_{n-1}) \right] \tag{3.14}$$

$$P(\mathbf{q}|\mathbf{B}) \approx P(q_1) \left[\prod_{n=2}^{N} P(q_n \mid q_{n-1}, B_{n-1}) \right] \tag{3.15}$$

$$P(\mathbf{r}|\mathbf{B}) \approx P(r_1) \left[\prod_{n=2}^{N} P(r_n \mid r_{n-1}, B_{n-1}) \right] \tag{3.16}$$

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)，$P(p_n \mid p_{n-1}, B_{n-1})$、$P(q_n \mid q_{n-1}, B_{n-1})$ 和 $P(r_n \mid r_{n-1}, B_{n-1})$ 分別表示各個不同韻律狀態，在給定停頓標記 B_{n-1} 的情況下，從第 n-1 個音節的韻律狀態到第 n 個音節韻律狀態的轉移機率(transition probability)。

最後化簡停頓標記語言模型 $P(\mathbf{B}|\mathbf{L})$ 為：

$$P(\mathbf{B}|\mathbf{L}) \approx P(\mathbf{B}|\mathbf{l}) = \prod_{n=1}^{N-1} P(B_n \mid \mathbf{l}_n) \tag{3.17}$$

並且以最大概似函數增益為分裂準則之決策數來實現此模型，每一個節點中將產生每一種停頓標記之機率，其問題集由語言參數 $\mathbf{l}$ 所產生。

3.4 中英夾雜 PLM 演算法

原本的 PLM 演算法是以中文音節為單位所建構出的模型，而依據上述英文字母的特性，本研究將每個英文字母視為一種特殊音節，分成 27 類(M 拆成 M_r、M_f)，將英文字母如同中文音節進行音節層次和上層韻律層次的拆解，並標記其對應的停頓標記與韻律狀態，其停頓標記集合與韻律標記集合皆和中文 PLM 一樣。以下將介紹中英文夾雜 PLM 如何應用原先中文 PLM 的四個子模型，進行中英文夾雜的韻律模型訓練與韻律標記。

首先，第一個模型：音節韻律模型 $P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})$。由 3.3 節可知此模型可以再化簡成 3 個子模型，如前述的(3.8)式~(3.10)式，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**。針對英文字母特性，將三個子模型改寫成(3.18)~(3.20)式，分別為音節基頻軌跡模型、音節音長模型及音節能量模型。

$$P(\mathbf{sp}_n|p_n,B_{n-1}^{n},t_{n-1}^{n+1})=N(\mathbf{sp}_n;\boldsymbol{\beta}_{t_n}+\boldsymbol{\beta}_{p_n}+\boldsymbol{\beta}^{f}_{B_{n-1},tp_{n-1}}+\boldsymbol{\beta}^{b}_{B_n,tp_n}+\boldsymbol{\mu},\mathbf{R}) \tag{3.18}$$

其中，$B_{n-1}^{n}=(B_{n-1},B_n)$，$t_{n-1}^{n+1}=(t_{n-1},t_n,t_{n+1})$；$\boldsymbol{\beta}_{t_n}$ 為音節之聲調(t_n)影響參數，除了考慮原本中文有的 5 種聲調外，也將每個英文字母視為一種特定的聲調，因此聲調共有 32 類，$t_n\in\{1,2,3,4,5,\mathrm{A},\mathrm{B},....\mathrm{M_r}.....\mathrm{Z},\mathrm{M_f}\}$，$\boldsymbol{\beta}_{p_n}$ 為音節基頻之韻律狀態(p_n)影響參數，**μ** 為 global mean，**R** 為殘餘值之共變數矩陣，$\boldsymbol{\beta}^{f}_{B_{n-1},tp_{n-1}},\boldsymbol{\beta}^{b}_{B_n,tp_n}$ 為音節間前後連音影響參數。由於將聲調分為 32 種類別，連音參數若依據原先演算法之定義，將會造成 7168(32×32×7) 種組合，以致於在估算此參數時會因資料量太過稀少，進而影響整個連音參數的求取，因此本研究採取 CART 演算法，將語料庫中所有連音參數組合(包括英文接英文)，藉由適當問題集 Θ_1、Θ_2(詳情請見附錄一)，分別建立兩顆決策樹(Forward，Backward)，將相似連音參數進行合併縮減，以達到降低參數量並得到更可靠的連音參數。所採用的分裂準則為最大平方總和誤差降低量(Maximum Sum Square Error Reduction)，詳細實作流程將在 3.6 節中介紹。

$$P(sd_n|q_n,t_n,s_n,u_n)=N(sd_n;\gamma_{t_n}+\gamma_{s_n}+\gamma_{q_n}+\gamma_{u_n}+\mu_d,R_d) \tag{3.19}$$

其中 γ' 表示 sd_n 之 APs。同上述所說，我們將每個英文字母視為一種特定聲調，及 $t_n \in \{1,2,...5,A,B,....M_r.....Z,M_f\}$。$s_n$ 原為中文 411 基本音節類別，但由於本身此語料庫偏少，因此將中文聲母化簡成 7 類，中文韻母化簡成 13 類，因此原先 411 種組合被簡化成 91 類，在此語料庫下，$s_n \in \{1\sim 82類\}$，但當 sd_n 為英文字母時，γ_{s_n} 則為零，因為英文字母的影響已經由聲調影響參數來模擬。q_n 為音節音長之韻律狀態，γ_{u_n} 為音長語句之影響參數，μ_d 與 R_d 分別表示 global mean 與音長殘餘值的共變數矩陣。

$$P(se_n \mid r_n, t_n, f_n, u_n) = N(se_n; \alpha_{t_n} + \alpha_{f_n} + \alpha_{r_n} + \alpha_{u_n} + \mu_e, R_e) \tag{3.20}$$

其中 α' 表示 se_n 之 APs。同理，$t_n \in \{1,2,...5,A,B,....M_r.....Z,M_f\}$，$f_n \in \{1,2,...40\}$，共有 40 個韻母，但當遇到英文字母時，$\alpha_{f_n}$ 則為 0，r_n 為音節能量之韻律狀態，α_{u_n} 為音節能量語句之影響參數，μ_e,與 R_e 分別表示 global mean 與音節能量殘餘值的共變數矩陣。

第二個模型：停頓標記聲學模型 $P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$ 採用的聲學參數都和中文 PLM 一樣，模型數學式同(3.11)式與(3.12)式，一樣採取 CART 演算法，使用分類準則為最大概似函數增益，將音節間停頓長度(pd_n)、音節間能量低點(ed_n)、相鄰兩音節之正規化基頻跳躍值(pj_n)、相鄰兩音節之正規化延長因子(dl_n 、 df)，對於不同停頓標記，根據問題集 Θ_3，詳情請見附錄二)進行分類，即每種停頓標記建立一顆決策樹，將每個節點裡的不同參數分別用不同 *pdfs* 來描述其分佈情形，用一個 gamma distribution 來描述 pd_n 之分佈情形，用四個 normal distribution 分別描述 ed_n、pj_n、dl_n 及 df 之分佈情形。此停頓標記聲學模型與原先中文 PLM 中最大不同在於語言參數的不一樣，由於多了英文字母，必須特定考量在英文字母的情況下，該如何設計語言參數，如 code-switch 處音節停頓會比一般 Non-code switch 處來的長一點，因此英文語言參數的考量將會大大影響此停頓標記聲學模型的效果。

第三個模型：韻律狀態轉移模型 $P(\mathbf{PS}|\mathbf{B})$，同中文 PLM，將此模型拆解成三個 bigram models，分別描述三種韻律狀態與停頓標記之轉移關係，數學式同(3.14)~(3.16)。

第四個模型：停頓標記語言模型 $P(\mathbf{B}|\mathbf{L})$，同中文 PLM，採用 CART 演算法，以最大概

似函數增益為分裂準則，統計在不同語言參數下，所蒐集到的停頓標記分佈狀況，其問題集(Θ_3)同停頓標記聲學模型。而所使用之語言參數一樣需要多考量英文字母，如 code-switch 處因為有比較長的音節停頓長度，因此產生停頓標記 B2-2 的機率就相對較高。而停頓標記也會受到下一個音節的聲母型態而有所影響，因此將英文字母依據起頭發音方式，分類成如同中文聲母型態的分類，如表 3.3 所示：

表 3.3：英文字母分類表

INULL	A、E、M、N、I、O、R、U、V、Y、F、H、S、X、L
b(ㄅ)、d(ㄉ)、g(ㄍ)	B、D、W
sh(ㄕ) 、x(ㄒ)	C
zh(ㄓ)、z(ㄗ)、j(ㄐ)	G、J、Z
p(ㄆ)、t(ㄊ)、k(ㄎ)	P、T、K、Q

其中 F、H、S、X、L 因起始發音也是濁音(voiced)，因此歸類在 INULL。

3.5 中英文夾雜韻律模型之訓練

本節將介紹如何實作 3.4 節所提出之演算法，中英夾雜 PLM 在建立時，是基於最大概似度準則，採用逐項最佳化程序來更新模型參數，標記出最佳的韻律標記，其標的函數(objective function)如下：

$$Q=\left(\prod_{n=1}^{N}P(\mathbf{sp}_n\mid p_n,B_{n-1}^{n},t_{n-1}^{n+1})P(sd_n\mid q_n,t_n,s_n,u_n)P(se_n\mid r_n,t_n,f_n,u_n)\right)$$
$$\left(P(p_1)P(q_1)P(r_1)\prod_{n=2}^{N}P(p_n\mid p_{n-1},B_{n-1})P(q_n\mid q_{n-1},B_{n-1})P(r_n\mid r_{n-1},B_{n-1})\right) \quad (3.21)$$
$$\left(\prod_{n=1}^{N-1}\left(p(pd_n,ed_n,pj_n,dl_n,df_n\mid B_n,\mathbf{l}_n)P(B_n\mid\mathbf{l}_n)\right)\right)$$

模型訓練的過程主要分成兩個步驟，為初始化(initialization)及重覆疊代(iteration)。將於 3.5.1 節和 3.5.2 節詳細介紹其流程。

3.5.1 初始化(Initialization)

Step1：計算總體平均值

計算音節基頻、音節長度、音節能量的總體平均值$\boldsymbol{\mu}$、μ_d、μ_e，其中$\mathbf{\mu}$並不包含求不出基頻的音節。

Step 2：計算聲調影響參數($\boldsymbol{\beta}_t$、γ_t、α_t)

分別計算聲調對音節基頻、音節長度、音節能量的影響參數，英文字母的影響歸類在聲調中，即每種英文字母看成一種聲調，計算公式如下:

$$\boldsymbol{\beta}_t=\frac{\sum_{n=1}^{N}(\mathbf{sp}_n-\boldsymbol{\mu})\delta(t_n=t)}{\sum_{n=1}^{N}\delta(t_n=t)},\quad \text{for } t=1,\cdots,5,\text{A,B...M}_\text{r}\text{...Z,M}_\text{f} \quad (3.22)$$

$$\gamma_t=\frac{\sum_{n=1}^{N}(sd_n-\mu_d)\delta(t_n=t)}{\sum_{n=1}^{N}\delta(t_n=t)},\quad \text{for } t=1,\cdots,5,\text{A,B...M}_\text{r}\text{...Z,M}_\text{f} \quad (3.23)$$

$$\alpha_t = \frac{\sum_{n=1}^{N}(se_n - \mu_e)\delta(t_n = t)}{\sum_{n=1}^{N}\delta(t_n = t)}, \quad \text{for } t = 1,\cdots,5,\text{A,B...M}_\text{r}\text{...Z,M}_\text{f} \tag{3.24}$$

Step 3：計算基本音節型態影響參數(γ_{s_n}、α_{f_n})

分別計算不同基本音節類型對於音節長度的影響，不同韻母類型對於音節能量的影響，基本音節型態由 411 種化簡成 91 總組合(7 種聲母類別×13 種韻母類別，此語料共只出現 82 類)，韻母則為分為 40 類，而在當前音節為英文時，這兩個 APs 都給 0。計算公式如下:

$$\gamma_s = \frac{\sum_{n=1}^{N}(sd_n - \mu_d - \gamma_t)\delta(s_n = s)}{\sum_{n=1}^{N}\delta(s_n = s)}, \quad \text{for base syllable type } s \tag{3.25}$$

$$\alpha_f = \frac{\sum_{n=1}^{N}(se_n - \mu_e - \alpha_t)\delta(f_n = f)}{\sum_{n=1}^{N}\delta(f_n = f)}, \quad \text{for final type } f \tag{3.26}$$

Step 4：標記初始化停頓標記(Initial labeling of break indices)

利用表 3.2 中的**Y**、**Z** 參數，包含了音節停頓長度(**pd**)，音節能量低點(**ed**)，正規化基頻跳躍值(**pj**)以及正規化音節延長因子(**dl**、**df**)，使用【10】提出之決策樹的方式，對所有音節邊界處標記初始的停頓型態(**B**)，如圖 3.10 所示。由於音節停頓時長是判斷是韻律邊界一個重要的聲學參數，而音節後為標點符號(PM)的邊界通常會有較大的停頓時長，因此往往屬於本研究所定義之 B3 與 B4。其次，大多數的詞外音節邊界有較短的停頓時長，通常被標記成本研究所定義之 B0 與 B1，然而 B0 是屬於音節間基頻停頓(pitch pause duration)很短的停頓標記，因此藉由很短的 pitch pause duration 和很高的音節能量低點區分 B0 和 B1。此外在 Non-PM 的詞外音節邊界中有中等程度以上的停頓時長、基頻跳躍值及音長延長，則分別歸類為 B2-2、B2-1 與 B2-3。藉由上述所說之語言參數與聲學參數的關係，我們可以制定一套演算法來決定此決策樹中的 threshold *Th*1~*Th*8，並透過這些 threshold *Th*1~*Th*8 自動化得到初始停頓標記，避免使用人工標記所帶來的不一致性或是錯誤標記。決定 threshold *Th*1~*Th*8 之

演算法細節請參閱附錄三。

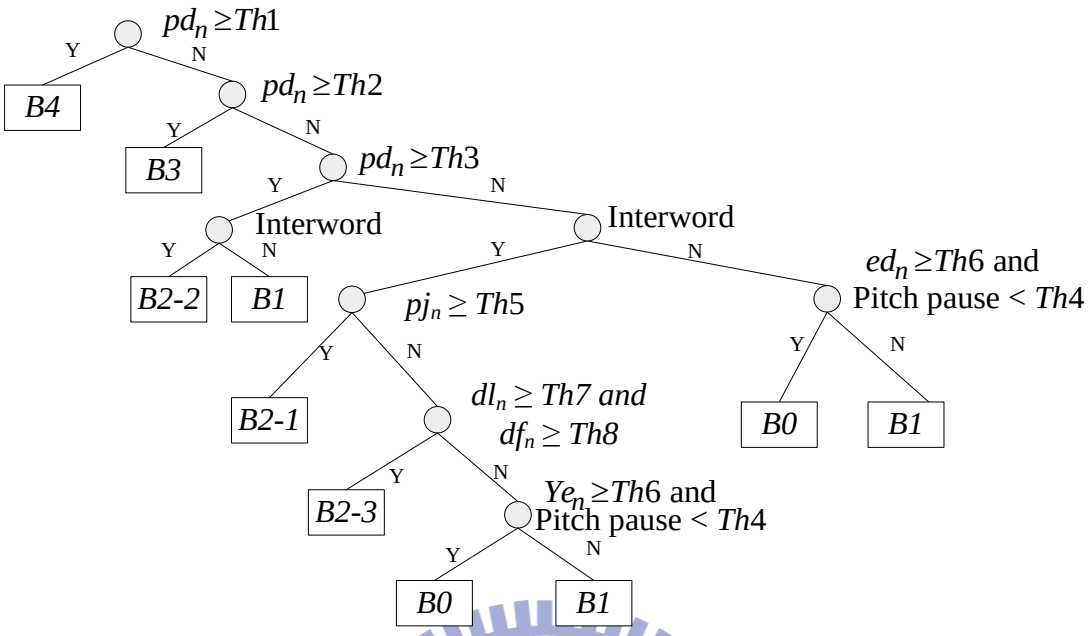


圖 3.10：Break Type 分類決策樹示意圖

Step 5：計算初始化連音影響參數($\boldsymbol{\beta}^{f}_{B_{n-1},tp_{n-1}}$,$\boldsymbol{\beta}^{b}_{B_n,tp_n}$)

因為語料庫太小且前後音節聲調組合(包含英文字母 27 種類別)太多的情況下，本研究採取 CART 演算法將分別建立兩顆決策樹(Forward、Backward)將特性相似的組合進行合併。首先，由數學式(3.27)與(3.28)計算訓練語料中所有組合的連音參數作為 CART 的母體(sample space)。

$$\boldsymbol{\beta}^{f}_{B,tp}=\frac{\sum_{n=2}^{N}\mathbf{sp}_n\delta(tp_{n-1}=tp)\delta(B_{n-1}=B)}{\sum_{n=2}^{N_k}\delta(tp_{n-1}=tp)\delta(B_{n-1}=B)}-\frac{\sum_{n=2}^{N}\mathbf{sp}_n\delta(t_n=j)\delta(B_{n-1}=B)}{\sum_{n=2}^{N}\delta(t_n=j)\delta(B_{n-1}=B)} \tag{3.27}$$

$$\boldsymbol{\beta}^{b}_{B,tp}=\frac{\sum_{n=1}^{N-1}\mathbf{sp}_n\delta(tp_n=tp)\delta(B_n=B)}{\sum_{n=1}^{N-1}\delta(tp_n=tp)\delta(B_n=B)}-\frac{\sum_{n=1}^{N-1}\mathbf{sp}_n\delta(t_n=i)\delta(B_n=B)}{\sum_{n=1}^{N-1}\delta(t_n=i)\delta(B_n=B)} \tag{3.28}$$

其中 $tp_n=(t_n,t_{n+1})$ ，$t_n\in\{1,2,3,4,5,A,B,....M_t.....Z,M_f\}$，若遇到特殊音節邊界($B_b$、$B_e$)則使用下列二式計算:

$$\boldsymbol{\beta}^{f}_{B_b,t}=\frac{\sum_{k=1}^{K}\mathbf{sp}_{k,1}\delta(t_{k,1}=t)}{\sum_{k=1}^{K}\delta(t_{k,1}=t)}-\boldsymbol{\beta}_t(t=t) \tag{3.29}$$

$$\boldsymbol{\beta}^{b}_{B_e,t}=\frac{\sum_{k=1}^{K}\mathbf{sp}_{k,N_k}\delta(t_{k,N_k}=t)}{\sum_{k=1}^{K}\delta(t_{k,N_k}=t)}-\boldsymbol{\beta}_t(t=t) \tag{3.30}$$

且令$\boldsymbol{\beta}^{f}_{B_b,t}$、$\boldsymbol{\beta}^{b}_{B_e,t}$的第一維係數為 0。其物理意義為音節邊界在句首或是句尾時，基頻軌跡的變化程度和整體平均值差了多少。

得到欲分類的母體後，根據問題集Θ_1、Θ_2，採用分裂準則為最大平方總和誤差降低量，定義在 node k 的情況下，Sum Square Error(SSE)如下:

$$\mathrm{E}(k)=\frac{\sum_{\hat{\boldsymbol{\beta}}^{f}_{B,tp}\in k}|\hat{\boldsymbol{\beta}}^{f}_{B,tp}-\mu|^2}{\sum_{N_{B,tp}\in k}N_{B,tp}} \tag{3.31}$$

其中$\mu=\frac{\sum_{\hat{\boldsymbol{\beta}}^{f}_{B,tp}\in k}\hat{\boldsymbol{\beta}}^{f}_{B,tp}\cdot N_{B,tp}}{\sum_{N_{B,tp}\in k}N_{B,tp}}$，$N_{B,tp}$為落在為落在 node k 的個數，$\hat{\boldsymbol{\beta}}^{f}_{B,tp}$為落在為落在 node k 的 Forward 連音參數(在此以 Forward 為例，Backward 同理)。 因此分裂準則(MSSER)之數學式為：

$$q^{*}=\arg\max_{q\in Q}\Delta\mathrm{E}_k(q)=\mathrm{E}(k)-(\ \mathrm{E}_{\mathrm{Y}}(k\mid q)+\mathrm{E}_{\mathrm{N}}(k\mid q)\ ) \tag{3.32}$$

其物理意義為尋找一個最適合問題(q^{*})使得 SSE 下降最多。其中: Y、N 為 node k 的子節點。根據以上數學式我們可以建立兩顆決策樹，將母體中所有連音參數進行分類，分類的組合數即為決策樹最後的葉節點個數，並將每個葉節點裡所蒐集到的連音參數取平均值，代表初始化的連音參數。

Step 6：計算初始化的韻律狀態($\boldsymbol{\beta}_p$、γ_q、α_r)

將音節基頻、音節長度以及音節能量分別扣掉先前所計算出來的影響參數後，將剩餘殘存值以向量量化(VQ)的方式計算音節基頻、音節長度以及音節能量之每一個韻律狀態的值$\boldsymbol{\beta}_p$、γ_q、α_r，並標記每個音節所屬的韻律狀態，而各剩於殘存值定義如下:

$$\begin{aligned}
\mathbf{sp}_n^r &= \mathbf{sp}_n - (\boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}^f_{B_{n-1},tp_{n-1}} + \boldsymbol{\beta}^b_{B_n,tp_n} + \boldsymbol{\mu}) \\
sd_n^r &= sd_n - (\gamma_{t_n} + \gamma_{s_n} + \mu_d) \\
se_n^r &= se_n - (\alpha_{t_n} + \alpha_{f_n} + \alpha_{r_n} + \mu_e)
\end{aligned} \tag{3.33}$$

其中，本研究將每種韻律狀態分為 16 種。

Step 7：內插法求取原先基頻為零的音節

為了避免求不出基頻的音節會影響到正規化基頻跳躍值(**pj**)的求取，進而在往後疊代過程中影響整個韻律模型，因此先透過內插法求取其韻律狀態的值($\boldsymbol{\beta}_p$)，再利用其他影響參數計算這些原本基頻為零的音節，給一個概略的基頻值($\mathbf{sp}'_n$)，避免完全是零的情況下，**pj** 變的超大，讓整個模型都被這些誤差干擾到。計算數學式如下：

$$\mathbf{sp}'_n = (\boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}^f_{B_{n-1},tp_{n-1}} + \boldsymbol{\beta}^b_{B_n,tp_n} + \boldsymbol{\beta}_p + \boldsymbol{\mu}) \tag{3.34}$$

Step 8：求取韻律狀態轉移模型($P(\mathbf{PS}|\mathbf{B})$)

將標記完的韻律狀態和停頓標記以相對計數率(relative count)的方式，計算在每種停頓標記下，韻律狀態的轉移機率。

Step 9：求取停頓標記語言模型($P(\mathbf{B}|\mathbf{L})$)

採用 CART 演算法實作出停頓標記語言模型之決策樹。以最大概似函數增益為分裂準則，區分各種停頓標記。在停頓標記語言模型中，在每一個節點裡使用相對計數率的方式產生每一種停頓標記之機率，即

$$P(B|i)=C(B|i)\Big/\sum_{B'\in\mathbf{B}}C(B'|i) \tag{3.35}$$

$P(B|i)$ 為落在第 i 個葉節點中停頓標記 B 的機率，$C(B|i)$ 為落在第 i 個葉節點中停頓標記 B 的個數。

Step 10：求取停頓標記聲學模型($P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$)

