國 立 交 通 大 學

電信工程研究所

碩 士 論 文

考慮語速影響之漢語韻律模型建立

與語音合成之應用

A Modeling of Speaking Rate Influences on Mandarin Speech Prosody and its Application to TTS

研究生：謝喬華

指導教授：王逸如 博士

中 華 民 國 一百零一 年 七 月

考慮語速影響之漢語韻律模型建立與語音合成之應用

A Modeling of Speaking Rate Influences on Mandarin Speech Prosody and its Application to TTS

研 究 生：謝喬華　　Student：Chiao-Hua Hsieh

指導教授 ：王逸如 博士　　Advisor：Dr. Yih-Ru Wang

國 立 交 通 大 學

電 信 工 程 學 系

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 of Science

in Communication Engineering

July 2012

Hsinchu, Taiwan, Republic of China

中 華 民 國 一百零一 年 七 月

考慮語速影響之漢語韻律模型建立與語音合成之應用

研 究 生：謝喬華　　指導教授：王逸如 博士

國立交通大學電信工程研究所碩士班

中文摘要

本論文提出一個新方法，考慮漢語說話速度對韻律變化的影響，建立一個語速相依的漢語階層式韻律模型(SR-HPM)。本方法修正了先前的非監督式韻律標記與模式(PLM)方法，將語速當作一新的連續獨立變數，讓韻律聲學參數及韻律模型參數受其影響。本研究之 SR-HPM 建構於一位專業女性播報員所錄製四種不同語速之平行語料庫。實驗結果顯示語速對於模型參數之影響符合現有的語言學知識，證實了本研究所提出之方法能系統化地量化語速對漢語韻律之影響。

最後將本研究所提出之韻律模型應用在文字轉語音上，我們製作了一個可控制語速的中文文字轉語音系統。實驗主觀測試結果顯示，我們所提出之方法在快、慢語速都明顯優於傳統 ML 為基礎的語速控制方法。

A Modeling of Speaking Rate Influence on Mandarin Speech Prosody and its Application to TTS

Student：Chiao-Hua Hsieh　　Advisor：Dr. Yih-Ru Wang

Institute of Communication Engineering
National Chiao Tung University

Abstract

In this thseis, a new approach of Mandarin-speech prosody modeling to consider the effects of speaking rate is proposed. The approach is a modification of previous prosody labeling and modeling (PLM) method to take speaking rate as a continuous independent vaiable and let prosodic-acoustic features and some parameters of prosodic models depend on it in order to account for its influences. A speaking rate-dependent hierarchical prosodic model (SR-HPM) is hence constructed from four speech corpra of a single female speaker with four different speaking rates. An analysis of the effects of speaking rate on the model parameters showed that they agreed well with our prior knowledge. So, the proposed approach provides a systematic and effective way to quantify the effects of speaking rate on Mandarin-speech prosody.

Last, an application to the prosody generation for Mandarin text-to-speech (TTS) is proposed. By using the well-trained SR-HPM, a speaking rate-controlled TTS system that can generate fluent speech for any given speaking rate is implemented.The subjective testing results indicated that the proposed methed was significantly better than the conveninal ML-based method for fast and slow rate.

致謝

對於本論文能順利完成，我要感謝很多人。首先是陳信宏老師，感謝陳老師這兩年的細心指導，即使身兼院長之職，仍常抽空與我共進研究，實在辛苦您了；感謝我的指導教授王逸如老師，王老師教我如何成為一個真正的研究生，而不是只會教作業的大學生，感謝您在這兩年所給予的扎實訓練。

接著我要感謝 707 的老大性獸，一個人掌控所有研究生的研究進度，感謝你不辭辛勞地指導我程式及研究上的問題，希望你在台北大學一切順利；感謝智合哥教我 Linux 指令，也恭喜你拿到博士學位了；感謝分享了實驗室很多八卦的輝哥；感謝帥氣的阿德哥在 8051 方面的指導，你真是我們北科的驕傲代表。在此也感謝前屆學長所給予的照顧：感謝最有學妹緣的文良，祝你早日練成 KOBE；感謝最愛宵夜攤的大胖，以後要常回來一起打球；感謝最愛周星馳系列喜劇電影的小瞎，相信你的研究終能修成正果；感謝健身王豆腐、舞林高手智障、707 小 AI 銘傑。感謝坐在我旁邊的 Syntax 之神睿詮，跟你一起修課讓我很安心；感謝人超好的 DD，以後我去華碩面試記得罩我；感謝臉胖身不胖的胖子軒，望你回國後四顆輪子能再次啟動；感謝南極來的聰明企鵝，學識淵博簡直是奇摩知識+；感謝樂觀開朗的 1 對 1 鬥牛球友昌祐；感謝人很隨和的雅婷；感謝籃球很秋的小高及很機靈的昂星。也感謝諸位”我最看好的”學弟妹：讓出 Battle 帳號的魔鬼筋肉人奕勳、聰明絕頂的優質子睿、脫臼不怕痛的柔道高手良基、講講話很很很有有韻律的小霸王蘇蘇蘇仲銘、愛裝可愛的可撈Ｄ～亞婉君。感謝我最愛的女朋友靖觀，一路上有妳的陪伴，讓我感到很溫暖，未來我們要繼續一起加油努力。感謝以上諸位這兩年的陪伴，我的碩士生涯因你們而感到充實快樂。

最後要感謝我父母對我從小的栽培，尊重我在求學階段的每一個決定，當我在外地求學時，不斷地給我關心及鼓勵，在此僅將此論文獻給你們。

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究動機

科技日新月益，各式創新的技術與應用使得生活越來越趨於便利與高效率，智慧型手機、平板電腦、衛星導航等電子資訊產品近年來被快速普及，成為現代人生活中不可或缺的一部份，而語音處理技術亦被大量使用在相關產品的人機介面溝通上，增加使用的便利性，例如語音訂票功能及電子有聲書等，人們可以用最直接、自然的方式與機器溝通，取代以往複雜的鍵盤輸入及文字輸出。

隨著隨藏式馬可夫模型為基礎(HMM-based)的文字轉語音(Text-To-Speech, TTS)技術興起，語音合成品質已有不錯的表現，當中語音合成該有的韻律掌握是重要的關鍵，然而不同的應用中會有不用的語