如同 Step 9 一樣，採取 CART 演算法實作停頓標記聲學模型，也是採取最大概似函數增益為分裂準則，區分 pd_n 、 ed_n 、 pj_n 、 dl_n 及 df_n ，並且以每一類停頓標記建立一顆決策樹，在每個節點分別將上述參數 pd_n 之伽瑪分佈 $g(pd_n;\alpha_{B_n,\mathbf{l}_n},\beta_{B_n,\mathbf{l}_n})$ ， ed_n 、 pj_n 、 dl_n 、 df_n 之高斯分佈 $N(ed_n;\mu_{B_n,\mathbf{l}_n},\sigma^2_{B_n,\mathbf{l}_n})$、 $N(pj_n;\mu^{pj}_{B_n,\mathbf{l}_n},\sigma^{2\,pj}_{B_n,\mathbf{l}_n})$ $N(dl_n;\mu^{dl}_{B_n,\mathbf{l}_n},\sigma^{2dl}_{B_n,\mathbf{l}_n})$、 $N(df_n;\mu^{df}_{B_n,\mathbf{l}_n},\sigma^{2df}_{B_n,\mathbf{l}_n})$ 相乘來表示。其問題集(Θ_3)基本上是採用和停頓標記語言模型一樣，但另外依據不同停頓標記來進行問問題，例如 B2-2、B3、B4 不會去問有關基本音節型態的問題，意即不會在音節邊界為 B3 的情況下，問下一個音節的子音型態。同理，B0、B1 也不會去問這個邊界是否在逗號處。

Step 11：求取音節韻律模型之共變異數矩陣

包含了音節基頻軌跡模型(**sp**)、音長模型(**sd**)以及音節能量(**se**)，經由扣除所有影響基頻、音節長度以及能量之 APs 後，將其殘存值以一個平均值為零的高斯分佈來描述它，共變異數矩陣分別如下:

$$\mathbf{R}=\frac{\sum_{n=1}^{N}\mathbf{sp}_n^r(\mathbf{sp}_n^r)^T}{N} \tag{3.36}$$

$$R_d=\frac{\sum_{n=1}^{N}sd_n^r(sd_n^r)^T}{N} \tag{3.37}$$

$$R_e=\frac{\sum_{n=1}^{N}se_n^r(se_n^r)^T}{N} \tag{3.38}$$

其中：

$$\mathbf{sp}_n^r = \mathbf{sp}_n - (\boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\beta}^f_{B_{n-1},tp_{n-1}} + \boldsymbol{\beta}^b_{B_n,tp_n} + \boldsymbol{\mu}) \tag{3.39}$$

$$sd_n^r = sd_n - (\gamma_{t_n} + \gamma_{s_n} + \gamma_{q_n} + \mu_d) \tag{3.40}$$

$$se_n^r = se_n - (\alpha_{t_n} + \alpha_{f_n} + \alpha_{r_n} + \mu_e) \tag{3.41}$$

3.5.2 重覆疊代(Iteration)

有了初始化模型後，接著進行重覆疊代，使得模型標的函數達到收斂，確保得到可靠的韻律標記及其他模型參數。詳細步驟如下：

Step 1：更新聲調影響參數($\boldsymbol{\beta}_t$、γ_t、α_t)，且其它 APs 固定。

Step 2：藉由初始化中 Step 5 的 CART 分類結果，更新連音影響參數($\boldsymbol{\beta}^f_{B,tp}$、$\boldsymbol{\beta}^b_{B,tp}$、$\boldsymbol{\beta}^f_{B_b,t}$、$\boldsymbol{\beta}^b_{B_e,t}$)，且其它 APs 固定。

利用初始化的結果，扣掉其餘 APs，以 Forward 為例，扣除聲調、基頻韻律狀態、Backward 連音參數及基頻平均值等 APs，計算公式如(3.42)式，我們可得到更精準的 Forward 連音參數($\hat{\boldsymbol{\beta}}^f_{B_{n-1},tp_{n-1}}$)，並以這些連音參數再重新建立一顆決策樹，得到更新的分類，同理 Backward 可由(3.43)式更新連音參數($\hat{\boldsymbol{\beta}}^b_{B_n,tp_n}$)並得到新的分類，最後再由這些組合一起更新連音參數。

$$\hat{\boldsymbol{\beta}}^f_{B_{n-1},tp_{n-1}} = \mathbf{sp}_n - (\boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\beta}^b_{B_n,tp_n} + \boldsymbol{\mu}) \tag{3.42}$$

$$\hat{\boldsymbol{\beta}}^b_{B_n,tp_n} = \mathbf{sp}_n - (\boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\beta}^f_{B_{n-1},tp_{n-1}} + \boldsymbol{\mu}) \tag{3.43}$$

Step 3：更新基本音節型態影響參數(γ_s、α_f)，且其它 APs 固定。

Step 4：更新共變異數矩陣($\mathbf{R}$、R_d、R_e)。

Step 5：利用維特比搜尋演算法(Viterbi search algorithm)重新標記韻律狀態，使得(3.21)式之 Q

最大。

Step 6：更新韻律狀態($\boldsymbol{\beta}_p$、γ_q、α_r)，並且更新韻律狀態轉移模型($P(\mathbf{PS}|\mathbf{B})$)和再次更新共變異數矩陣($\mathbf{R}$、$R_d$、$R_e$)。

Step 7：利用維特比搜尋演算法重新標記停頓標記，使得(3.21)式之Q最大，並且再次更新韻律狀態轉移模型($P(\mathbf{PS}|\mathbf{B})$)和共變異數矩陣($\mathbf{R}$、$R_d$、$R_e$)。

Step 8：使用 CART 演算法以及問題集Θ_3，重新建立停頓標記聲學模型($P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$)和停頓標記語言模型($P(\mathbf{B}|\mathbf{L})$)之決策樹。

Step 9：重複 step 1~step 8 直到收斂為止。

第四章 韻律模型訓練結果與分析

本實驗訓練語料為中英夾雜語料庫 539 句中，拿掉編號尾數為 7 之句子，共 485 句，總音節數為 12185 個音節，包含 10504 個中文音節與 1681 個英文字母。採逐項最佳化程序重覆疊代 51 次收斂，其對應之標的函數值(total likelihood of objective function)如圖 4.1 所示。在本章節將介紹四個子模型之參數觀察結果以及韻律標記之分析。4.1 節將介紹音節韻律模型，包含各影響參數對音節韻律的貢獻與影響，4.2 節將分析停頓標記聲學模型的結果，4.3 節將分析韻律狀態轉移模型的結果，4.4 節則是分析停頓標記語言模型的結果，最後，4.5 節則是分析韻律標記的結果。

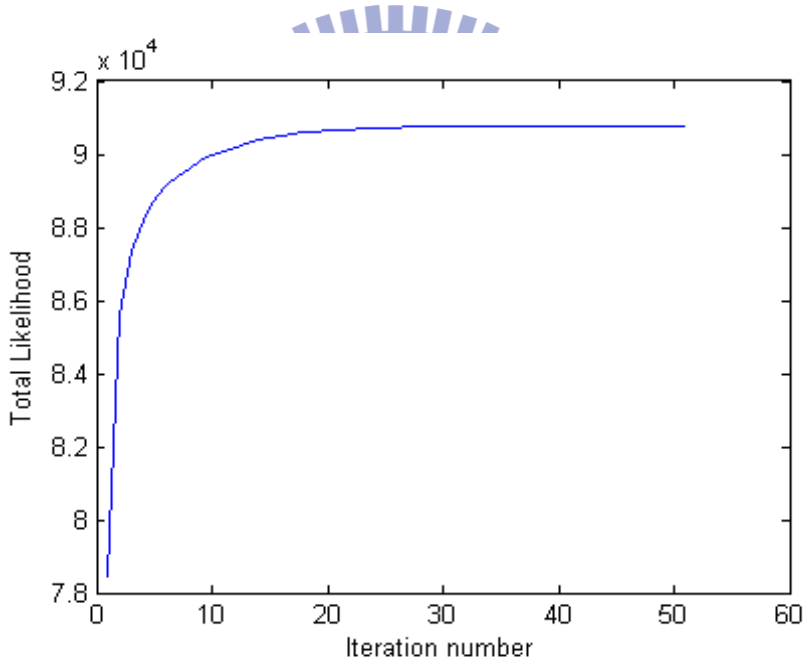


圖 4.1：疊代次數與標的函數值

4.1 音節韻律模型

音節韻律模型又可分成三個子模型，分別描述音節基頻軌跡、音節長度以及音節能量受到各個影響因素的變化與影響。首先觀察三個子模型訓練前後之共變數矩陣，分別如下：

$$\mathbf{R}_{\mathbf{sp}}\begin{bmatrix}421.7 & 21.3 & -22.2 & -1.53\\ 21.3 & 98.6 & 4.61 & -4.84\\ -22.2 & 4.61 & 16.2 & -1.21\\ -1.53 & -4.84 & -1.21 & 5.98\end{bmatrix}\times 10^{-4} \Rightarrow \mathbf{R}_{\mathbf{sp}^r}=\begin{bmatrix}2.95 & 0.26 & -0.04 & -0.03\\ 0.26 & 23.4 & 1.55 & 0.3\\ -0.04 & 1.55 & 9.93 & 0.02\\ -0.03 & 0.3 & 0.02 & 4.53\end{bmatrix}\times 10^{-4}$$

$$R_{sd}=402.2\times 10^{-5} \quad \Rightarrow \quad R_{sd^r}=1.71\times 10^{-5}$$

$$R_{se}=31.09 \quad \Rightarrow \quad R_{se^r}=0.1949$$

由以上數據可發現各個模型之共變數矩陣之值在收斂前後都明顯變小，也證明的確可以使用各個影響因素來描述音節中韻律參數的變化。接著表 4.1、4.2 及 4.3 則分別列出所有音節、只計算中文音節以及只計算英文字母在扣除不同影響因素(APs)之下，各韻律參數的總殘餘誤差值(Total Residual Error, TRE)，意即扣除各種 APs 後，殘餘值之變異數與原始資料之變異數的比值，藉此觀察各種 APs 對音節中各韻律參數變化的影響程度。由表 4.2 可以得知中文音節之基頻軌跡的變化貢獻由大到小分別為韻律狀態、聲調以及連音現象；而中文之音節長度以及音節能量的變化貢獻由大至小分別為：韻律狀態、基本音節型態、聲調以及語句。至於英文字母之各 AP 貢獻程度則可由觀察表 4.3 得知，可發現英文字母之基頻軌跡變化貢獻由大至小分別為韻律狀態、英文字母型態以及連音現象，此外也可發現到連音現象對英文字母的影響程度其實不大，此部分會在 4.1.1.1 節中深入討論。而英文字母之音節長度以及音節能量的變化貢獻由大到小分別為：韻律狀態、英文字母型態以及語句。

表 4.1：所有音節在不同 APs 下音節韻律模型參數之 TRE

Pitch		**Duration**		**Energy**	
APs	TRE	APs	TRE	APs	TRE
		+Utterance	97.3%	+Utterance	87.7%
+Tone	65.1%	+Tone	80.5%	+Tone	79.8%
+Coarticulation	56.8%	+Base-syllable	55%	+Final	58.4%
+Prosodic state	7.53%	+Prosodic state	4.3%	+Prosodic state	0.63%

表 4.2：中文音節在不同 APs 下音節韻律模型參數之 TRE

Pitch		**Duration**		**Energy**	
APs	TRE	APs	TRE	APs	TRE
		+Utterance	98%	+Utterance	88.5%
+Tone	66.5%	+Tone	86.2%	+Tone	82%
+Coarticulation	57.8%	+Base-syllable	56.7%	+Final	58.9%
+Prosodic state	7.64%	+Prosodic state	4.3%	+Prosodic state	0.62%

表 4.3：英文字母在不同 APs 下音節韻律模型參數之 TRE

Pitch		**Duration**		**Energy**	
APs	TRE	APs	TRE	APs	TRE
		+Utterance	97.4%	+Utterance	78%
+Tone (English type)	47.7%	+ Tone (English type)	42.6%	+Tone (English type)	53.1%
+Coarticulation	45.3%				
+Prosodic state	6.22%	+Prosodic state	3.7%	+Prosodic state	0.75%

4.1.1.1 音節層次中基頻之影響型態

首先我們先觀察聲調(包含不同英文字母)對基頻軌跡的影響，由圖 4.2 可看到中文五種聲調對中文基頻軌跡的影響，此圖也與第三章中的圖 3.1(過去學者研究中文聲調之基頻軌跡)相符合，其中也可觀察到三聲和五聲的聲調影響因子之基頻軌跡雷同，但五聲之基頻軌跡本身容易受到前後音節聲調的影響，因此本研究語料庫剛好有此現象。圖 4.3 則為不同英文字母型態對英文字母之基頻軌跡影響，其中圖(a)為有 Tone 1 Borrowing 現象的英文字母基頻軌跡，圖(b)為剩餘其他英文字母，觀察後可發現英文字母 Tone 1 Borrowing 的情形最多，有{A、B、C、D、E、G、I、J、K、N、O、P、Q、T、U、V、Y}，Tone 2 Borrowing 有{ F、H、L、M_r、R、S、X }；Tone 4 Borrow 有{M_f}，剩餘字母 W、Z 則不具有 Tone Borrowing 的現象。以上結果也再次驗證 3.1 節中所討論的英文 Tone Borrowing 現象。

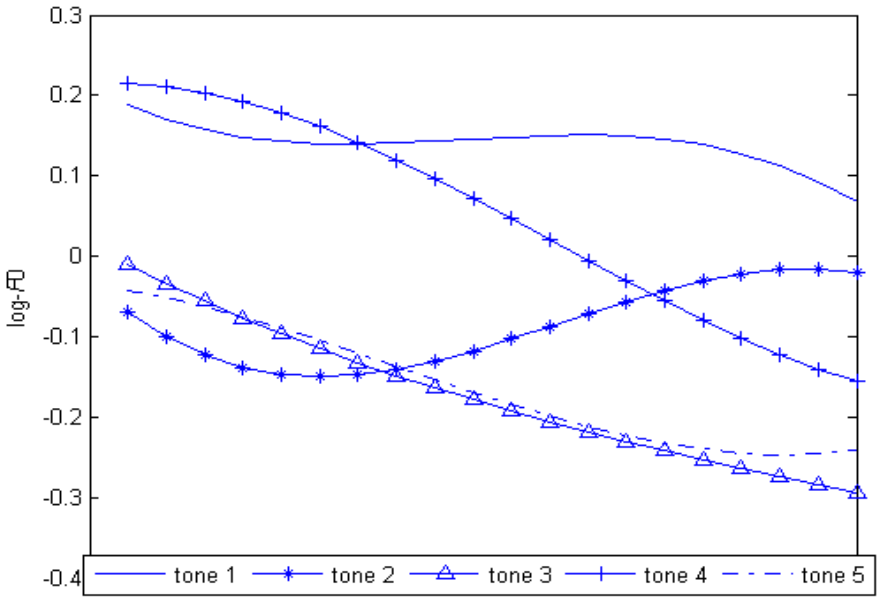


圖 4.2：基頻之中文五種聲調 AP

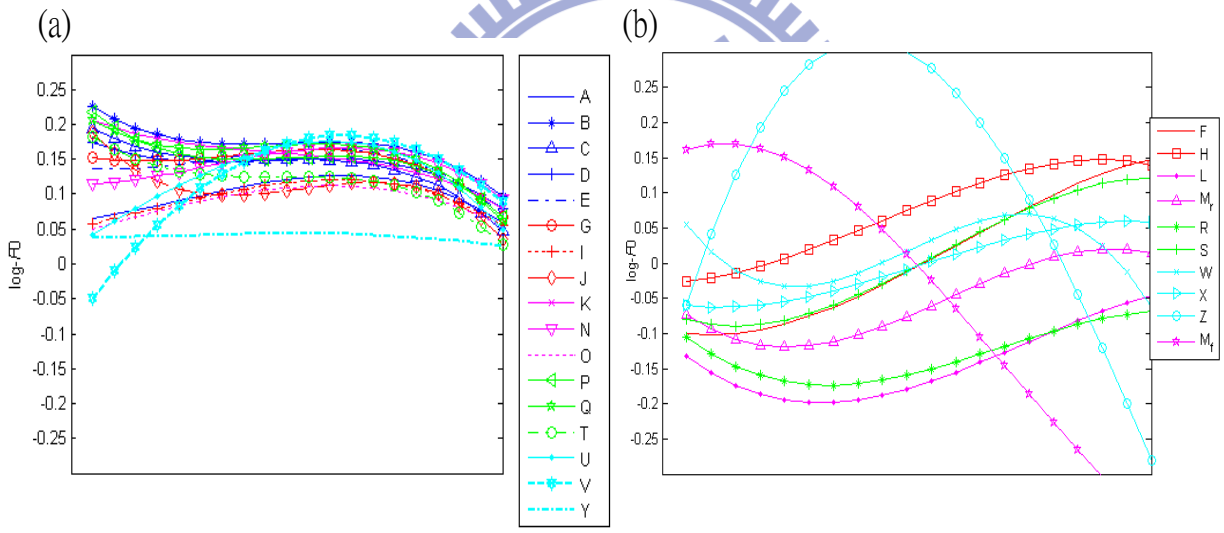


圖 4.3：基頻之英文字母型態 AP，(a)英文字母為 Tone1 Borrowing，(b)剩餘英文字母

此外圖 4.4 為中文聲調、英文字母型態對基頻 AP 之第一維數值，對中文而言，以一聲和四聲的基頻平均值較高，而有 Tone 1 Borrowing 現象的英文字母其基頻平均值也較高，同理中文二聲和 Tone 2 Borrowing 的英文字母其基頻平均值都偏低。

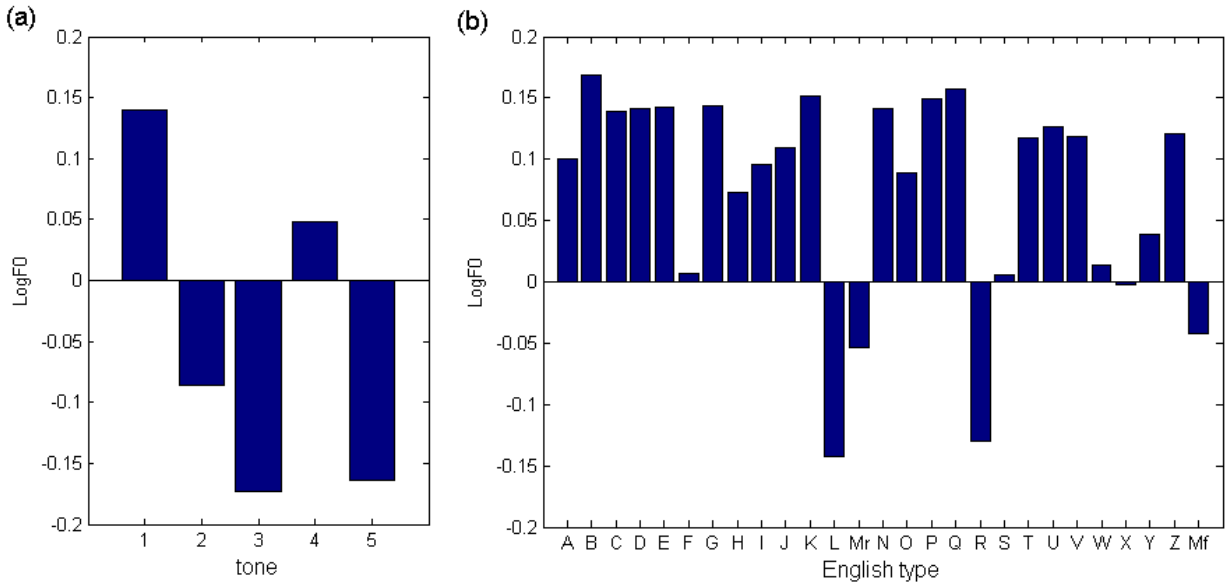


圖 4.4：(a)中文五種聲調和(b)英文型態基頻 AP 之第一維數值

最後觀察連音影響參數對中、英文之音節基頻影響程度。透過 CART 演算法及適當的問題集將原本高達 7168(32×32×7)種組合的連音影響參數(Forward、Backward)分別大幅化簡成 79 類和 57 類，而在此補充說明問題集設計的概念，本研究利用前後音節基頻起始與節尾高度的不同，搭配停頓標記的強度，進行問題集的設計，其示意圖如圖 4.5 所示，圖中 1~5 為中文聲調一到五聲，E-t1 為 Tone 1 Borrowing 之英文字母，E-t2 為 Tone 2 Borrowing 之英文字母，我們將當前這個音節相似的英文基頻軌跡合併看待，如 Tone 1 Borrowing 之英文字母，並考慮將相似中英文之聲調合併，如中文一聲和 Tone 1 Borrowing 之英文字母，亦或單獨只考慮英文，而考慮下一個音節影響時，則是將下一個音節基頻起始高度相似的合併，如中文一聲、中文四聲、Tone 1 Borrowing 之英文字母與 M_f。此外由於中文三聲接中文三聲有特殊變調現象，故特別獨立分開處理。同理在考慮前一個音節影響時，將前一個音節基頻結尾高度相似的合併在一起。如中文 2 聲和 Tone 2 Borrowing 之英文字母。而停頓標記的合併則是使用下列分類：{B0}、{B0,B1}、{B0,B1,B2-1}、{B0,B1,B2-1,B2-3}、{B2-2}、{B2-2,B3,B4}。使用上述問題集的好處在於，若是英文字母之連音影響參數和中文一樣強烈，自然會和中文合併在一起，反之則應會獨立出來，同理若在不同停頓標記下，其連音影響程度差不多，自然不會被分開。因此透過問題集的設計與 CART 演算法，我們可以將原本

大量的連音參數組合進行自動化的合併，達到化簡又不失一般性的效果。

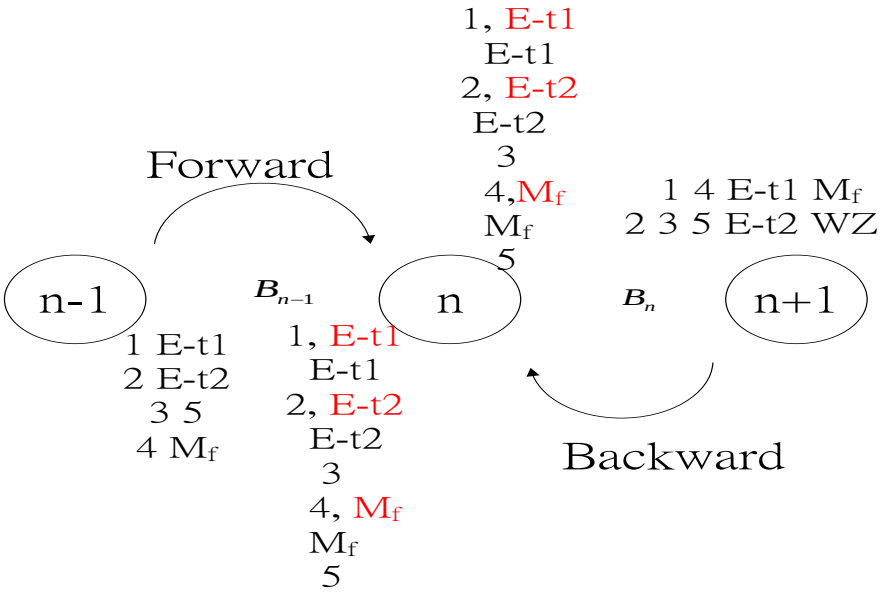


圖 4.5：連音參數之問題集設計概念

接下來透過幾張圖來觀察連音現象的一些分類結果，首先圖 4.6 為 Backward 之決策樹結果，每個節點都有對應的節點編號和其問題，編號 1 即代表根節點(root node)，每個節點內的圖示為這組分類的連音型態，另外實線表示父節點問題為“是”，虛線則表示父節點問題為“否”。由圖(a)之節點 1 可知道第一個被問到的問題為中文三聲接中文三聲的變調現象，此結果也是預料中的事，因為變調為特殊現象，和其他聲調組合之連音現象明顯不同，故首先被分類出來，由節點 2 之問題，可知又將變調現象以停頓標記{B0}和{B1,B2-1,B2-3}區分開來，代表在 B0 之下的變調現象又來得比其他停頓標記更強。觀察節點 7 之問題到節點 15、18 及 19 之結果，可發現到一開始節點 7 先將中文二聲和 Tone 2 Borrowing 之英文字母合併一起詢問下一個音節是否都是基頻起始較低的中文聲調或英文字母，然而節點 10 則是再將中、英文分開來詢問，代表 Tone 2 borrowing 之英文字母雖然有和中文二聲相似的基頻軌跡，但其連音現象並沒有和中文二聲相仿。類似之現象可由圖(b)之節點 57 觀察到，節點 57 先將中文一聲和 Tone 1 Borrowing 之英文字母一起詢問下一個音節是否為基頻軌跡起始較高的中文或是英文字母，然而往下走到節點 58 時，又將其拆開詢問，代表 Tone 1 Borrowing 之英文字母並不會擁有像中文一聲那麼強烈的連音現象。

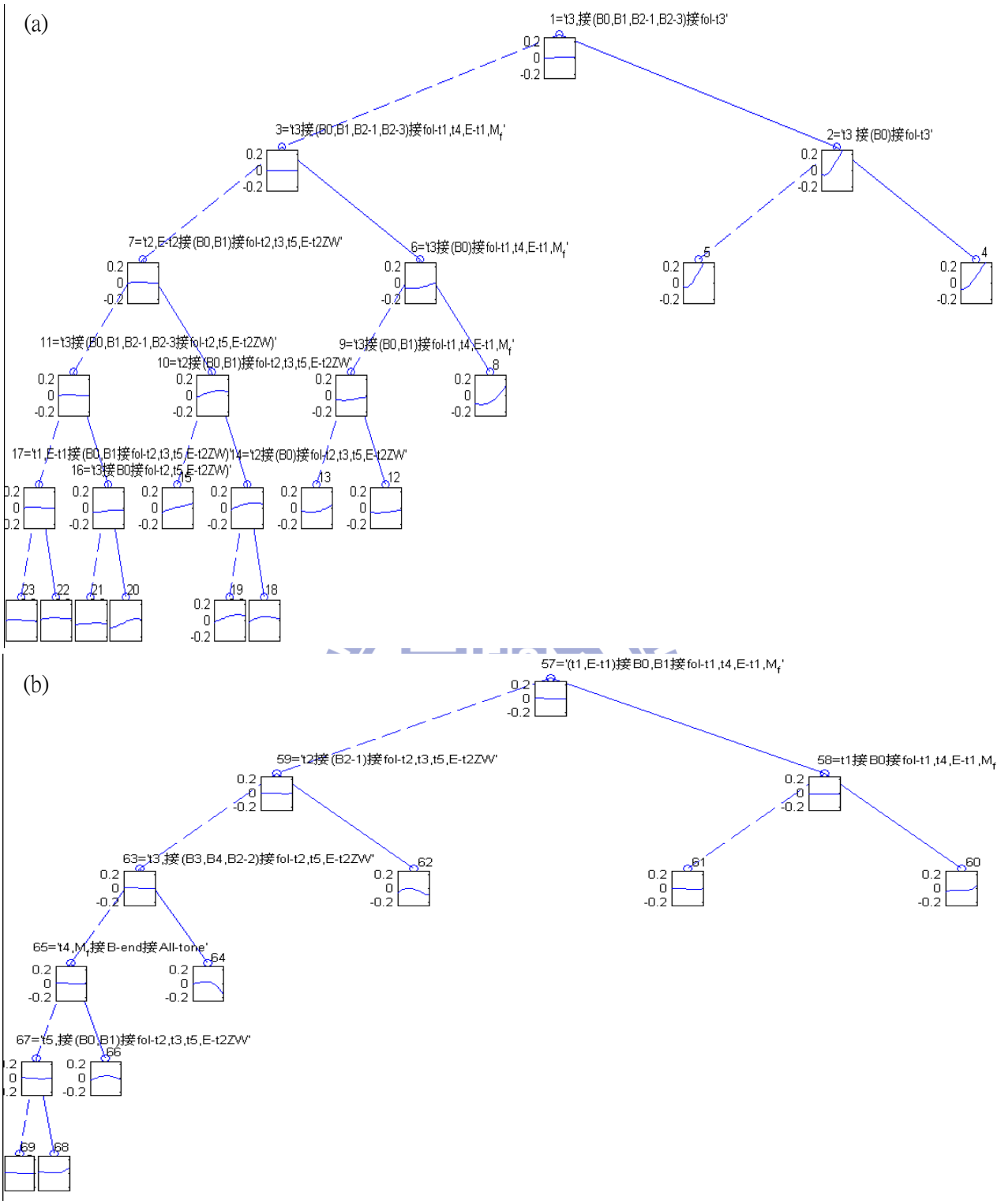


圖 4.6：Backward 連音參數之決策樹

接著圖 4.7 為 Forward 連音參數之決策樹部分結果，觀察節點 33、34 及 37 所詢問之問題，可發現節點 33 先將中文二聲和 Tone 2 Borrowing 之英文字母一起詢問前面音節是否為中文二聲且 Tone 2 Borrowing 之英文字母且在前面邊界為 B0、B1 的情況下 ，往下走到節點

34 時，則是將 B0 獨立出來，而節點 37 則是又將中文二聲還有 Tone 2 Borrowing 之英文字母分開，此結果也再度顯示若是中英文特性差異太大，決策樹自動會將其分開。因此不僅可以省去人工化簡的時間，即便換了另一個中英夾雜語料庫，此方法仍然適用，不會因語料庫不同而需要再次根據語料庫特性進行人工分類的動作，這也意味著此方法相當彈性。

綜合以上討論和 4.1 節中的表 4.2 與表 4.3 之 TRE 值，可得知英文字母的連音現象在本語料庫裡並不顯著，可能原因為本語料庫在念英文字母時，多半音節間停頓時長比較長，此推論也將在稍後停頓標記分析中探討到。

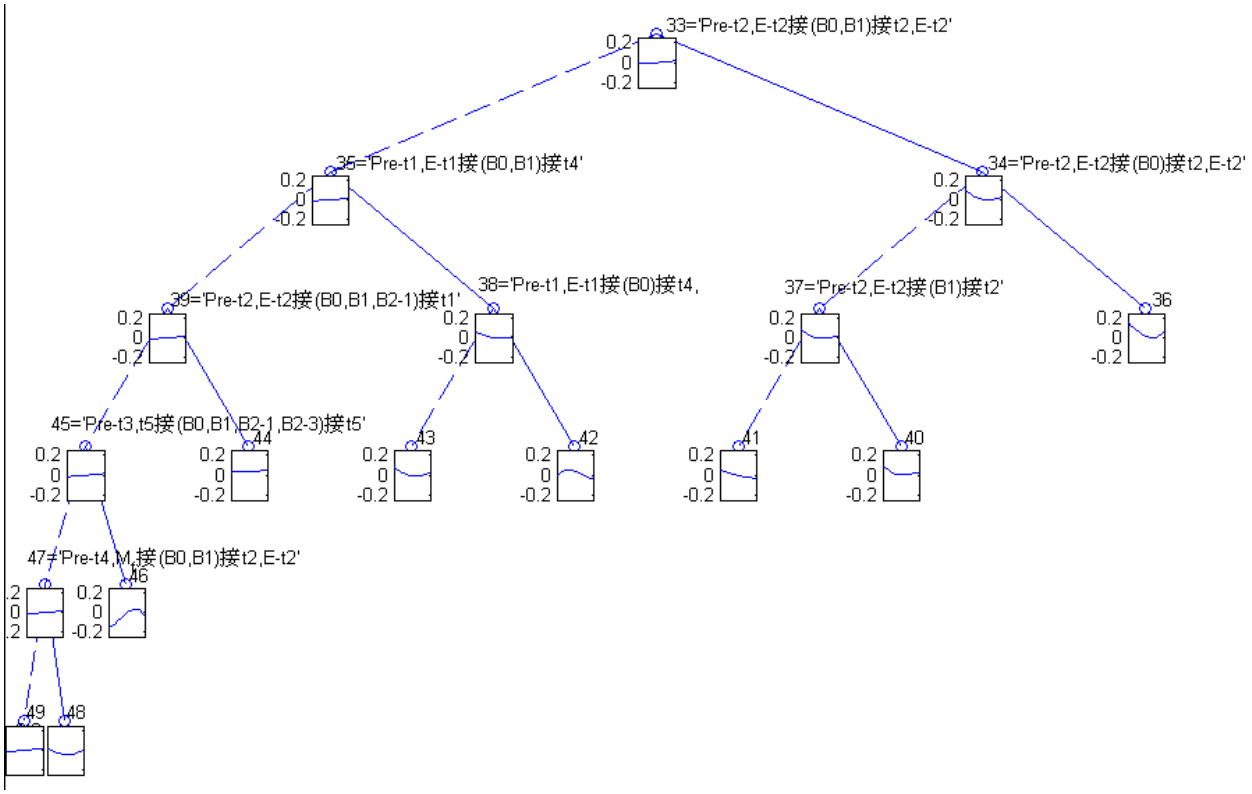


圖 4.7：Forward 連音參數之決策樹

4.1.1.2 音節層次中音節長度之影響型態

接著分析音節長度之 APs 對音節長度的影響，由圖 4.8 中(a)顯示中文二聲的音節長度比較長，再來為一聲和四聲，而比較圖(b)和(c)會發現英文字母型態的音節長度多數明顯大於一般基本音節，其原因為基本音節為中文已先扣掉聲調的影響參數，意即英文字母的影響因素等同於中文字的聲調加上基本音節，故其值會大一點。而在不同英文字母型態中會發現 B 和 D 的音節長度相對較短，其原因在於 B 和 D 發音方式類似中文聲母中的爆破音，如ㄅ(b)、ㄉ(d)這個類別，故整體音節長度偏短。類似現象還有 P、T 類似中文聲母ㄆ(p)、ㄊ(t)，其音節長度亦相近。另外，F、H、M_r、R、S 及 X 音節長度都偏長，而這些英文字母的音高軌跡都類似中文二聲(Tone Borrowing)，而二聲的中文音節長度也正是最長的。綜合以上，可知不同中文聲調和不同基本音節對中文音節長度明顯有所不同，而英文字母也會因其發音方式的不同，呈現不同的音節長度，但本研究並不採取將英文字母分類成相似的中文聲調和相似的中文基本音節類型，因為中文聲調加上基本音節類型的組合太多，一旦將英文也歸類在其中，其音節長度很有可能被中文字干擾，故仍將其獨立成 27 種類別來模擬。

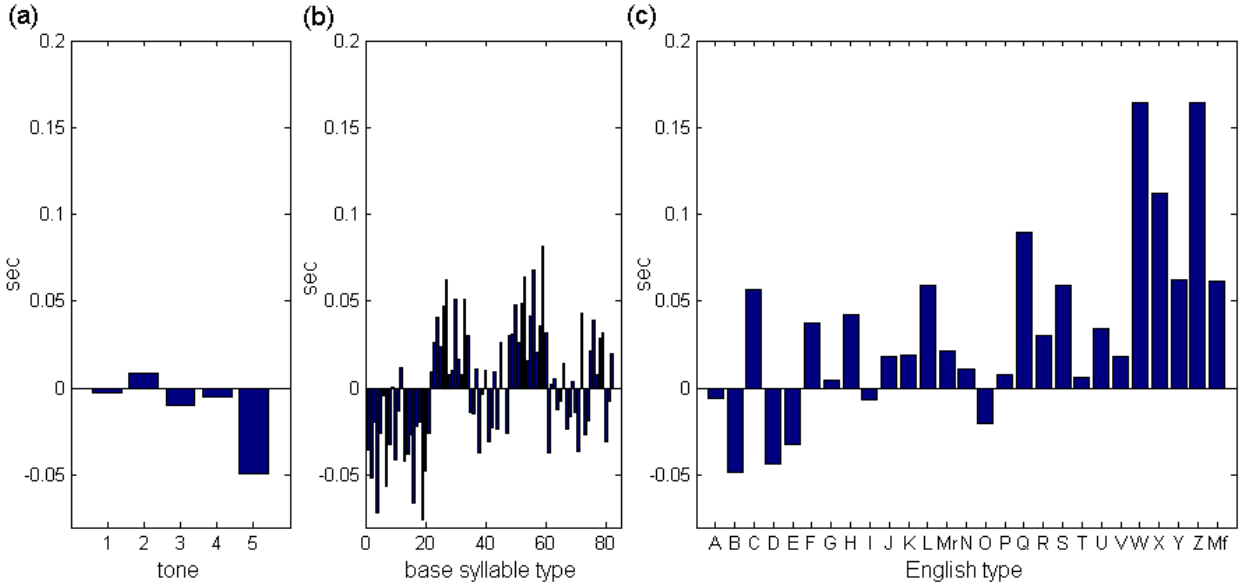


圖 4.8：(a)中文五種聲調(b)基本音節型態以及(c)英文型態之音節長度 AP

4.1.1.3 音節層次中音節能量之影響型態

最後觀察音節能量之 APs 對音節能量的影響，如圖 4.9 所示。在中文部分聲調對能量的影響以一聲和四聲來的最大，而中文韻母型態和英文字母型態相比，其動態範圍差不多，也顯示英文字母並沒有因為不同語言而造成音節能量上有所太大差異。

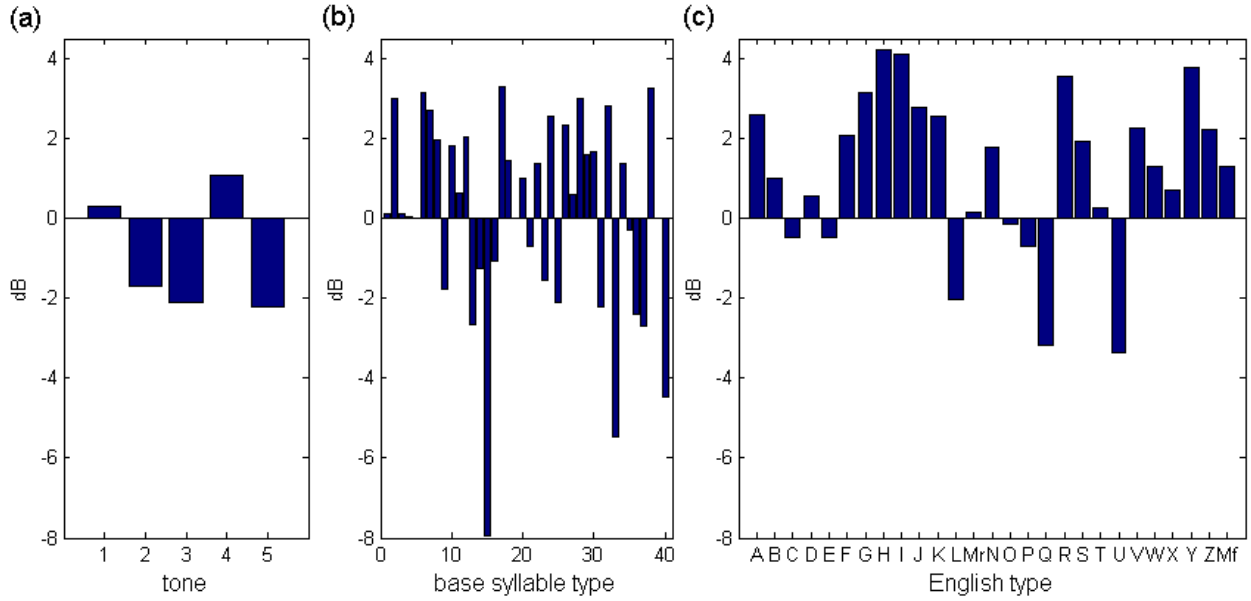


圖 4.9：(a)中文五種聲調(b)韻母型態以及(c)英文型態之音節能量 AP

4.1.2 上層韻律狀態之影響型態

韻律狀態為扣除音節層次之 APs，上層韻律之貢獻與變化，本研究重點在於觀察中文和英文之各類韻律狀態分佈情形。圖 4.10 中(b)與(c)分別顯示了音節基頻之中文和英文字母韻律狀態的分佈狀況，以相對數量的角度觀察，英文字母的韻律狀態並不會特別落在某個狀態，而是近似於中文韻律狀態的分佈，這也證明了英文字母基頻的上層韻律資訊受到前後中文韻律變化的牽制與影響，並跟隨中文的韻律變化有著不同的高低起伏。(d)和(e)則分別為中、英文基頻正規化後的韻律狀態值分佈，一樣可發現中、英文的分佈相近，此外也可觀察到其動態範圍皆大於其他音節層次的 APs，也顯示出韻律的變化主要還是取決高層次(韻律詞、韻律短語和呼吸群/韻律群)的貢獻。而由圖 4.11 與圖 4.12 一樣可觀察出音節長度和音節能量之韻律狀態分佈情況並不會因中、英文而有所太大的差異。

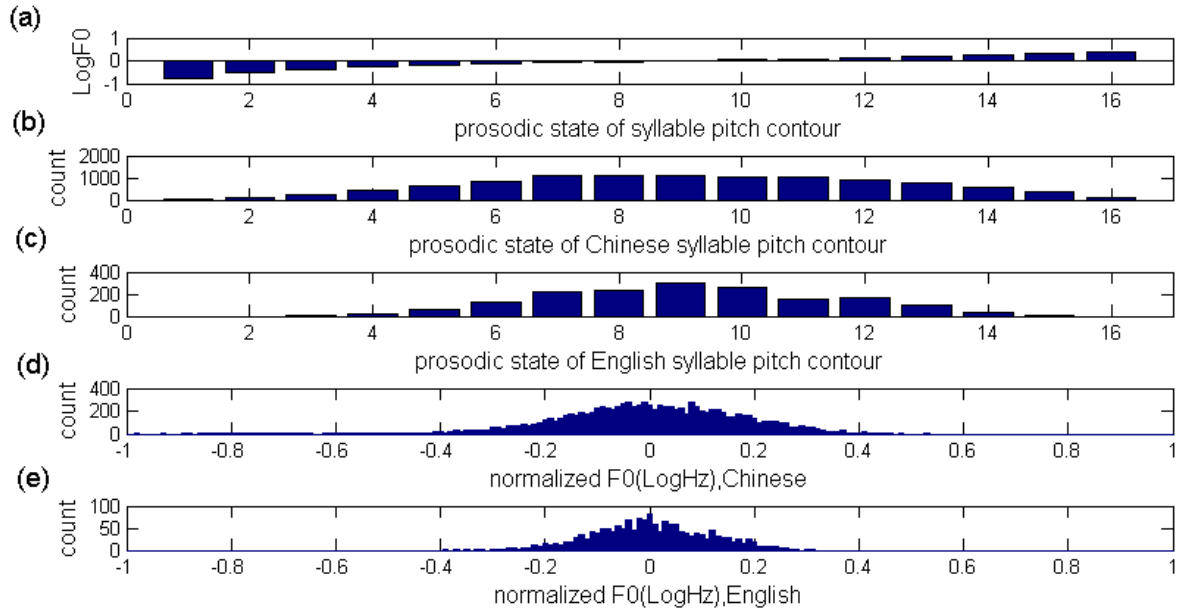


圖 4.10：基頻之(a)各韻律狀態之數值(b)中文韻律狀態分佈圖(c)英文字母韻律狀態分佈圖(d)中文正規化基頻數值分佈圖(e)英文字母正規化基頻數值分佈圖之比較

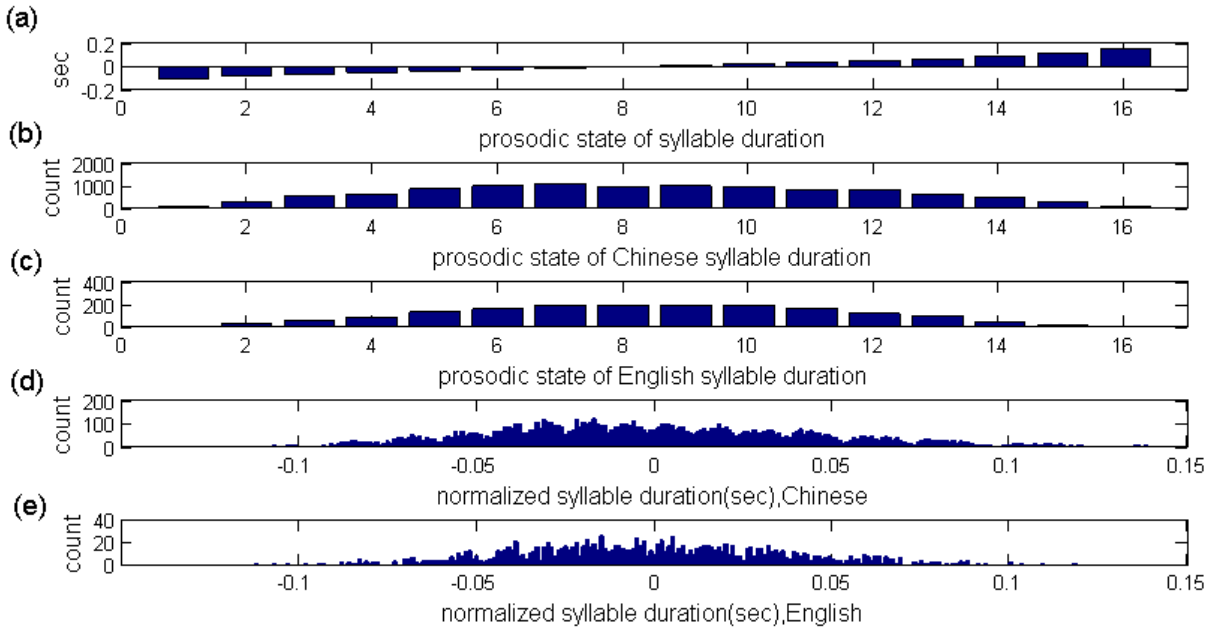


圖 4.11：音節長度之(a)各韻律狀態之數值(b)中文韻律狀態分佈圖(c)英文字母韻律狀態分佈圖(d)中文正規化音節長度分佈圖(e)英文字母正規化音節長度分佈圖之比較

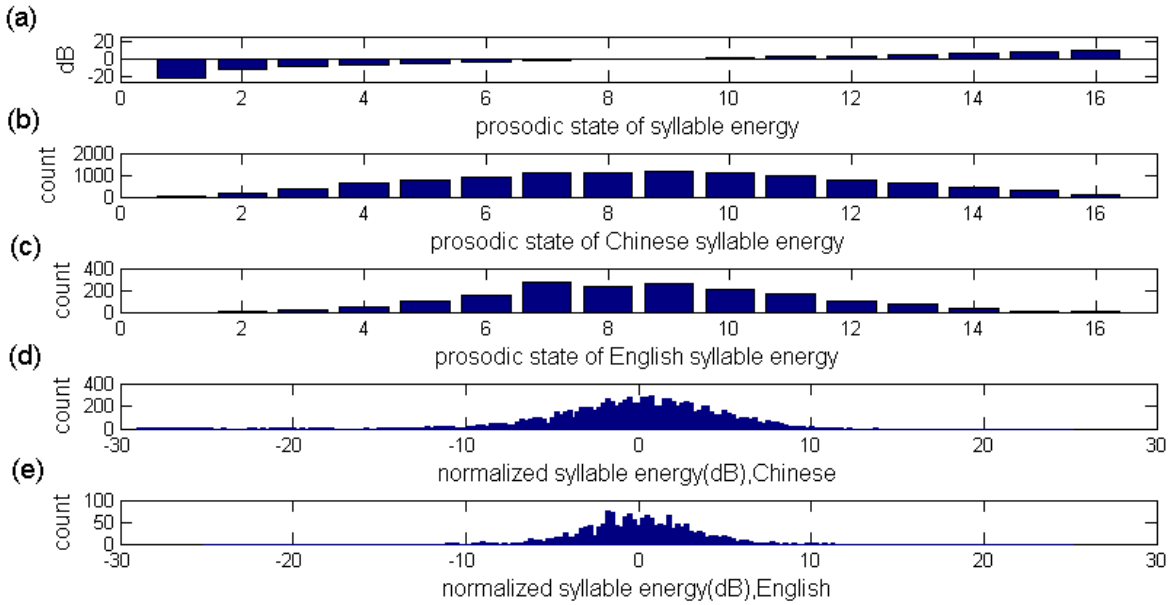


圖 4.12：音節能量之(a)各韻律狀態之數值(b)中文韻律狀態分佈圖(c)英文字母韻律狀態分佈圖(d)中文正規化音節能量分佈圖(e)英文字母正規化音節能量分佈圖之比較

4.2 停頓標記聲學模型

停頓標記聲學模型為利用 CART 演算法描述七種停頓標記**B**、語言參數**I**與音節間參數**Y**及音節差韻律參數**Z**之間的相互關係。圖 4.13 顯示在不同停頓標記下，決策樹根節點中各參數的分佈圖。由此圖可以觀察到愈上層(BG/PG)的停頓標記如 B3、B4 有較長的停頓時長、較低的音節間能量低點、較大的正規化基頻跳躍值以及較大的正規化音節延長因子。而 B0 與 B1 則都具有非常短的停頓時長，但 B0 擁有最大的音節能量低點，表示 *B0* 為兩音節緊密連接的邊界。B2-2 的停頓時長和音節能量低點則介於 B0/B1 和 B3/B4 之間。B2-3、B2-1 及 B1 有相似的停頓長度與音節能量低點。B2-1、B2-2、B3 及 B4 比起 B0、B1 及 B2-3 具有較大的正規化基頻跳躍值，這也顯示音高重置(pitch reset)分別在 intra-PW 與 inter-PW 音節邊界的影響。而有較長的正規化延長因子則為 B2-2、B2-3、B3 與 B4，這也表示音節長度拉長現象通常在 PW、PPh、BG/PG 的最後一個音節。而這些參數表現也符合本研究一開始所定義之停頓標記特性。

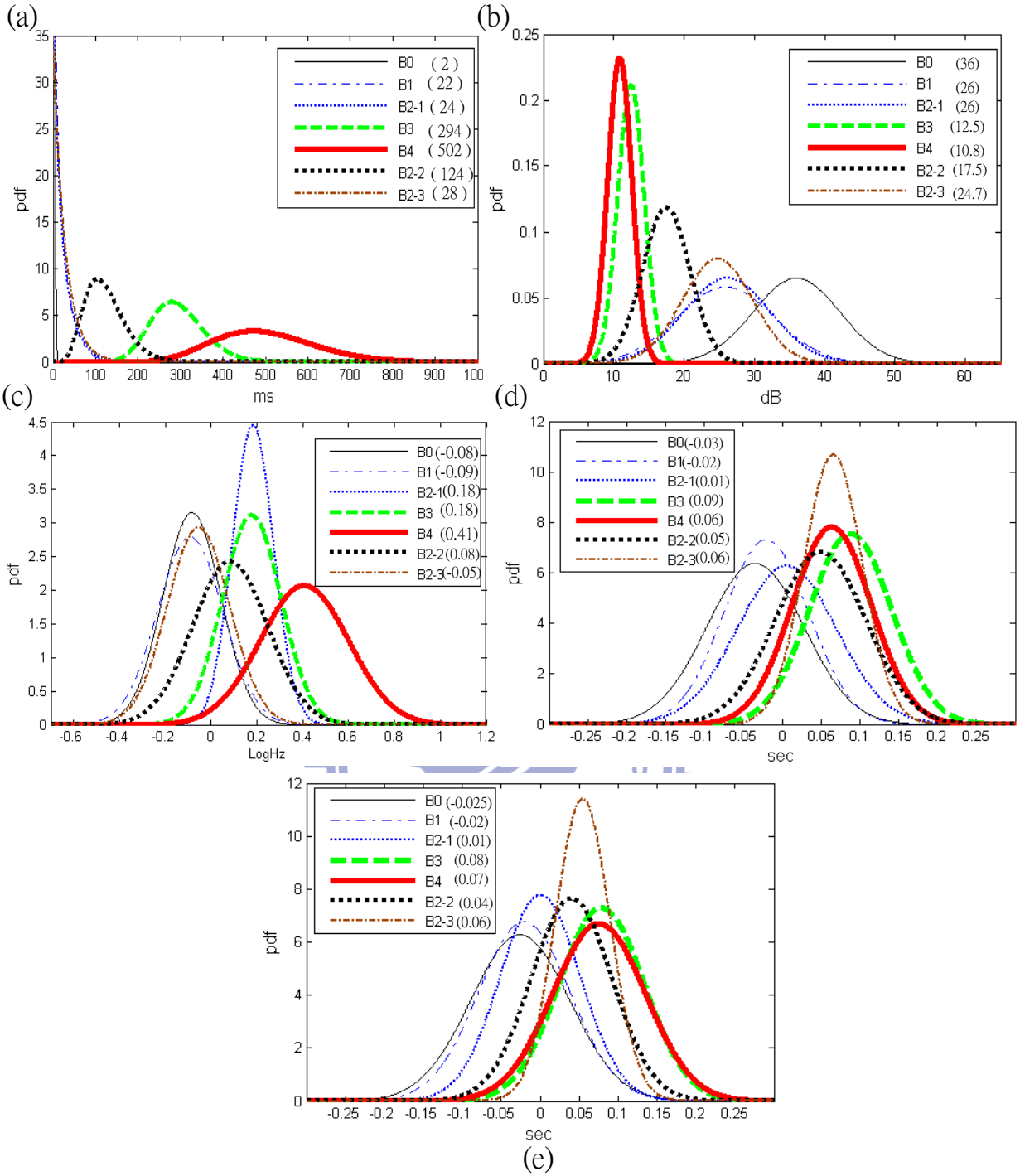


圖 4.13：(a)停頓音節長度 (b)音節能量低點 (c)正規化基頻跳躍值 (d)正規化音節延長因子 1(e)正規化音節延長因子 2 之分佈圖，其中括號中之數值為平均值

4.3 韻律狀態轉移模型

韻律狀態轉移模型描述在已知前一個音節停頓標記之下，基頻、音節長度及音節能量的韻律狀態轉移情形，圖 4.14 至圖 4.16 分別為基頻、音節長度及音節能量的韻律狀態轉移圖。觀察圖 4.14 可發現 B0、B1 之基頻韻律狀態轉移多由大至小或是維持不變，這也代表在一個 PW 之內，基頻上層之韻律變化由高至低。而 B2-1、B3 及 B4 則由低階的韻律狀態轉移至高階的韻律狀態，顯示在這些邊界情況下，會有明顯的音高重置現象。至於 B2-3 則是比較接近 B0/B1 的轉移狀態，也顯示出 B2-3 不是使用音高重置現象來代表韻律詞邊界。

觀察圖 4.15 可發現在 B0/B1 的情況下，音節長度的韻律狀態多由低至高，而 B3 與 B4 則是明顯由高階韻律狀態降至低階韻律狀態，代表在 PPh、BG/PG 等大韻律單元邊界會有顯著的音節延長效應。至於 B2-3 與 B2-2 之轉移狀態也是由高轉至低，代表在沒有明顯音節停頓的情況下，仍然可以透過音節長度的延長，反應出韻律詞邊界。

最後觀察圖 4.16 可發現在 B3/B4 的情況下，音節能量的韻律狀態會由低至高大幅提升，表示在 PPh、BG/PG 等邊界其能量會降至很低，再由新的韻律單元起始將能量提高。也代表一個 PPH、BG/PG 中的能量走勢是由高至低。

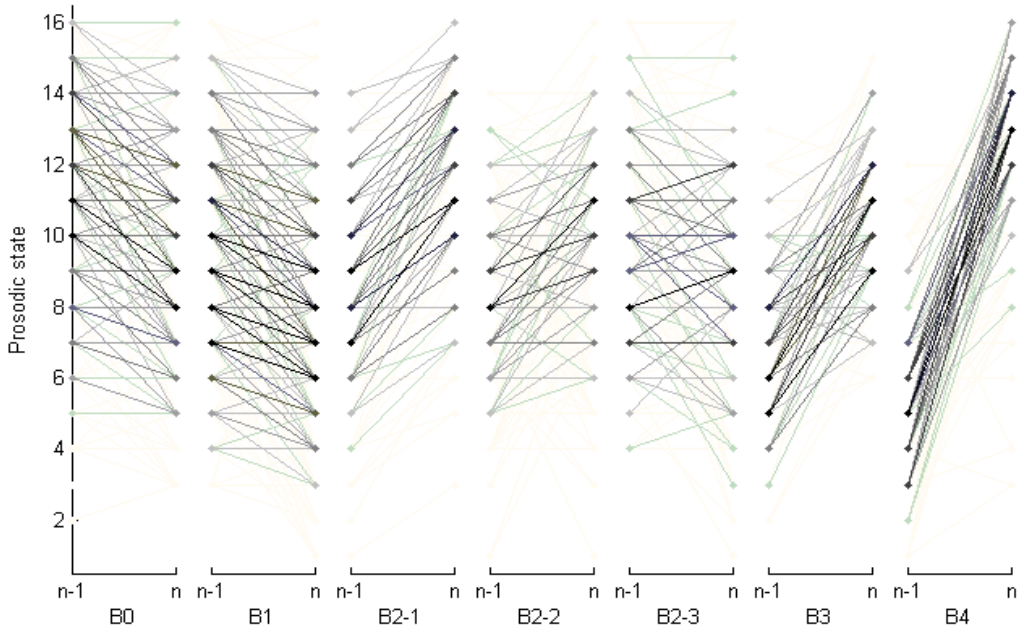


圖 4.14：各停頓標記下，基頻韻律狀態轉移的狀況，顏色越深表示此狀態轉移的機率越大

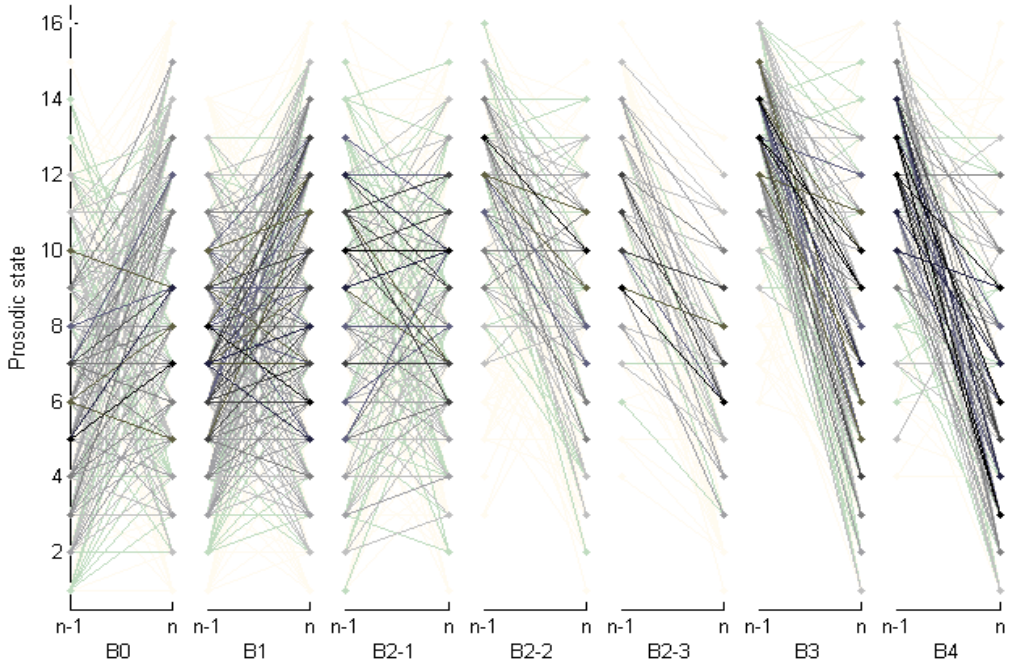


圖 4.15：各停頓標記下，音節長度韻律狀態轉移情形，顏色越深表示此狀態轉移的機率越大

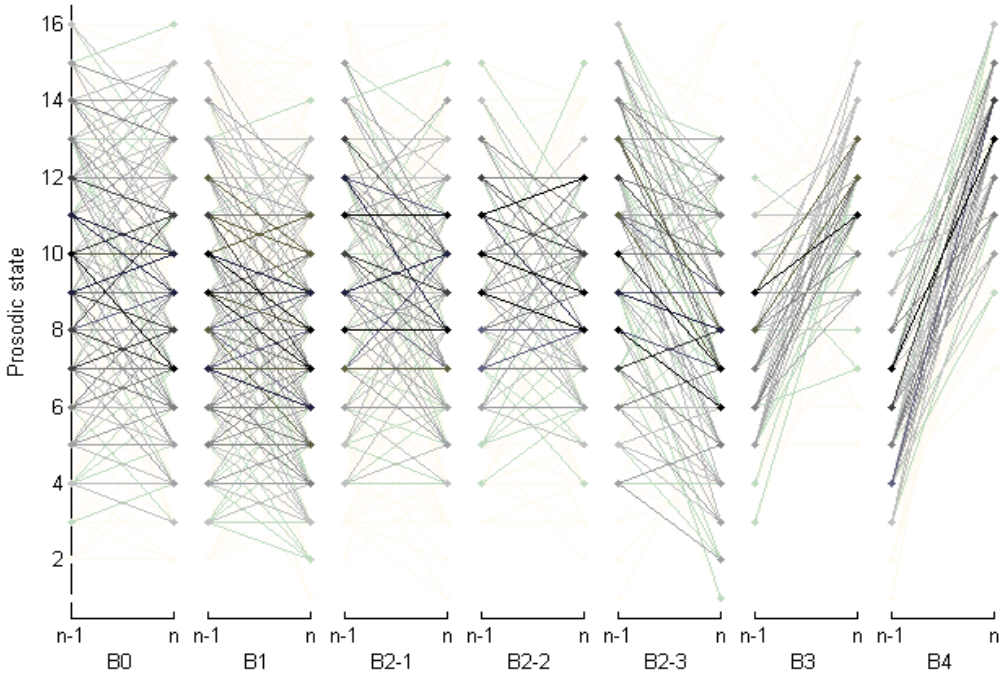


圖 4.16：各停頓標記下，音節能量韻律狀態轉移情形，顏色越深表示此狀態轉移的機率越大

4.4 停頓標記語言模型

由停頓標記聲學模型的結果可知道停頓標記會受到聲學參數的影響而有所不同。而語言參數也是影響停頓標記的一個重要因數，因此本研究透過停頓標記語言模型來驗證此觀點。停頓標記語言模型描述在不同語言參數的情況之下，停頓標記的分佈情形。圖 4.17 為本研究透過不同語言參數的問題下，所實作出的停頓標記語言模型之決策樹，每個節點都有對應的節點編號和其問題，編號 1 即代表根節點，每個節點內的長條圖表示落在此節點的停頓標記個數，由左至右分別是 B0,B1,B2-1,B2-2,B2-3,B3,B4，另外實線表示父節點問題為“是”，虛線則表示父節點問題為“否”。由圖(a)可觀察到一開始在根節點先問到此標記所對應之語言參數是否為標點符號(PM)，因為大部分 B3 與 B4 都位於標點符號處，因此藉由此問題可將大部分 B3 與 B4 與其他停頓標記區隔開來，由節點 2 之長條圖分佈也顯示出此現象。而由落在節點 4 之長條圖分佈，可以確信在標點符號為逗號之時，其對應之停頓標記幾乎都是 B3 與 B4，代表語者通常會以逗號當作是一個 PPh 或 BG/PG 的結尾。將位於標點符號之邊界區隔開後，節點 3 則是以是否在處在音節詞內邊界來區分其於停頓標記，若不是則往節點六走，接著並詢問此邊界之下一個音節的聲母型態是否為{ INULL,m,n,l,r , AEFHILMNORSUVXY }，此問題包含了中文和英文字母，其意義為下一個音節起始是濁音，因為若下一個音節起始為濁音，則音節與音節間容易形成 tightly coupling，使得此邊界標記成 B0 的機率較高。

然而由接續下來的節點 10 問題到節點 14 與節點 15 的結果，可發現中、英文的特性還是有所不同，落在節點 15 表示此邊界下一個音節起始為濁音的英文字母型態，反之節點 14 則是濁音的中文型態，落在節點 14 中停頓標記以 B0、B1 數量最多，而節點 15 中則以 B2-2 和 B3 最多，這顯示了即便當下一個音節為濁音之英文字母型態時，其停頓時長比中文來的長，因此並不會有 tightly coupling 的現象，而由於此語料中大部分的詞都是以單一語言為主，意即只有相當少量的字會在詞內出現 code-switch，因此上述結果也顯示在中文詞轉英文詞之邊界大多以 B2-2 和 B3 居多，代表在 code-switch 處之停頓時長通常都較長。接著再看音節詞內邊界的韻律標記分佈情形，由落在節點 7 之分佈狀況，可發現到若是音節邊界為詞內邊

界則其停頓標記不分中英文，皆多數為 B0 與 B1，而由圖(b)繼續探討下去，節點 13 在詞內邊界的前提下詢問下一個音節之聲母類型是否為{INULL,m,n,l,r}，而由節點 24 與節點 25 的結果顯示大部分停頓標記以 B0 為主，其原因如同上述所說，在此條件的情況下，音節間容易形成 tightly coupling，故大多標記成 B0。而由節點 27 與節點 35 所問到之問題，也可發現當下一個音節為英文字母且發音方式類似中文聲母某種類型時，其對應到的停頓標記也會與此中文之停頓標記相似。詳細英文字母所對應到之停頓標記分析將在 4.5 節中呈現。

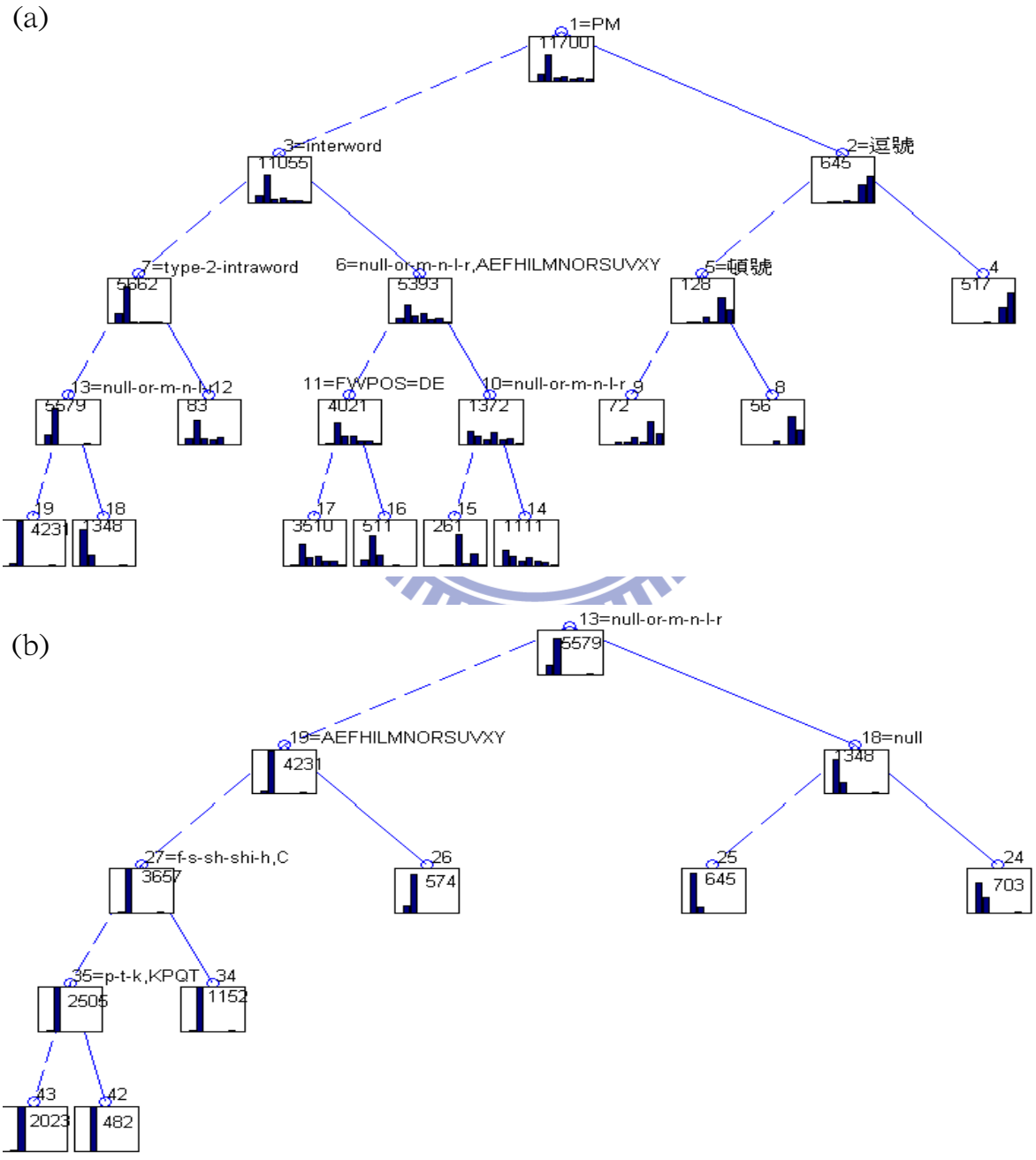


圖 4.17：停頓標記語言模型之決策樹

4.5 韻律標記結果之分析

分析停頓標記時，本研究重點在於英文字母所對應之停頓標記是否有特殊現象，圖 4.18 中(a)、(b)、(c)分別顯示出所有音節、中文字和英文字母所對應之停頓標記分佈比例，可觀察到英文字母所對應之停頓標記為 B2-1 與 B0 的比例相對於中文來說較低，而 B1 之比例則較中文來的高，其他停頓標記比例則和中文相仿。B2-1 偏低的可能原因為英文字母 27 種類型中高達 17 種為 Tone 1 Borrowing，其音節基頻本身就偏高，因此不太容易再有音高重置的現象。而由第三章中英文字母特性可知英文詞內間停頓長度比中文詞內間停頓長度來的高，因此英文詞內邊界大部分都標記成 B1，使得英文字母 B0 之比例相對降低不少。

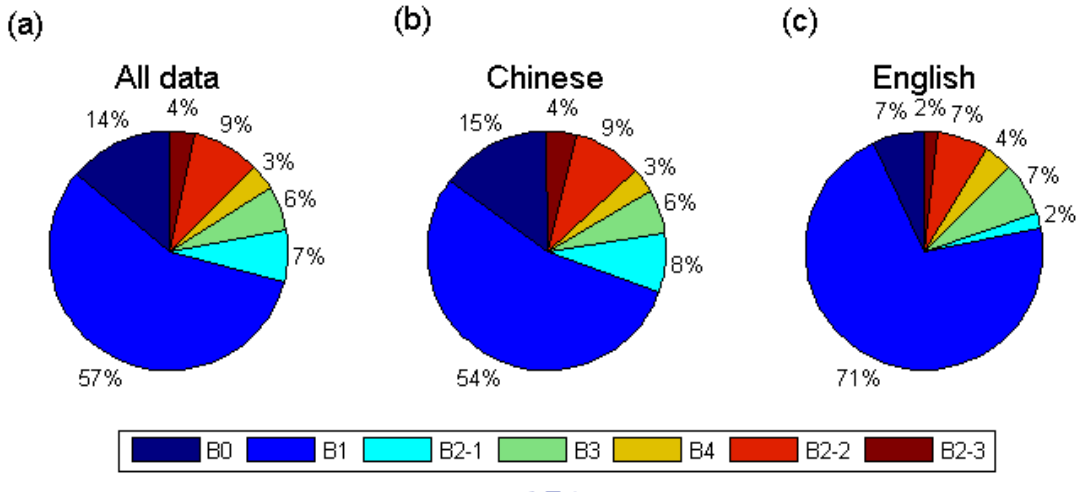


圖 4.18：音節邊界之停頓標記分佈：(a)所有音節邊界(b)中文音節邊界(c)英文字母音節邊界

接著表 4.4 與表 4.5 分別列出當 Non-code switch 與 code-switch 時各類停頓標記的統計量，由觀察表 4.4 可知當 Non-code switch 時，只要此停頓標記位於詞內邊界處，無論是中文接中文還是英文接英文幾乎全部為 B0 與 B1，若是在詞外邊界處，中文因數量龐大使得各種停頓標記皆有出現，而英文則是因本身語料庫英文詞接英文詞數量就偏少，且多半介於標點符號之間，因此皆為 B2 類以上的標記。最後觀察表 4.5 可發現當停頓標記為 code-switch 且位於中文詞轉英文詞之詞外邊界(C-E-Interword)時，以 B2 類以上的標記最多，其中又以 B2-2 和 B3 數量為大宗，這也再度顯示由中文詞轉英文詞時，為了強調英文詞，往往會有較大的停頓長度。以下列舉幾個此類別的範例：

範例 1：**透過**,P　(**B2-2**)　ＳＭＴＰ,Nb　的,DE　(B2-3)　功能,Na

範例 2：少女,Na　**團體**,Na　(**B3**) ＳＥＳ,Nb

而當邊界為 code-switch 且位於英文詞轉中文詞之詞外邊界(E-C-Interword)時，其標記仍舊以 B2 類以上為大宗，但會發現 B1 的數量也占了此類別將近三分之一，這是因為不少英文詞接中文詞往往會形成一個詞組或是片語，因此其邊界之韻律斷點就不會那麼強，使得標記成 B1 的機率大幅提升。以下也列舉幾個此類別的範例：

範例 1：ＩＢＭ,Nb　(**B1**)　**公司**,Nc

範例 2：ＴＣＰ,Nb(／PM)　(B4) ＩＰ,Nb　的,DE　(B3) ＤＮＳ,Nb　(**B1**) **選項**,Na

表 4.4：Non-code switch 處之停頓標記統計

	Non-code-switch Break			
	C-C-Interword	E-E-Interword	C-C-Intraword	E-E-Intraword
B0	437	0	1074	96
B1	2002	0	3437	961
B2-1	778	0	6	0
B3	445	10	0	0
B4	294	6	0	0
B2-2	676	5	5	0
B2-3	353	1	11	0

表 4.5：code-switch 處之停頓標記統計

	Code Switch Break			
	C-E-Interword	E-C-Interword	C-E-Intraword	E-C-Intraword
B0	0	18	0	0
B1	20	**163**	14	48
B2-1	17	35	3	1
B3	**135**	**105**	1	1
B4	47	60	0	0
B2-2	**272**	**108**	1	2
B2-3	25	26	0	1

由以上討論可知道英文字母與 code-switch 處所對應的停頓標記確實與中文及 Non-code switch 有些許差別，然而這些統計都是針對當前這個音節在中、英文或是 code-switch 與否所呈獻的結果，意即最多只考慮到前後音節型態的變化，但本韻律標記模型是建構在階層式的韻律架構上，代表我們不應該只關心英文字母或是 code-switch 處的標記情形，而是分析一個 PPh 或 BG/PG 等大韻律單元內的停頓標記相對情形，因為我們認為在此韻律架構下，韻律斷點的強度不僅僅是只受到當下這個音節是英文或是 code-switch 處的影響，也受到前後文語法的影響，而有相對韻律斷點強度。為了驗證此觀點，本研究將進行以下的分析討論。

由於中英夾雜語料庫中含有大量的名詞片語，我們藉由分析這些名詞片語的層次結構和其對應的韻律結構，來驗證停頓標記間的相互關係，並輔助未來使用在語音合成中的韻律產生器。首先我們對中英夾雜語料庫標示所有的名詞片語，並標示這些名詞片語的層次結構，經過標示後，我們將名詞片語的類型分為以下三類：

(1) 雙詞結構：名詞片語由兩個組成，此類的數量在名詞片語中佔最多數。舉例如下：

美國,Nc 國會,Nc (偏正結構)

台灣,Nc 產業,Na (偏正結構)

ＷＴＩ,Nb 價格,Na (偏正結構)

ＡＹＯ,Nb 新聞網站,Na (偏正結構)

ＡＴM,Nb 網路,Na (偏正結構)

ＶＨＳ,Nb 影帶,Na (偏正結構)

ＦＩＴ,Nb 今日百貨,Nb (並列結構)

ＤＮＳ,Nb 選項,Na (偏正結構)

外星人,Na ＥＴ,Nb (並列結構)

(2) 三詞結構：一個名詞片語由三個詞組成，其中詞和詞之間的組合是有層次的，片語中有一個為中心語的詞或名詞片語，另外還有這個中心語的形容或是附加成分。下為範例，其中 “{}”代表片語邊界； “[]”代表中心語； “()”代表形容中心語的附加成分。

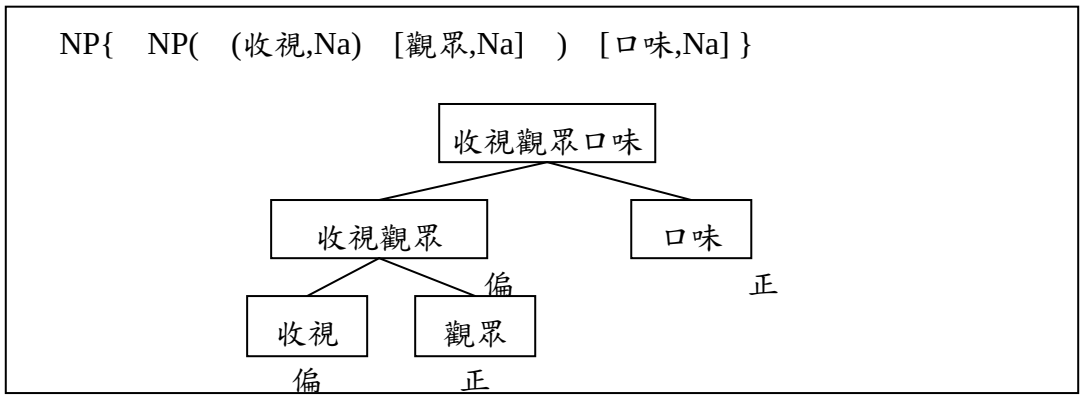


圖 4.19：名詞片語 “收視,Na 觀眾,Na 口味,Na”的層次結構

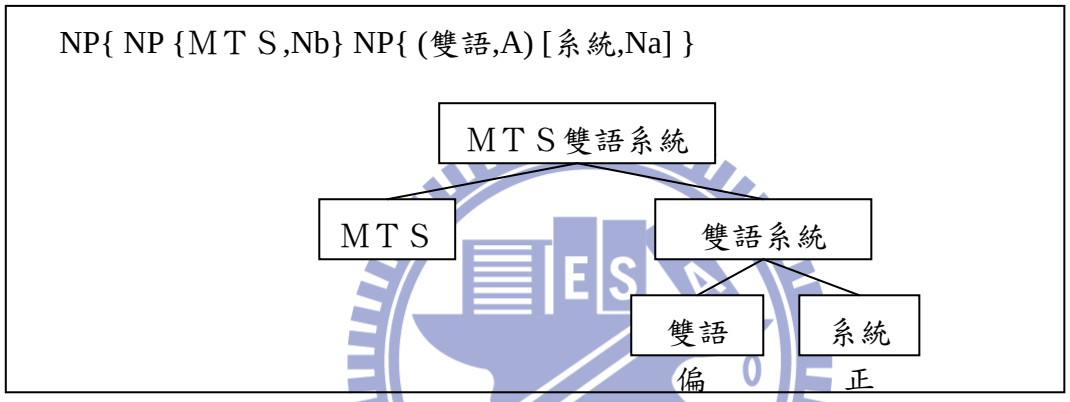


圖 4.20：名詞片語 “ＭＴＳ,Nb 雙語,A 系統,Na” 的層次結構 (並列結構)

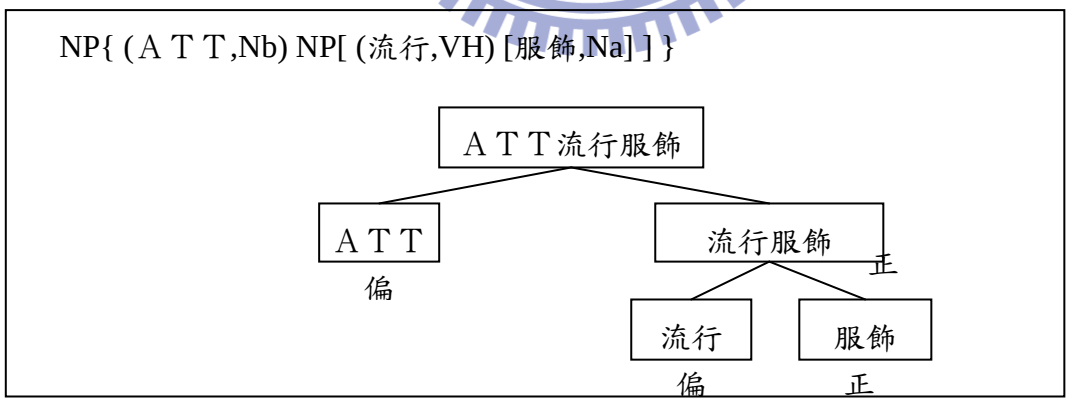


圖 4.21：名詞片語“ＡＴＴ,Nb 流行,VH 服飾,Na” 的層次結構

由上面三個圖的範例可以知道，在三詞結構下的名詞片語有明顯的層次結構，如：“收視,Na 觀眾,Na 口味,Na” 先是由 “收視,Na 觀眾,Na” 組成名詞片語，再由 “收視,Na 觀眾,Na” 和 “口味,Na “ 組成更大的名詞片語； “ＭＴＳ,Nb 雙語,A 系統,Na”，其中 “雙

語,A 系統,Na” 為一個名詞片語，而 “ＭＴＳ” 就是所謂的雙語系統，因此 “ＭＴＳ,Nb” 和 “雙語,A 系統,Na”為並列結構的名詞詞組；在“ＡＴＴ,Nb 流行,VH 服飾,Na” 之中，“流行,VH 服飾,Na” 為一個名詞片語且為中心語，而 “ＡＴＴ,Nb” 為修飾 “流行,VH 服飾,Na” 的部份，因此構成更大的語意語法單位 “ＡＴＴ流行服飾”。

(3) 複雜結構：此類的名詞片語由三個詞以上組成，其層次結構較雙詞和三詞複雜許多，以下為一範例：

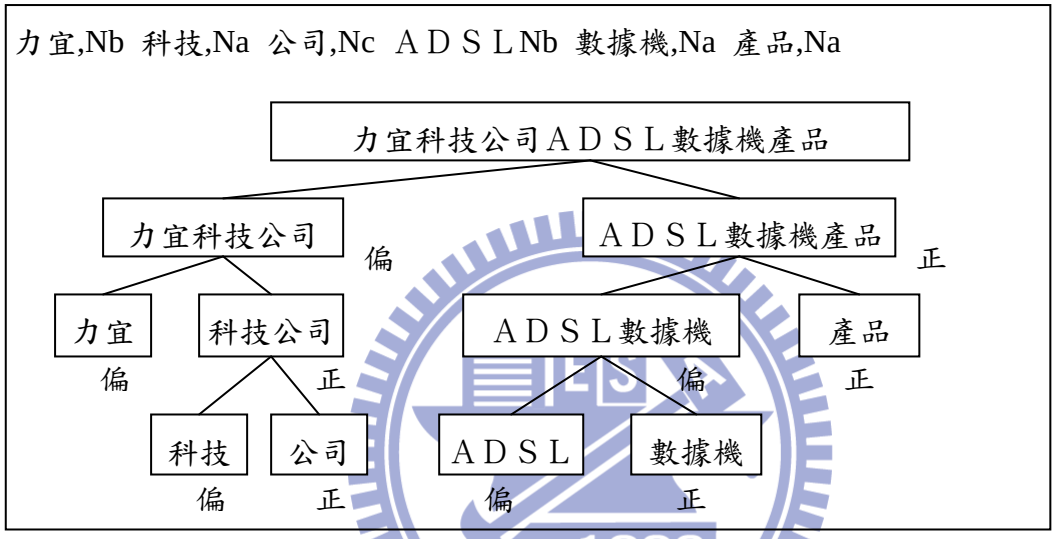


圖 4.22：名詞片語 “力宜,Nb 科技,Na 公司,Nc ＡＤＳＬNb 數據機,Na 產品,Na”的層次結構

表 4.6 為這三種名詞片語結構的統計數量及其總量，可以發現到名詞片語佔整個語料庫詞數的三分之一，所以將名詞片語的層次結構和韻律結構的關係建立起來，對於語音合成中的韻律產生有很大的幫助。

表 4.6：名詞片語結構於中英夾雜語料的統計數量

	片語數	詞數
雙詞結構	476	952
三詞結構	163	489
複雜結構	123	544
名詞片語總數	777	1985
整個語料庫		6606

在了解本研究對名詞片語的分類後，我們將就這三類名詞片語作語法結構上的層次結構

標示。

(1) 雙詞結構：

由於雙詞結構是由兩個詞組成，相對於三詞結構和複雜結構，其詞的層次結構只有兩層(詞為下一層，片語為上一層)，片語中的詞邊界位置，相對於片語邊界來說，其語法上的層次較低的語法邊界，我們以下面的篇章為例：

由於,Cbb **ＷＴＩ,Nb 價格,Na** 並,D 未,D 下降,VA(，PM) 中油公司,Nb 油價,Na 上漲,VH 的,DE 計畫,Na 將,D 暫時,D 延期,VB(。PM)

在第一句的名詞片語前後文「由於,Cbb **ＷＴＩ,Nb 價格,Na** 並,D」，詞邊界「ＷＴＩ * 價格」 的層次較 「由於 * ＷＴＩ」 和 「價格 * 並」 低。由於雙詞結構的型態簡單，由片語的邊界標記便可知道片語內和片語邊界相對的層次關係，並不需要特別對其標示相對的層次結構。

(2) 三詞結構和複雜結構

依據上面名詞片語的分析，三詞結構和複雜結構的片語多可以用樹狀結構表示其層次關係，我們利用詞邊界在樹狀結構中的深度(depth in a tree)來定義詞邊界於片語中的層次，以下為範例：

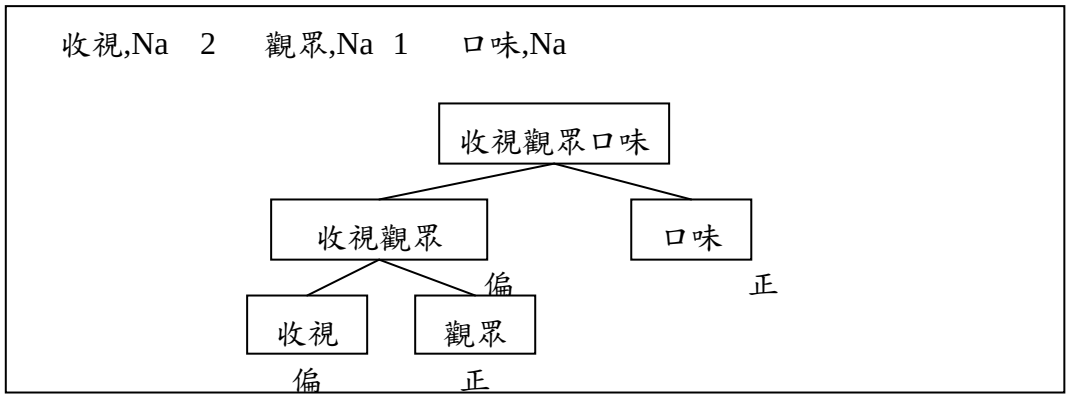


圖 4.23：名詞片語 “收視,Na 觀眾,Na 口味,Na”的層次結構標示

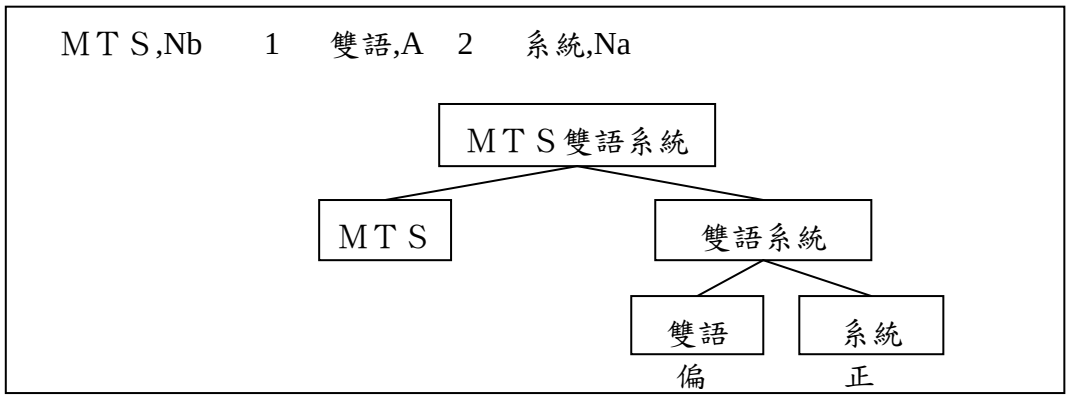


圖 4.24：名詞片語“ＭＴＳ,Nb 雙語,A 系統,Na”的層次結構標示

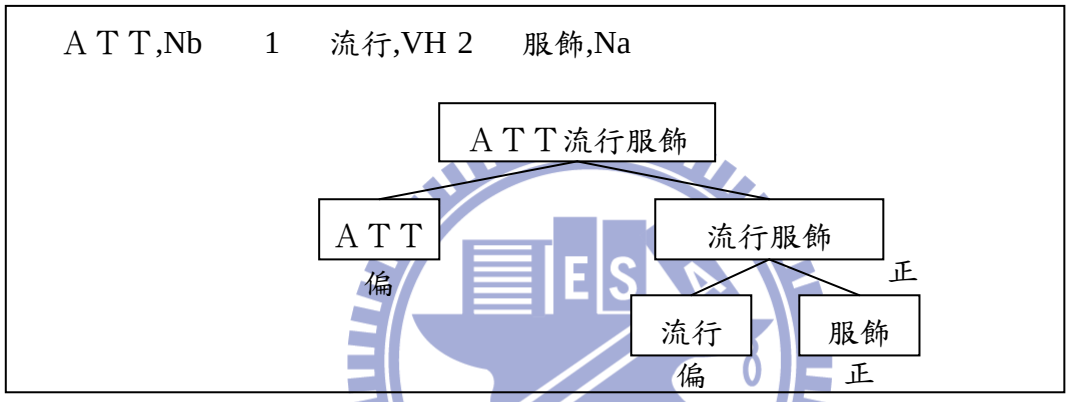


圖 4.25：名詞片語“ＡＴＴ,Nb 流行,VH 服飾,Na”的層次結構標示

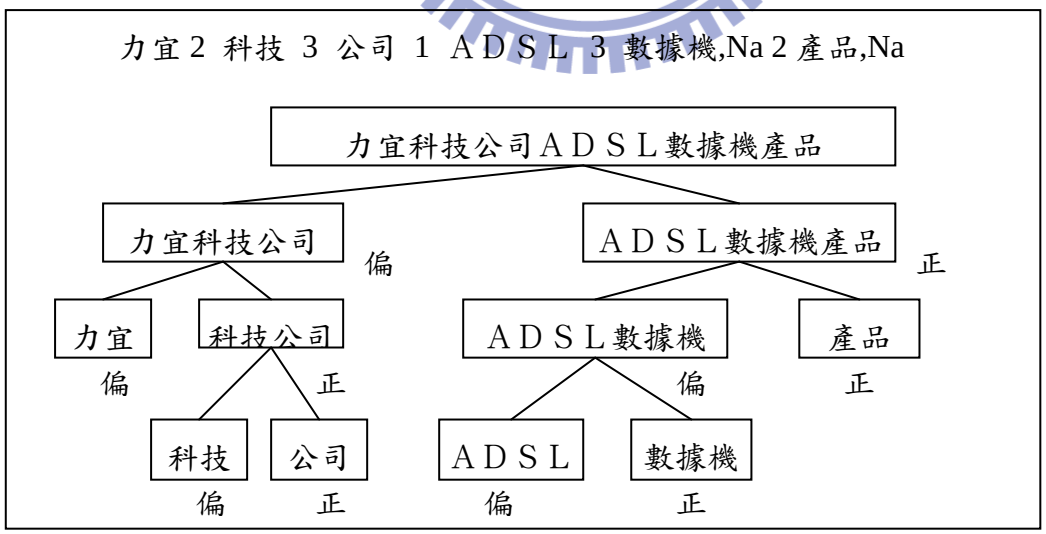


圖 4.26：名詞片語“力宜,Nb 科技,Na 公司,Nc ＡＤＳＬNb 數據機,Na 產品,Na”的層次結構標示

表 4.7 為片語層次結構標示的列舉：

表 4.7：片語層次結構標示範例

各個,DM 1 應用,Na 2 系統,Na

每一個,DM 網站,Nc

每一張,DM 網頁,Na

一個,DM 1 專用,VI 2 地址,Na

電子,Na 郵件,Na

ＡＴＭ,Nb 網路,Na

未來,Nd 1 網路,Na 2 技術,Na

ＫＧＮＶ,Nb 1 新任,VG 2 董事長,Na

中信銀,Nc 2 董事長,Na 1 辜濂松,Nb

一個,DM 1 雙環,Na 2 結構,Na

光纖,Na 網路,Na

虛擬,VC ３Ｄ,Nb

任何,Neqa 1 運動,Na 2 比賽,Na

人工,Na 智慧,Na

這次,DM 世界杯,Nb

預賽,Na 2 分組,VB 1 Ｅ組,DM

少女,Na 2 團體,Na 1 ＳＥＳ,Nb

少男,Na 歌迷,Na

Ｂ型,Na 肝炎,Na

很多,Neqa 人,Na

ＶＨＳ,Nb 影帶,Na

統一,VHC 1 ＡＤＥ,Nb 2 沙拉油,Na

一旦有了名詞片語的層次結構標示後，我們就可以利用停頓標記所對應的韻律結構做對應關係的分析。以下將針對(1)雙詞結構、(2)三詞結構作層次結構和韻律結構的相對關係分析，至於複雜結構因組合較複雜且占名詞片語比例只有 16%，故在此先不討論。

(1) 雙詞結構：

名詞片語在句子中通常扮演著主語或是賓語，也可以是修飾謂語的狀語，或是修飾主語或賓語的定語，一般來說雙詞結構的名詞片語大部分為偏正結構，也就是詞組中的第一個詞為修飾成分(定語)，而第二個詞為主語，通常以二字詞加上二字詞構成一個四字片語為多數。雖然雙詞結構的片語較一般詞在語法結構上有著較大的語法和語意單位，但它的語法功能和詞是類似的，有些片語的認定甚至常和詞混淆，因此，我們假設語音的韻律階層也會反映以上所述的語法層次結構，觀察片語內詞邊界和片語邊界的韻律結層和韻律斷點強度關係，進而建立由語法預估韻律的模型，應用於韻律產生，以下我們先對韻律階層和韻律斷點的強度進行說明：

(a) 韻律階層

韻律階層由高至低為 BG/PG、PPh、PW 以及 SYL，它們所對應的邊界(也就是韻律斷點)分別為 B4、B3、B2 及 B0/1，一般來說，越大的語法單位邊界，越容易對應到越上層的韻律邊界，例如：標點符號容易對應到 B3 或 B4，而詞內的邊界容易對應到 B0/1。

(b) 韻律斷點強度

韻律斷點強度在這個研究裡，定義為各韻律斷點所對應的聽感上不連續程度大小，很明顯地，各韻律斷點的強度由大至小分別是 B4、B3、B2、B0/1，而在 B2 裡面子類別的斷點強度由大至小為 B2-2、B2-1、B2-3，因此，總括來說在本研究中所採用的韻律斷點之韻律強度排序為：

B4 > B3 > B2-2 > B2-1 > B2-3 > B1 > B0

由上述的定義和說明，我們要驗證以下論點是否成立：韻律斷點強度反映了語法的層次結構，雙詞結構名詞片語內詞邊界之韻律斷點強度會等於或是小於雙詞結構片語邊界的韻律斷點強度，如下所示：

$LW_{n-1}\ B_{n-1}\ LW_n\ B_n\ LW_{n+1}\ B_{n+1}\ LW_{n+2}$

$NP = LW_n + LW_{n+1}$，NP 表示名詞片語，LW 代表一般詞，統計結果則如表 4.8 所示：

表 4.8：韻律斷點強度和雙詞結構名詞片語的層次關係

韻律斷點相對強度	範例	數量	百分比
$B_n \le B_{n+1}$ and $B_n \le B_{n-1}$	到,P B3 **ＥＭＢＡ,Nb B2-2 研究所,Nc** B3 進行,VC	246	51.68
$B_n \le B_{n+1}$ and $B_{n-1} = B_b$	Bb **長谷,Nc B2-3 科技,Na** B4 將,D	60	12.60
$B_{n+1} = B_e$ and $B_n \le B_{n-1}$	的,DE B2-2 **ＵＳＢ,Nb B2-3 介面,Na** (。PM) Be	88	16.80
$B_n > B_{n+1}$ and $B_n \le B_{n-1}$	身穿,VJ B3 **ＲＯＣ,Nb B2-2 字樣,Na** B1 的,DE	31	6.51
$B_n \le B_{n+1}$ and $B_n > B_{n-1}$	在,P B2-1 **澄清湖,Nc B2-2 棒球場,Nc** B3 進行,VC	22	4.62
$B_n > B_{n+1}$ and $B_n > B_{n-1}$	在,P B1 **台灣,Nc B2-1 青少年,Na** B1 間,Ng	29	6.09
Total		476	100.00

由表 4.8 的統計可以知道，將近有 82.77% (51.68+12.60+16.80) 的雙詞結構名詞片語符合我們的論點，只有 6.09%的片語，片語的左端和右端皆不符合語法結構和韻律斷點強度的相關論點。經過觀察這些不符合我們論點的 17.22% (6.51 + 4.62 + 6.09)，大部分是因為片語前後有短詞相接，這些短詞地通常是“的”、介詞、後置詞等等，這些短詞容易和片語中的詞結合成韻律詞，因此容易造成不符合我們論點的結果。相反的，符合我們論點的名詞片語，大多數是因為名詞片語本身是主語且後面是直接連著謂語，或名詞片語為賓語且前面直接連著述語。

(2) 三詞結構：

在此結構下，我們只先探討片語內語法層次結構和韻律斷點強度的關係。三詞結構的名詞片語分析表示如下：

$LW_n\ B_n\ LW_{n+1}\ B_{n+1}\ LW_{n+2}$
$LW_n\ R_n\ LW_{n+1}\ R_{n+1}\ LW_{n+2}$
$NP = LW_n + LW_{n+1} + LW_{n+2}$

其中 R_n 代表詞邊界於片語中的層次。在三詞結構下的名詞片語，我們認為韻律斷點強度的排序會和語法層次相反，也就是說韻律的結構會反映名詞片語在語法的層次結構關係。表 4.9 為韻律斷點強度和語法層次關係的統計。

表 4.9：韻律斷點強度和三詞結構名詞片語的層次關係

韻律斷點相對強度及語法層次	範例	數量	百分比
$B_n \le B_{n+1}$ and $R_n > R_{n+1}$	多媒體 B2-1 儲存卡 B3 MMC	143	87.73
$B_n \ge B_{n+1}$ and $R_n < R_{n+1}$	OECD B2-2 主要 B2-1 會員國		
$B_n > B_{n+1}$ and $R_n > R_{n+1}$	砷化鎵 B2-1 晶圓 B1 代工	20	12.27
$B_n < B_{n+1}$ and $R_n < R_{n+1}$	全球 B2-1 臭氧層 B2-2 破洞		
Total		163	100.00

由表 4.9 中的統計可以清楚看到，有將近 88%的三詞結構名詞片語的韻律斷點強度反映了我們的論點。由以上的統計數據可以合理的推論：語法的相對層次關係，語者會利用韻律斷點相對強度來表現，而不是說在同一種語言參數下(equivalent linguistic context, ELC)，語者就會使用同一種韻律斷點強度來表現語音韻律，因此為了韻律產生的目的，除了在 ELC 下產生靜態的韻律斷點預估外，也應產生因為語法相對層次關係而造成的相對韻律斷點動態變化，透過此方法應可由語言參數產生更自然的合成韻律。

第五章 基於 PLM 演算法之韻律產生器

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，我們需要更精確的韻律預估。在本章節中提出兩種韻律產生的方法，第一種方法為利用本研究提出之韻律模型，由文字直接預估音節停頓標記，藉由停頓標記產生韻律詞與韻律短語的邊界，並利用這些邊界產生更多韻律層次的文脈相關資訊，最後利用這些文脈資訊由 HTS 產生韻律參數，包括音節基頻、音節長度等。第二種方法為應用訓練完的韻律模型，直接由文字預估音節停頓標記與韻律狀態，進而由文字搭配對應的 APs 與預估出的韻律狀態，疊加合成出韻律參數。5.1 節將介紹停頓標記的預估，5.2 節將介紹透過韻律標記的預估，進而產生韻律參數。5.3 節將比較上述兩種方法與傳統 HTS 之韻律產生結果。

5.1 停頓標記預估

在第四章停頓標記結果分析中，可得到中英文夾雜語料庫之停頓標記的特點，無論是在 code-switch 處的特性，亦或是整個 BG/PG 中停頓標記的相對關係，一旦有了停頓標記之後，我們就可以區分出韻律詞、韻律短語等上層韻律邊界，並利用這些停頓標記產生出更多文脈相關資訊，藉由更豐富的文脈相關資訊，使得 HTS 在產生韻律參數上能有更精確的數值。5.1.1 節將介紹停頓標記預估方法一：All-in-one CART-based，5.1.2 節將介紹停頓標記預估方法二：Two-stage CART-based。

5.1.1 All-in-one CART-based

方法一演算法的方塊圖如圖 5.1 所示，藉由語言參數和七類停頓標記之間的相互關係，訓練一顆停頓標記語言模型決策樹，使用分類準則為最大概似函數增益。透過此決策樹即可由輸入文字之語言參數預估出七類停頓標記，而使用的語言參數則如表 5.1 所示，包含音節邊界種類、當前詞長、詞類、前後兩個(±1、±2)詞長、詞類、現在音節前後之標點符號，及當前句長、前後句子長度等，其中詞類分類是依據中研院 46 類詞類，依實詞、虛詞、八大

詞類及其他特殊詞類集合。

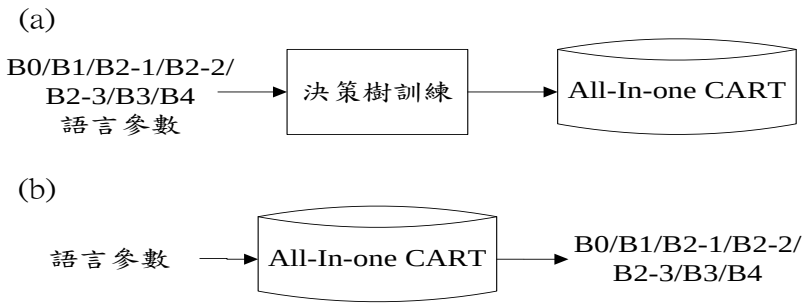


圖 5.1：停頓標記預估之 All-in-one CART-based，(a)決策樹訓練階段，(b)停頓標記預估

表 5.1：All-in-one CART-based 之語言參數列表

SB	Type of syllable boundary: inter-phrase intra-phrase inter-word, Type-1 intra-word, Type-2 intra-word.
POS0	Broad class of preceding/following word: substantive word, function word
POS1	11-type POS: A, C, D, N, I, P, T, V, DE, SHI, DM
POS2	19-type POS : A, C, Dfa, Dfb, D, N, Nd, Ne, Ng, Nh, P, T, VA, VC, VH, V_2, DE, SHI, DM
POS3	46-type POS : A, Caa, Cab, Cba, Cbb, Da, Dfa, Dfb, Di, Dk, D, Na, Nb, Nc, Ncd, Nd, Neu, Nes, Nep, Neqa, Neqb, Nf, Ng, Nv, Nh, I, P, T, VA, VAC, VB, VC, VCL, VD, VE, VF, VG, VH, VHC, VI, VJ, VK, VL, V_2, DE, SHI, DM
WL	Length of word in syllable: 1, 2, 3, 4, 5, 6
PM	Type of PM: comma, period, question mark, dun hao and others
LS	Length of sentence in syllable
LPS	Length of previous sentence
LFS	Length of following sentence
DPP	Distance to previous PM (the beginning of the sentence)
DFP	Distance to following PM (the end of the sentence)
CS	Code-switch: C-E-Interword, E-C-Interword, C-C-Interword, E-E-Interword

表 5.2 顯示出此演算法在訓練語料之韻律標記的辨認結果，可觀察到 B1 有最高的辨認率，依次再來為 B0、B4、B3 及 B2-2，而 B2-1 和 B2-3 則是辨認率偏低。分析 B2-2、B2-1 與 B2-3 之錯誤原因為大部分都辨認成 B1，因 B1 本身數量過大且 B1 是有可能標記在詞外邊界，導致單靠語言參數所找到之葉節點中 B1 的相對數量(即機率)通常最大。至於 B3 則是容易和 B4 與 B2-2 混淆，其原因則是 B2-2、B3 及 B4 所對應到之聲學特性相似(停頓時長)，故

其本身對應到之語言參數也類似。而 B0 之錯誤情形則大都標記成 B1，但 B0 與 B1 皆定義為韻律詞之詞內邊界(Intra-PW)，所以此錯誤對往後韻律參數預估影響不大。

在原本韻律架構中，B0、B1 為一個韻律詞內的邊界，B2-1、B2-2、B2-3 為區分韻律詞的邊界，B3、B4 則是區分更大韻律單元包含 PPh、BG/PG 等邊界，因此我們將七類停頓標記簡化成三類，即 Non-Break{B0,B1}，Minor Break{B2-1,B2-2,B2-3}，Major Break{B3,B4}，如此一來其停頓標記辨認率如表 5.3 所示。簡化後的三類辨認率明顯提升許多，但 Minor Break 仍然受到 Non-Break 的影響甚巨，一旦辨認不出 Minor Break，將無法區分出韻律詞的邊界，代表抓不到 BG/PG 中更細微的韻律結構變化，因此必需加強提升 Minor Break 的辨認率。

表 5.2：All-in-one CART-based 之停頓標記預估辨認率

Tar\Pre	B0	B1	B2-1	B3	B4	B2-2	B2-3	Total
B0	**84%**	13%	2%	0%	0%	1%	0%	1625
B1	6%	**90%**	2%	0%	0%	2%	0%	6645
B2-1	6%	38%	**44%**	2%	0%	8%	2%	840
B3	2%	4%	2%	**66%**	9%	16%	1%	697
B4	0%	1%	0%	17%	**77%**	5%	0%	407
B2-2	3%	15%	7%	8%	0%	**64%**	2%	1069
B2-3	9%	43%	12%	3%	0%	17%	**16%**	417

表 5.3：All-in-one CART-based 之三類韻律標記預估辨認率 ，(NB: Non-Break, MiB: Minor Break, MB: Major Break)

Tar\Pre	NB	MiB	MB	Total
NB	**96%**	4%	0%	8270
MiB	33%	**62%**	5%	2326
MB	4%	13%	**82%**	1104

5.1.2 Two-stage CART-based

由前 5.1.1 節中的方法可發現 Minor Break 很容易被辨認成 Non-Break，而太少的 Minor Break 會使得合成語音的韻律太過呆板且急速，因而降低其自然度，為了改善 Minor Break 的辨認率，本研究提出 Two-stage CART-based 的方法，其方塊圖如圖 5.2 所示。此方法是採取兩階段式辨認，首先第一級辨認是透過一顆簡化成三類停頓標記訓練出的三類決策樹(MB/MiB/NB)分類器，辨認出屬於三類中某一類後，再餵給第二級由三類停頓標記各自訓練出的分類器，辨認出屬於七類停頓標記中的哪一類。

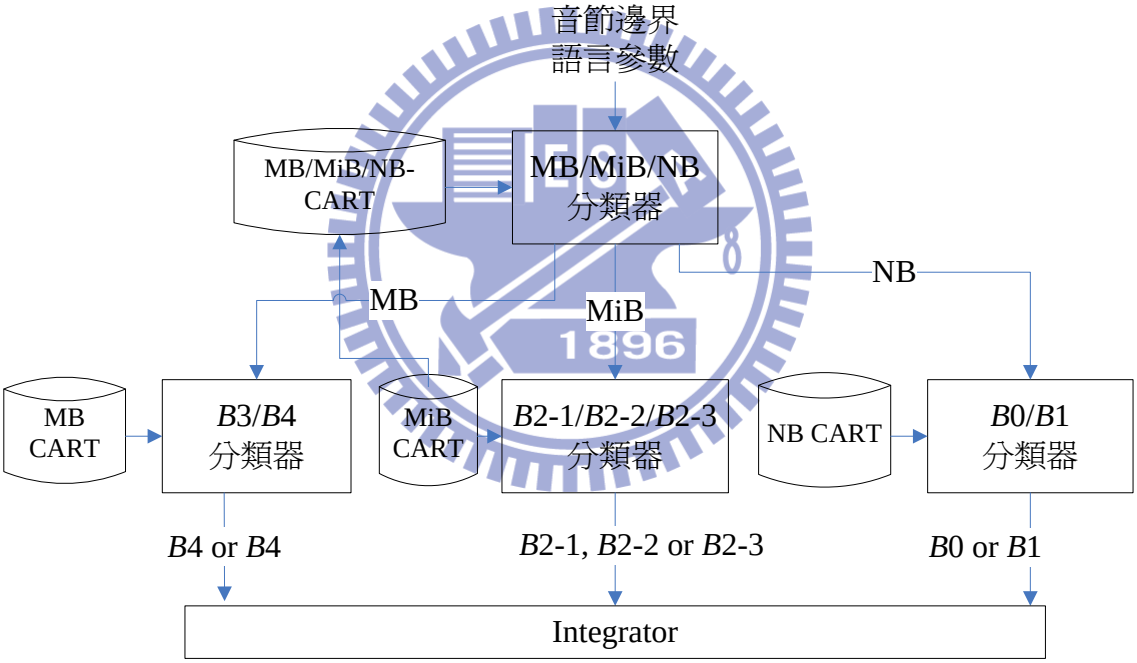


圖 5.2：Two-stage 停頓標記預估之方塊圖

此方法的辨認結果列在表 5.4 與表 5.5。由表 5.4 可發現到 Minor Break 的辨認率大幅提升，由 62%提升至 81%，Major Break 小幅提升 5%，Non-Break 則是維持 96%的高辨認率。表 5.5 則顯示由三類辨認後進而辨認出的七類辨認率，可觀察到 B2-1、B2-2、B2-3 分別提升 19%、11%以及 22%，而 B3 與 B4 則是提升 7%和 6%，由以上結果可以證明 Two-stage 的方法的確可以大幅改善 Minor Break 的辨認率，並進而提升七類辨認率的結果，尤其在 B2-1、

B2-1 與 B2-3。而此方法之所以可行的原因在於，原本是將七類停頓標記對應到所有語言參數，但此方法先將七類簡化成三類，使得只需要用三類標記去對應原本那麼多的語言參數，故其辨認率自然會提升。

利用上述之音節停頓標記預估演算法，即可由文字直接產生韻律詞與韻律短語等邊界，並產生更豐富文脈相關資訊幫助傳統 HTS 韻律產生。利用此結果產生之韻律階層文脈相關資訊如表 5.6 所示。

表 5.4：Two-stage 之三類韻律標記預估辨認率

Tar\Pre	NB	MiB	MB	Total
NB	**96%**	4%	0%	8270
MiB	16%	**81%**	3%	2326
MB	2%	11%	**87%**	1104

表 5.5：Two-stage 之停頓標記預估辨認率

Tar\Pre	B0	B1	B2-1	B3	B4	B2-2	B2-3	Total
B0	**85%**	12%	1%	0%	0%	1%	1%	1625
B1	6%	**90%**	2%	0%	0%	1%	1%	6645
B2-1	3%	19%	**63%**	1%	0%	8%	5%	840
B3	2%	2%	2%	**73%**	7%	13%	1%	697
B4	0%	0%	0%	15%	**83%**	1%	0%	407
B2-2	1%	6%	9%	5%	0%	**75%**	3%	1069
B2-3	4%	24%	16%	0%	0%	18%	**38%**	417

表 5.6：韻律階層之文脈相關

Prosody level	/a:F_Syl_in_PW	Syllable position in a PW(forward)
	/b:B_Syl_in_PW	Syllable position in a PW(backward)
	/c:F_Syl_in_PPh	Syllable position in a PPh(forward)
	/d:B_Syl_in_PPh	Syllable position in a PPh(backwrad)
	/e:Pre_1_PW	Word length of previous PW
	/f:Cur_PW	Word length of current PW
	/g:Fol_1_PW	Word length of following PW
	/h:Pre_1_PPh	Word length of previous PPh
	/i:Cur_PPh	Word length of current PPh
	/j:Fol_1_PPh	Word length of following PPh
	/k:Cur_B_PW	Break type bounding current Prosodic word

5.2 PLM 之韻律參數預估

由 4.1 節中各韻律參數之 TRE 值得知，一旦挑選到適當的韻律狀態，其 TRE 值將降得很低，因此韻律狀態預估是否恰當，將決定最後韻律參數預估的好壞。在第三章中我們利用(3.21)式採逐項最佳化程序標記出最佳的韻律標記，但在實際語音合成時，並沒有聲學參數，意即並無音節韻律模型 $P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})$ 與停頓標記聲學模型 $P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$ 中的 **X** 、 **Y** 、 **Z** 等參數。因此我們修改(3.21)式，提出(5.1)式韻律標記預估之標的函數，以句子為單位，利用維特比演算法，動態搜尋出最佳音節停頓標記與音節韻律狀態。

$$\begin{aligned}&\mathbf{p}^*,\mathbf{q}^*,\mathbf{B}^*\\&=\arg\max_{\mathbf{p},\mathbf{q},\mathbf{B}}\left(P(p_1)P(q_1)\prod_{n=2}^{N}P(p_n\mid p_{n-1},B_{n-1})P(q_n\mid q_{n-1},B_{n-1})\right)\\&\quad\left(\prod_{n=0}^{N}\left(P(pj_n\mid B_n,\mathbf{l}_n)P(df_n\mid B_n,\mathbf{l}_n)P(B_n\mid\mathbf{l}_n)\right)\right)\left(\prod_{n=1}^{N}P(p_n\mid\mathbf{l}_n)P(q_n\mid\mathbf{l}_n)\right)\end{aligned}\tag{5.1}$$

其中 $P(p_n\mid p_{n-1},B_{n-1})$ 、 $P(q_n\mid q_{n-1},B_{n-1})$ 分別為音節基頻、音節長度之韻律狀態轉移模型，可由原本中英夾雜 PLM 模型中得到， $P(B_n\mid\mathbf{l}_n)$ 為停頓標記語言模型，其產生方法如同 5.1.1 節中產生 All-in-one CART 一樣，即利用語言參數與正確停頓標記之關係訓練一顆決策樹。而 $P(pj_n\mid B_n,\mathbf{l}_n)$ 、 $P(df_n\mid B_n,\mathbf{l}_n)$ 則為停頓標記聲學模型之基頻跳躍值(**pj**)與音節長度延長因子2(**df**)，這兩個模型產生方式則是如同第三章，採取 CART 演算法，使用分裂準則為最大概似函數增益，以不同停頓標記對應之聲學參數各自建立語言參數、停頓標記與聲學參數之間相關的決策樹，不同之處在於第三章中是每個葉節點裡同時有 **pj** 之 *pdf* 與 **df** 之 *pdf*，但在此是將兩種參數用不同決策樹來實現。然而在預估韻律標記時，並無真正基頻跳躍值與音節長度延長因子，因此我們將基頻跳躍值以前後音節基頻韻律狀態差($pj_n\approx\beta_{p_{n+1}}-\beta_{p_n}$)來取代，音節長度延長因子則可用前後音節長度韻律狀態差($df_n\approx\gamma_{q_{n+1}}-\gamma_{q_n}$)來表示，此外當 $n=0,N$ 時，即 pj_0,pj_N,df_0,df_N ，則額外考慮句首和句尾的基頻跳躍值與音節長度延長因子。 $P(p_n\mid\mathbf{l}_n)$ 音節基頻韻律狀態語言模型和 $P(q_n\mid\mathbf{l}_n)$ 音節長度韻律狀態語言模型則是為了預估韻律參數時

所額外加入的模型，其實現方式一樣採用 CART 演算法，使用分類準則為最大概似函數增益，建立 16 種韻律狀態和語言參數間的關係。透過(5.1)式，將可得到一序列音節停頓標記、音節基頻韻律狀態以及音節長度韻律狀態，最後每個音節的韻律參數由各音節之語言參數對應到相關之音節層次 APs，並加上預估之韻律狀態，即可疊加出來，其數學式如(5.2)與(5.3)式。

$$\mathbf{sp}_n^* = \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n^*} + \boldsymbol{\beta}^f_{B^*_{n-1},tp_{n-1}} + \boldsymbol{\beta}^b_{B^*_n,tp_n} + \boldsymbol{\mu} \tag{5.2}$$

$$sd_n^* = \gamma_{t_n} + \gamma_{s_n} + \gamma_{q_n^*} + \mu_d \tag{5.3}$$

此外音節間靜音停頓(short pause, sp)預測也是語音合成中重要的一環，因此我們利用 PLM 訓練模型中之停頓標記聲學模型 $P(pd_n|B_n,\mathbf{l}_n)$ 的停頓時長和停頓標記與語言參數的關係，由上述預估之停頓標記為 B2-2、B3 及 B4 的音節，找到 $P(pd_n|B_n,\mathbf{l}_n)$ 中對應之停頓時長的 *pdf*，並採用此 *pdf* 的平均值為此音節之停頓時長。

最後在此補充說明(5.1)式之物理意義，$P(p_n|\mathbf{l}_n)$、$P(q_n|\mathbf{l}_n)$ 可視為韻律狀態和語言參數間的靜態(static)模型，可以確保韻律狀態之定位點，若不加入此模型可能會造成韻律狀態飄移不定，$P(p_n|p_{n-1},B_{n-1})$、$P(q_n|q_{n-1},B_{n-1})$ 則是限制了韻律狀態轉移的關係，可以避免韻律狀態不合理的轉移，$P(pj_n \approx \beta_{p_{n+1}}-\beta_{p_n}|B_n,\mathbf{l}_n)$、$P(df_n \approx \gamma_{q_{n+1}}-\gamma_{q_n}|B_n,\mathbf{l}_n)$ 則可視為韻律狀態和停頓標記與語言參數間的動態(dynamic)模型，利用此模型可以描述韻律狀態在不同停頓標記與語言參數下的跳動變化。

5.3 語音合成實驗結果與分析

本小節將比較使用傳統 HTS 方法與前兩小節所提出之韻律產生方法作比較，並採用 I. 客觀評估(方均根誤差)與 II.主觀評估(MOS, Preference)評量各種韻律產生方法。

第一種方法為使用傳統 HTS 韻律產生的方法，即第二章所介紹之方法，且只在標點符號處給予靜音停頓，稱此方法為 HTS-PM。然而靜音停頓的預估好壞對語音合成的流暢度是

相當重要的，因此本研究利用文字和靜音停頓(sp≥25ms)關係，訓練一顆靜音停頓預估決策樹，透過文字直接預估每個音節後使否有靜音停頓，並結合 HTS 文本標示，由前後文脈相關語言參數預估靜音停頓長度與產生語音，稱此方法為 HTS-SP。第三種方法為利用 5.1.2 節 Two-stage 之演算法預估停頓標記，藉此產生更豐富的韻律層次文脈相關資訊，並在 B2-2、B3 及 B4 才給予靜音停頓，最後利用 HTS 演算法產生韻律參數，稱此方法為 HTS-Break。最後一個方法則是採用 5.2 節的 PLM 之韻律產生方法，稱此方法為 PLM。

上述四種方法之訓練語料皆為中英夾雜語料庫 539 句中，拿掉編號尾數為 7 之句子，共 485 句，12185 個音節，包含 10504 個中文音節與 1681 個英文字母。測試語料皆為編號尾數為 7 之句子，共 54 句，1355 個音節，包含 1164 個中文音節與 191 個英文字母。以下將依序介紹客觀評估與主觀評估的結果。

I. 客觀評估：方均根誤差(Root Mean Square Error, RMSE)

整體訓練語料與測試語料之 RMSE 統計如 5.7 所示，其中 PLM-Correct Break 與 HTS-Correct Break 分別為正確停頓標記的情況下，透過 PLM 和 HTS-Break 方法的結果，而 HTS-Correct SP 則是正確靜音停頓標記的情況下，透過 HTS-SP 方法的結果。

表 5.7：整體語料之 RMSE 值

	Training		Testing	
	Pitch (LF0)	Duration (ms)	Pitch (LF0)	Duration (ms)
HTS-PM	0.1269	49.1	0.1393	48.9
HTS-SP	0.1216	47.5	0.1360	48.4
HTS-Correct SP	0.1178	46.4	0.1351	47.1
HTS-Break	0.1221	45.8	0.1375	46.4
HTS-Correct Break	0.1142	44.6	X	X
PLM	0.1335	39.3	0.1429	44.3
PLM-Correct Break	0.1231	37.4	X	X

由表 5.7 顯示在訓練語料中以 HTS-Correct Break 在基頻的 RMSE 會有最好結果，其次依照 RMSE 值由小到大分別為：HTS-Correct SP <HTS-SP< HTS-Break <PLM-Correct Break < HTS-PM< PLM。而測試語料之基頻 RMSE 值由小到大分別為 HTS-SP<HTS-Break< HTS-PM<

PLM。至於音長在訓練語料中的 RMSE 則是 PLM-Correct Break 表現最佳，其次依照 RMSE 值由小到大分別為：PLM < HTS-Correct Break < HTS-Break < HTS-Correct SP <HTS-SP<HTS-PM。而測試語料之音長 RMSE 則如同訓練語料中的結果，以 PLM 表現最佳，其次依序為 HTS-Break<HTS-SP<HTS-PM。

以上結果也顯示了給予正確 Break 的情況下，因加入更多韻律層次的資訊，HTS-Correct Break 在基頻與音長預估上確實能改善 HTS-PM 和 HTS-SP，然而一但停頓標記預估不夠準確的情況下(HTS-Break)，其基頻預估效能就會略輸給 HTS-SP，但仍然比 HTS-PM 來的要好，代表錯誤的停頓標記比起錯誤的靜音停頓所帶給基頻的影響更大，至於本研究所提出之 PLM 方法在預估基頻時的效能並不佳，代表在此方面仍需加強改進，但在音節長度的預估上則是比起各種 HTS 的方法都要來得出色，推測會有此結果的原因在於，根據第四章韻律模型訓練結果中的表 4.2 與 4.3 之 TRE 值顯示，基頻之 TRE 值還有 7.53%，而音節長度則只剩下 4.3%，意即原本韻律模型中，基頻韻律狀態的標記就沒有音節長度來的準確，另外根據【10】中對於全中文語料之韻律標記結果，可將基頻 TRE 值降到 1.1%，這也顯示中英夾雜語料庫在韻律標記上仍然有進步的空間，若是能改進此部分，使用 5.2 節所提出基於 PLM 演算法之韻律參數預估將更加準確。分析中英夾雜之 TRE 值偏高的可能原因為此語料庫的大小相對於全中文語料而言來的較小，因此若是透過增加語料量或是經由調適的方式或許能改善此缺點。

以上為整個語料之客觀評估分析，以下將中文與英文字母個別分開作客觀實驗的分析。表 5.8 與表 5.9 分別列出中文、英文字母之基頻 RMSE 與音長 RMSE 分開統計之結果，其中 English mean 是指預估英文字母的基頻與音長時，使用訓練語料中各英文字母的平均基頻與平均音長。觀察表 5.8 可發現在訓練語料中的中文之基頻 RMSE 如同表 5.7 結果一樣，仍以 HTS-Correct Break 最佳，並可知道英文字母的基頻誤差相對於中文來說是非常小的，這是相當合理的結果，畢竟所有中文字種類遠大於英文字母，故其中文基頻變化劇烈程度遠大於英文字母，因此預估的誤差自然較大。但在測試語料時，PLM 預估英文字母的能力則是比 HTS-PM 和 HTS-Break 來得較佳，但比 HTS-SP 還差。由此可知 PLM 雖然在中文預估能力沒有比其他方法好，但在英文字母的掌控上卻有不錯的效果。接著觀察表 5.9，在測試語料

時，其音長 RMSE 不論是中文或是英文字母，都以 PLM 為最佳，其次為 HTS-Break，再來為 HTS-SP 和 HTS-PM，而此結果也證明在加入停頓標記後，多了韻律層次的資訊的確可以提升音節長度的預估能力。

此外觀察使用各英文字母平均(English mean)的結果，可發現到在基頻部分則是比各方法都要來的差很多，意外的是音長部分只比 PLM 要來的差，但這不代表使用各英文字母的平均長度所合成出的語句就會比較自然。

表 5.8：中文、英文字母之基頻 RMSE

	Training LF0 RMSE		Testing LF0 RMSE	
	Chinese	English	Chinese	English
English mean	X	0.1201	X	0.1119
HTS-PM	0.132	0.0881	0.1449	0.0987
HTS-SP	0.1266	0.0833	0.1415	0.0954
HTS-Correct SP	0.1227	0.0808	0.1405	0.0956
HTS-Break	0.1268	0.0873	0.1444	0.1012
HTS-Correct Break	0.1186	0.0822	X	X
PLM	0.1393	0.0896	0.1487	0.0985
PLM- Correct Break	0.1281	0.0855	X	X

表 5.9：中文、英文字母之音長 RMSE

	Training duration RMSE (ms)		Testing duration RMSE (ms)	
	Chinese	English	Chinese	English
English mean	X	42	X	39.2
HTS-PM	49.1	48.2	49.6	46
HTS-SP	48	44.8	48.6	43.3
HTS-Correct SP	46.9	41.8	48.1	42.2
HTS-Break	46.2	42.4	47.3	40.9
HTS-Correct Break	44.8	41.9	X	X
PLM	39.8	34.3	45.2	38.5
PLM- Correct Break	37.9	34.5	X	X

圖 5.3 與圖 5.4 則是原始韻律參數(音節平均基頻軌跡、音節長度)和四種方法預估出來的比較圖範例一與範例二，其中 Predict Break 為 PLM 方法所預估出 B2 類以上的停頓標記。

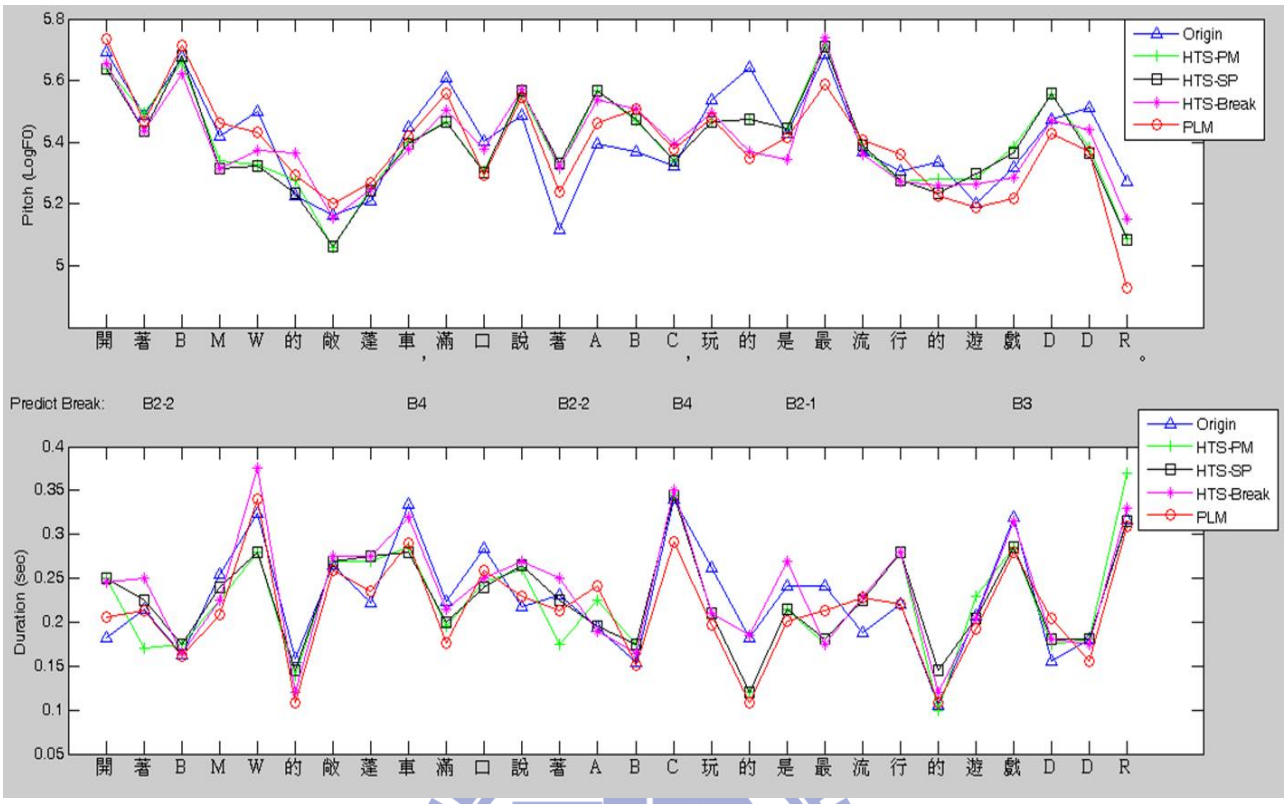


圖 5.3：原始韻律參數和四種方法預估出來的比較圖範例一。(a)音節平均基頻值(b)音節長度

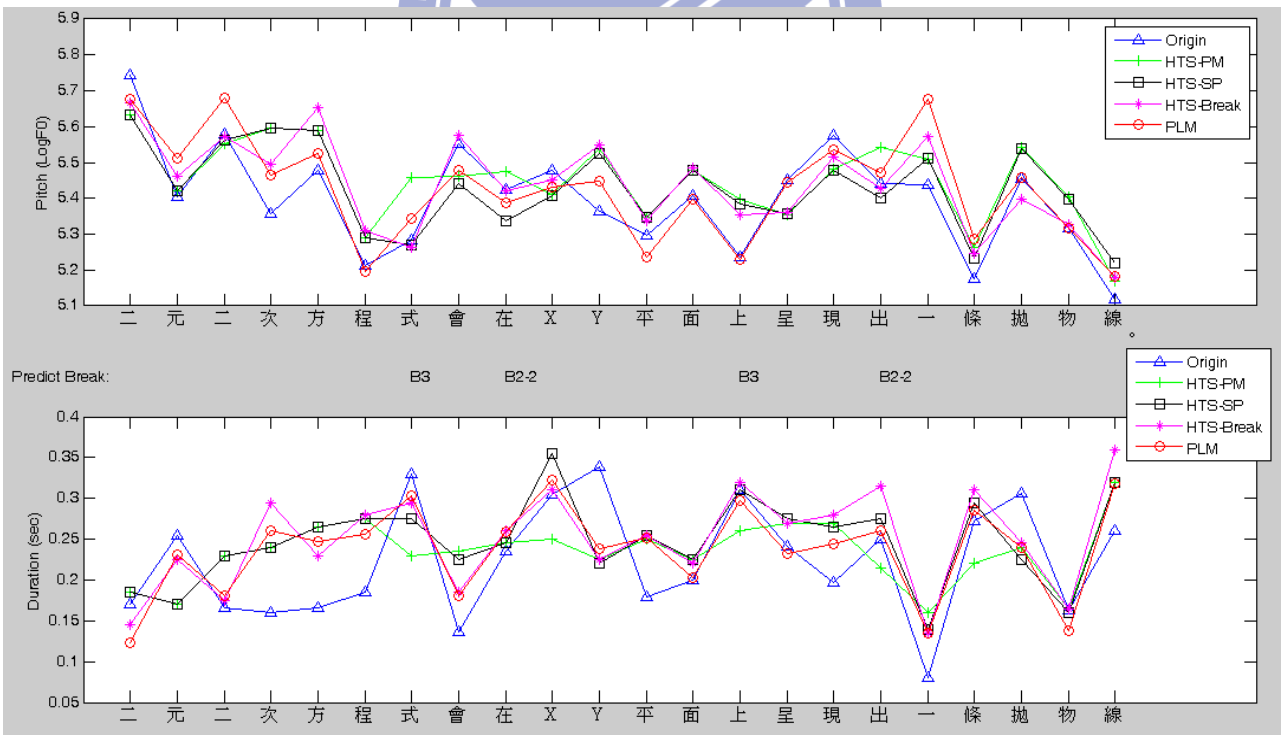


圖 5.4：原始韻律參數和四種方法預估出來的比較圖範例二。(a)音節平均基頻值(b)音節長度

II. 主觀評估：

(1) 平均主觀值分數(Mean Opinion Source, MOS)：

四種方法各自挑選相同編號的 15 句合成音檔，交由九位受試者評估四種方法在聽覺上自然度的分數。每組編號相同的音檔是採取隨機撥放的方式，交由受試者評估四種方法的自然度分數，評分標準如表 5.10 所示：

表 5.10：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

MOS 評分結果如圖 5.5 所示：

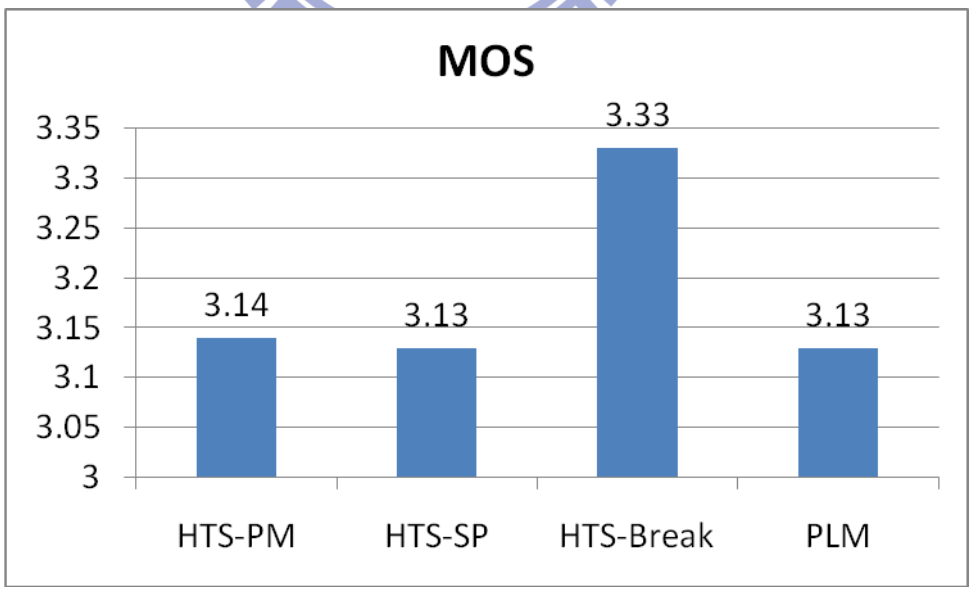


圖 5.5：MOS 主觀評估結果

由圖 5.5 可觀察到 MOS 量測的結果以 HTS-Break 在聽覺上的自然度有最好的表現，其餘三種方法則是不相上下。

(2) 偏好測定(Preference test)：

將四種方法兩兩一組，產生六組配對，每組配對中各方法選取相同編號的 15 句合成音檔，由九位受試者進行偏好測定實驗。每組編號相同的音檔是採取隨機撥放的方式，交由受試者選取兩句中比較喜歡的一句，即認為聽覺上比較自然的語句，評估結果如圖 5.6：

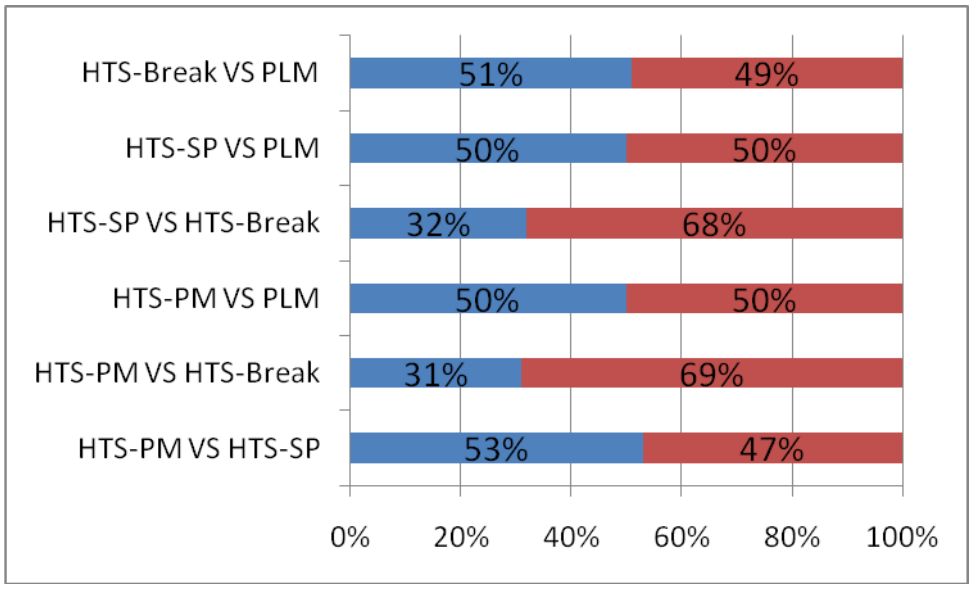


圖 5.6：偏好測定評估結果

由圖 5.6 可觀察到在偏好測定中，仍然以 HTS-Break 有最好的結果，而其餘方法仍是不相上下，此結果也符合 MOS 評量的結果。

綜合以上客觀評估與主觀評估的結果，可觀察出四種方法在客觀評估上各有好壞，如 HTS-SP 在基頻預測中有較好的結果，而 PLM 則是在音長預測中有最好的表現，但真正聽覺上的自然度則是以 HTS-Break 有最好的結果，這也意味著四種方法在基頻與音長預測上的差異並不是主要影響聽覺感受的原因，關鍵則是在於靜音停頓的預估，不恰當的靜音停頓會造成語流上的中斷，進而影響受試者在自然度上的評分，也代表我們所預估出的停頓標記確實能更準確地抓到整個韻律變化的節奏，使受試者感受到更自然的語音。

第六章 結論與未來展望

6.1 結論

本論文利用聲學參數以及語言參數，建立一個中英夾雜語料(英文字母鑲嵌於中文文句中)的韻律模型，並透過此模型自動完成韻律標記。藉由此韻律模型的分析，也證明了華人在念中英文夾雜時，英文字母往往會有Tone Borrowing之現象，且上層韻律變化是受到整體中文韻律架構的影響與限制。此模型也嘗試在有限語料下，使用更一般化的方式來模擬中英夾雜之連音現象。此外我們也發現在code-switch處且為詞外邊界時，其停頓標記之強度通常較大(B2-2/B3/B4)，代表在code-switch時，此語者會利用韻律斷點較強的停頓標記來表示中英文詞的轉換。更重要的是，本論文也發現，停頓標記的分佈情形不僅僅只受到前後音節是否為不同語言的影響，更受到整個PPh、BG/PG等大韻律單元之間的停頓標記相對關係影響。

藉由此韻律模型，本研究提出兩種中英夾雜文句之韻律產生器的方法，第一種方法為利用停頓標記的預估，產生更多韻律層次的文脈相關資訊，幫助傳統HTS方法提高韻律預估的效能。第二種方法則是利用韻律模型直接由文字預估停頓標記與韻律狀態，進而產生韻律參數，最後結合HTS所產生頻譜參數(MGC)，透過MLSA filter還原語音，完成整個中英夾雜語音合成系統。由客觀評估的實驗結果得知，第一種方法確實能改善使用傳統HTS在韻律參數上的預估能力，而第二種方法則是在音長預測中有很好的結果，而主觀評估的結果也顯示第一種方法在聽覺上有最佳的自然度表現，代表透過本研究所預估的停頓標記能抓到更自然的韻律節奏變化。

6.2 未來展望

由於中英夾雜語料量的不足，使得韻律模型在標記韻律狀態的能力較差，若是能透過增加語料量，或是利用中文語料庫來調適此模型，相信可以建立更強健的韻律模型，使得模型訓練後的 TRE 值降得更低，並進一步應用在韻律參數預估時，提高音節基頻的預估能力。

本研究在預估停頓標記與韻律參數時，目前並沒有引入大韻律單元內停頓標記間的相對關係，未來希望能將此部份量化成具有相對物理意義的數學式，相信此作法可以再提升停頓標記與韻律參數的預估能力。而目前已完成英文字母之中英夾雜語音合成系統，希望在未來日子裡能再完成逐音標發音之中英夾雜語音合成系統。

參考文獻

【1】 F. Deprez, J. Odijk, and J. D. Moortel, “Introduction to Multilingual Corpus-based Concatenative Speech Synthesis,” Proc. of Interspeech, pp.2129-2132, August 2007.

【2】 M. Chu, H. Peng, Y. Zhao, Z. Y. Niu, and E. Chang, “Microsoft Mulan - A Bilingual TTS System, ” *Proc. of ICASSP*, vol.1, pp.264-267, 2003.

【3】 A. W. Black, and K. A. Lenzo, “Multilingual Text-to-Speech Synthesis,” *Proc. of ICASSP*, vol.3, pp.761-764, 2004.

【4】 Wei-Chih Kuo, Yih-Ru Wang, Hung-Mao Lu, and Sin-Horng Chen, “An NN-based Approach to Prosody Generation for English Word Spelling in English-Chinese Bilingual TTS, ” in *Eurospeech*-2003, 3109-3112

【5】 Sin-Horng Chen, Shaw-Hwa Hwang, and Yih-Ru Wang, “An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to-Speech,” *IEEE Trans*. Speech Audio Processing, vol.6, no.3, pp.226-239,1998.

【6】 Yi Zhang, Jianhua Tao, “Prosody Modification on Mixed-Language Speech Synthesis, ” Chinese Spoken Language Processing, 2008 *ISCSLP*

【7】 Hui Liang, Yao Qian, Frank K. Soong, Gongshen Liu, “A Cross-Language State Mapping Approch to Bilingual (Mandarin-English) TTS,” *ICASSP* 2008

【8】 T. A. Myrvoll, and F. K. Soong, “Optimal Clustering of Multivariate Normal Distributions Using Divergence and Its Application to HMM Adaptation,” *Proc. of ICASSP*, vol.1,pp.552-555, April 2003.

【9】 Y. Zhao, C. Zhang, F. K. Soong, M. Chu, and X. Xiao, “Measuring Attribute Dissimilarity with HMM KL-Divergence for Speech Synthesis, ” *Proc. of the 6th ISCA* Speech Synthesis Workshop, pp.206-210, August 2007

【10】 江振宇，“非監督式中文語音韻律標記及韻律模式”，國立交通大學博士論文，民國

九十八年三月

【11】 S. Imai, “Cepstral analysis synthesis on the mel frequency scale,” *Proc. of ICASSP*, pp.93–96, Feb. 1983.

【12】 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.

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

【14】 S.H. Chen and Y.R. Wang, “Vector Quantization of Pitch Information in Mandarin Speech”, *IEEE Transactions on Communications*, Vol. 38, No. 9, pp. 1317-1320, 1990.

【15】 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

【16】 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

【17】 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/

【18】 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

【19】 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

【20】 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).

【21】 Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi ,and Tadashi

Kitamura , “Speech parameter generation algorithms for HMM-Based speech synthesis” *Proc. of ICASSP*, pp.1315-1318, June 2000

【22】 吳仲耘，“應用韻律階層及動態參數之音高預測在基於HMM之中文語音合成器”，國立成功大學碩士論文，民國九十七年七月。

附錄一

問題集 Θ_1與問題集Θ_2分別為連音參數(Forward,Backward)兩顆決策樹之問題集。

Θ_1設計概念是以兩大分類方式：

(1)：相似聲調的合併，又分(a)當前音節的合併與(b)前一個音節基頻結尾高度相似的合併

(a)當前音節的合併：{tone1}、{E-t1}、{tone1,E-t1}、{tone2}、{E-t2}、{tone2,E-t2}、{tone3}、{tone4}、{M_f}、{tone4,M_f}、{tone5}

(b)前一個音節的合併：{tone1,E-t1}、{tone2,E-t2}、{tone3, tone5}、{tone4,M_f}

其中 E-t1，E-t2 分別為 Tone 1 Borrowing 和 Tone 2 Borrowing 之英文字母

(2)：依停頓標記強度合併，共分成{B0}、{B0,B1}、{B0,B1,B2-1}、{B0,B1,B2-1,B2-3}、{B2-2},{B2-2,B3,B4},{B-stat}。

透過此方式可設計出 275 個有關於 Forward Coarticulation 之問題集，以下列舉部分問題集：

Q：Pre-t1,E-t1 接(B0,B1,B2-1,B2-3)接 t1,E-t1

Q：Pre-t1,E-t1 接(B0)接 E-t1

Q：Pre-t2,E-t2 接(B0,B1)接 t4,M_f

Q：Pre-t2E-t2 接(B0,B1,B2-1)接 t3

Q：Pre-t2,E-t2 接(B3,B4,B2-2)接 t1,E-t1

Q：Pre-t3,t5 接(B0,B1)接 t1,E-t1

Q：Pre-t3,t5 接(B3,B4,B2-2)接 E-t1

Q：Pre-t3,t5 接(B0,B1,B2-1,B2-3)接 E-t2

Q：Pre-t4, M_f 接(B0,B1,B2-1,B2-3)接 t3

Q：Pre-t4, M_f 接(B0,B1)接 M_f

Q：Pre-t4, M_f 接(B0,B1,B2-1,B2-3)接 t5

Θ_2 設計概念則是和 Θ_1 雷同，一樣分成兩大類：

(1)：相似聲調的合併，又分(a)當前音節的合併與(b)下一個音節基頻起始高度相似的合併

(a)當前音節的合併：{tone1}、{E-t1}、{tone1,E-t1}、{tone2}、{E-t2}、{tone2,E-t2}、{tone3}、{tone4}、{M_f}、{tone4,M_f}、{tone5}

(b)前一個音節的合併：{tone1,tone4,E-t1, M_f }、{tone2, tone3,tone5,E-t2,WZ}

(2)：依停頓標記強度合併，共分成{B0}、{B0,B1}、{B0,B1,B2-1}、{B0,B1,B2-1,B2-3}、{B2-2},{B2-2,B3,B4},{B-end}。

其中 tone 3 接 fol-tone3 則是獨立考慮。

透過此方式可設計出 152 個有關於 Backward Coarticulation 之問題集，以下列舉部分問題集：

Q：t1,E-t1 接(B0,B1,B2-1,B2-3)接 fol-t1,t4,E-t1,M_f

Q：E-t2 接(B0,B1)接 fol-t1,t4,E-t1,M_f

Q：M_f 接(B3,B4,B2-2) 接 fol-t2,t3,t5,E-t2,Z,W

Q：t3 接(B0,B1) 接 fol-t3

Q：E-t2 接(B0,B1,B2-1)接 fol-t2,t3,t5,E-t2,Z,W

Q：E-t2 接(B3,B4,B2-2) 接 fol-t2,t3,t5,E-t2,Z,W'

Q：t3 接(B0,B1,B2-1,B2-3)接 fol-t3

Q：t4, M_f 接(B0,B1,B2-1,B2-3)接 fol-t2,t3,t5,E-t2,Z,W

Q：M_f 接(B0,B1) 接 fol-t2,t3,t5,E-t2,Z,W

Q：t5 接(B0,B1,B2-1) 接 fol-t1,t4,E-t1, M_f

Q：E-t1 接(B0,B1) 接 fol-t2,t3,t5,E-t2,Z,W

Q：t1,E-t1 接(B3,B4,B2-2) 接 fol-t2,t3,t5,E-t2,Z,W

Q：E-t2 接(B2-2)接 fol-t2,t3,t5,E-t2,Z,W

Q：t4, M_f 接(B0) 接 fol-t1,t4,E-t1, M_f

Q：t3 接(B3,B4,B2-2) 接 fol-t3

附錄二

The question set Θ_3 used to construct the decision trees for building the break-acoustic model $p(pd_n, ed_n, pj_n, dl_n, df_n | B_n, \mathbf{l}_n)$ and break-syntax model $P(B_n | \mathbf{l}_n)$ is listed below:

1. Syllable Level

$Q_3$1.1: Is the initial of the following syllable a null one or in {*m, n, l, r*}?

$Q_3$1.2 : Is the initial of the following syllable a null one?

$Q_3$1.3 : Is the initial of the following syllable in {*b, d, g*}?

$Q_3$1.4 : Is the initial of the following syllable in {*f, s, sh, shi, h*}?

$Q_3$1.5 : Is the initial of the following syllable in {*m, n, l, r*}?

$Q_3$1.6 : Is the initial of the following syllable in {*ts, ch, chi*}?

$Q_3$1.7 : Is the initial of the following syllable in {*tz, j, ji*}?

$Q_3$1.8 : Is the initial of the following syllable in {*p, t, k*}?

$Q_3$1.9 Is the initial of the following syllable a null one{A,E,I,O,R,U,V,Y,F,H,S,X,L,M,N }?

$Q_3$1.10 : Is the initial of the following syllable in {B, D, W}?

$Q_3$1.11: Is the initial of the following syllable in {C}?

$Q_3$1.12 : Is the initial of the following syllable in {G,J,Z}?

$Q_3$1.13: Is the initial of the following syllable in {P, T, K, Q}?

$Q_3$1.14 Is the initial of the following syllable a null one or in {*m, n, l, r,* M,N A,E,I,O,R,U,V,Y,F,H,S,X,L }?

$Q_3$1.15 Is the initial of the following syllable in {*b, d, g,* B, D, W}?

$Q_3$1.16 Is the initial of the following syllable in {*f, s, sh, shi, h,* C}?

$Q_3$1.17 Is the initial of the following syllable in {*tz, j, ji,* G, J, Z}?

$Q_3$1.18 Is the initial of the following syllable in {*p, t, k,* P, T, K, Q}?

$Q_3$1.19 : Is the inter-syllable location an inter-word?

: Is the inter-syllable location a Type-1 intra-word?

$Q_3$1.20 : Is the inter-syllable location a Type-2 intra-word?

$Q_3$1.21 是否出現在 code switch 處?

$Q_3$1.22 code switch,中文+英文，且為 Inter-word (C-E-Inter)?

$Q_3$1.23 code switch, 英文+中文，且為 Inter-word (E-C-Inter)?

$Q_3$1.24 code switch,{中文+英文，英文+中文}，且為 Inter-word (C-E-Inter, E-C-Inter)?

$Q_3$1.25 code switch, 中文+英文，且為 Type1 Intra-word (C-E-Intra1)?

$Q_3$1.26 code switch,英文+中文，且為 Type1 Intra-word (E-C-Intra1)?

$Q_3$1.27 code switch,{中文+英文，英文+中文}，且為 Type1 Intra-word (C-E-Intra1, E-C-Intra1)?

$Q_3$1.28 code switch,中文+英文，且為 Type2 Intra-word (C-E-Intra2)?

$Q_3$1.29 code switch,英文+中文，且為 Type2 Intra-word (E-C-Intra2)?

$Q_3$1.30 code switch,{中文+英文，英文+中文}，且為 Type2 Intra-word (C-E-Intra2, E-C-Intra2)?

$Q_3$1.31 Non code switch,中文+中文，且為 Inter-word (C-C-Inter)?

$Q_3$1.32 Non code switch,英文+英文，且為 Inter-word(E-E-Inter)?

$Q_3$1.33 Non code switch,{中文+中文，英文+英文}，且為 Inter-word (C-C-Inter, E-E-Inter)?

$Q_3$1.34 Non code switch, 中文+中文，且為 Type1 Intra-word (C-C-Intra1)?

$Q_3$1.35 Non code switch,英文+英文，且為 Type1 Intra-word (E-E-Intra1)?

$Q_3$1.36 Non code switch,{中文+中文，英文+英文}，且為 Type1 Intra-word (C-C-Intra1,E-E-Intra1)?

$Q_3$1.37 Non code switch,中文+中文，且為 Type2 Intra-word (C-C-Intra2)?

$Q_3$1.38 Non code switch,英文+英文，且為 Type2 Intra-word (E-E-Intra2)?

$Q_3$1.39 Non code switch,{中文+中文,英文+英文}，且為 Type2 Intra-word (C-C-Intra2, E-E-Intra2)?

2. Word Level

All the following questions are subject to a prerequisite condition that the current inter-syllable location is an inter-word.

2.1 PM

In the following questions, we define major PMs ={“。”, “！”, “；”, “？”} and minor PMs={“，”, “、”, “：”, ”-”, ”」”, ”「”}.

$Q_3$2.1.1: Does a PMs exist at the inter-syllable location?

$Q_3$2.1.2: Does a major PM exist at the inter-syllable location?

$Q_3$2.1.3: Does a minor PM exist at the inter-syllable location?

$Q_3$2.1.4: Does a comma exist at the inter-syllable location?

$Q_3$2.1.5: Does a period exist at the inter-syllable location?

$Q_3$2.1.6: Does an exclamation exist at the inter-syllable location?

$Q_3$2.1.7: Does a semi colon exist at the inter-syllable location?

$Q_3$2.1.8: Does a question mark exist at the inter-syllable location?

$Q_3$2.1.9: Does a dot exist at the inter-syllable location?

$Q_3$2.1.10: Does colon exist at the inter-syllable location?

$Q_3$2.1.11: Does a hyphen exist at the inter-syllable location?

$Q_3$2.1.12: Does a parenthesis exist at the inter-syllable location?

$Q_3$2.1.13: Does a dot or colon exist at the inter-syllable location?

2.2 Word length

$Q_3$2.2.1~4: Is the preceding word an $n \in \{1, 2, 3, 4\}$-syllable word?

$Q_3$2.2.5~8: Is the following word an $n \in \{1, 2, 3, 4\}$-syllable word?

$Q_3$2.2.9: Is the length of the preceding word in syllable greater than 4?

$Q_3$2.2.10: Is the length of the following word in syllable greater than 4?

2.3 Substantive/function words

$Q_3$2.3.1~2: Is the preceding word a substantive word/function words?

$Q_3$2.3.3~4: Is the following word a substantive word/function words?

2.4 Level-1 POS and special tags

$Q_3$2.4.1~11: Is the POS of the preceding word A/C/D/N/I/P/T/V/DE/SHI/DM?

$Q_3$2.4.12~22: IS the POS of the following word A/C/D/N/I/P/T/V/DE/SHI/DM?

2.5 Level-2 POS

$Q_3$2.5.1~33 : Is the POS of the preceding word Ca/Cb/Da/Db/Dc/Dd/Df/Dg/Dh/Di/Dj/Dk/Na /Nb/Nc/Nd/Ne/Nf/Ng/Nh/VA/VB/VC/VD/VE/VF/VG/VH/VI/VJ/VK/VL/V_2?

$Q_3$2.5.34~66 : Is the POS of the following word Ca/Cb/Da/Db/Dc/Dd/Df/Dg/Dh/Di/Dj/Dk/Na /Nb/Nc/Nd/Ne/Nf/Ng/Nh/VA/VB/VC/VD/VE/VF/VG/VH/VI/VJ/VK/VL/V_2?

2.6 Level-3 POS

$Q_3$2.6.1~15 :Is the POS of the preceding word Caa/Cab/Cba/Cbb/Dfa/Dfb/Ncd/Neu/Nes/Nep /Neq/VA2/VC1/VH16/VH22?

$Q_3$2.6.16~30 : Is the POS of the following word Caa/Cab/Cba/Cbb/Dfa/Dfb/Ncd/Neu/Nes

/Nep/Neq/VA2/VC1/VH16/VH22?

2.7 Combination of POS

Q_3 2.7.1~7: Does the POS of the preceding word belong to {Da, Db, Dc, Dd, Dg, Dh, Di, Dj, Dk}/{Na, Nb, Nc}/{Ncd, Ng}/{I, T}/{VA, VG}/{VB, VC, VD, VE, VF, VJ, VK, VL}/{VH, VI}?

Q_3 2.7.8~14: Does the POS of the following word belong to {Da, Db, Dc, Dd, Dg, Dh, Di, Dj, Dk}/{Na, Nb, Nc}/{Ncd, Ng}/{I, T}/{VA, VG}/{VB, VC, VD, VE, VF, VJ, VK, VL}/{VH, VI}?

附錄三

(1) *Th*1、*Th*2 和 *Th*3 的定義

*Th*1、*Th*2 和 *Th*3 是分別用來界定 B4、B3、B2-2 和 B0/B1 停頓時長的 threshold。由於 B4 和 B3 長長是標點符號的邊界，有較長的停頓時長，因此可將此類型音節邊節的停頓時長收集起來，以 vector quantization(VQ)分成兩類，用 Gamma distribution 去 fitting，令 mean 比較大的一群為 B4 之機率分佈 $f_{B4}(pd)$，另一個則為 B3 之機率分佈 $f_{B3}(Pd)$。另外由於 B0 和 B1 的停頓時長通常都比較短，因此將 intra-word 音節邊界的停頓時長收集起來，用 Gamma distribution 去 fitting，得到機率分佈 $f_{B0/B1}(pd)$。最後將屬於非標點符號之 inter-word 邊界的停頓時長收集起來，一樣使用 Gamma distribution 去 fitting，得到 B2-2 的機率分佈 $f_{B2-2}(pd)$，由於 B2-2 被定義為有明顯的韻律詞邊界，因此我們再加上 $f_{B3}(pd_n) > f_{B0/B1}(pd_n)$ 的條件，藉此將停頓時長太短的非標點符號之 inter-word 過濾，並將不滿足這條件的停頓時長歸類到 B0/B1。最後令 $f_{B0/B1}(pd)$、$f_{B2-2}(pd)$、$f_{B3}(Pd)$ 和 $f_{B4}(pd)$ 的交叉點分別為 *Th*3、*Th*2 和 *Th*1。

(2) *Th*5 的定義

*Th*5 是用來界定 B2-1 和 B0/B1 的 threshold，由於 B2-1 和 B0/B1 在停頓時長的差距不大，但是 B2-1 有明顯的音高重置，因此定義 normalized log-F0 level jump 如下式：

$$\xi_n = (\mathbf{sp}_{n+1}(1) - \boldsymbol{\beta}_{t_{n+1}}(1)) - (\mathbf{sp}_n(1) - \boldsymbol{\beta}_{t_n}(1))$$

$\mathbf{x}(1)$ 定義 $\mathbf{x}$ 向量的第一維度，藉由平均各個聲調的音高可以求得五種聲調的 $\boldsymbol{\beta}_t$。接著將 intra-word 邊界的邊界參數基頻差收集起來並用高斯分佈 fitting，得到 $f_{\text{intra}}(\xi)$；同時也將標點符號邊界的基頻差收集起來用高斯分佈 fitting，得到 $f_{\text{PM}}(\xi)$，再來將非標點符號之 inter-word 邊界的基頻差收集起來歸類為 B2-1，利用我們已知 B2-1 有明顯的音高重置，因此再加上條件 $f_{\text{PM}}(\xi) > f_{\text{intra}}(\xi)$，滿足此條件者我們將其收集起來並用高斯分佈 fitting，得到

$f_{\text{B2-1}}(\xi)$，最後令 *Th*5 為 $f_{\text{intra}}(\xi)$和 $f_{\text{B2-1}}(\xi)$的交叉點。

(3) *Th*4 和 *Th*6 的定義

在這個部分要從 B0/B1 這類資料再細分出 B0 和 B1，然而我們知道由於 B0 音節邊界屬於 tightly coupling，其連音情形比 B1 嚴重，導致音高停頓(pitch pause)比較短且 engery-dip 也比較大，因此我們用 *Th*4 作為 F0 pause duration threshold，*Th*6 作為 engery-dip level threshold，達到區分 B0 和 B1 的目的。令 *Th*4 為 1 個 frame 長(=10ms)，意即被歸類為 B0 的音高停頓長度為零，接著將剩餘未分類的資料用 VQ 將其 engery-dip 分為兩類，用高斯分佈去 fitting 其 engery-dip，令 mean 比較大的那群為 B0，engery-dip 機率分佈為 $f_{\text{B0}}(Pe)$，而 mean 較小的那群為 B1，ngery-dip 機率分佈為 $f_{\text{B1}}(Pe)$，則 *Th*6 即為這兩個高斯機率分佈 $f_{\text{B0}}(Pe)$和 $f_{\text{B1}}(Pe)$的交叉點。

(4) *Th*7 和 *Th*8 的定義

*Th*7 和 *Th*8 是用來區分 B2-3 和 B0/B1，我們已知 B2-3 為 inter-word 音節邊界，有相對明顯的音節長度拉長效應，因此判斷是否屬於 B2-3 的依據在於正規化的音節長度拉長因子 1 和 2(即 dl_n 和 df_n)是否大於 *Th*7 和 *Th*8。首先將 intra-word 和標點符號音節邊界之邊界參數的正規化音節長度拉長因子收集起來用高斯分佈 fitting，分別得到四個高斯分佈 $\{ f^{dl}_{\text{intra}}(\tau)/ f^{df}_{\text{intra}}(\tau)\}$和$\{ f^{dl}_{\text{PM}}(\tau)/ f^{df}_{\text{PM}}(\tau)\}$，接著針對符合非標點符號、inter-word 且有明顯音節拉長效應的音節邊界，將其正規化的音節長度因子 1 和 2 的資料收集起來分類成 B2-3，用高斯分佈去 fitting 而得到$\{ f^{dl}_{\text{B2-3}}(\tau)/ f^{df}_{\text{B2-3}}(\tau)\}$，然而為了避免所收集到的資料其正規化音節長度拉長因子與 intra-word 音節邊界的情形相似，因此再增加了一個條件：

$f^{dl}_{\text{PM}}(\tau) > f^{dl}_{\text{intra}}(\tau)$和 $f^{df}_{\text{PM}}(\tau) > f^{df}_{\text{intra}}(\tau)$，藉此條件將非標點符號且為 inter-word，但不與 B2-3 音節邊界特性相似的資料過濾掉。最後，令 *Th*7 為 $f^{dl}_{\text{intra}}(\tau)$和 $f^{dl}_{\text{B2-3}}(\tau)$的交叉點；令 *Th*8 為 $f^{df}_{\text{intra}}(\tau)$和 $f^{df}_{\text{B2-3}}(\tau)$的交叉點。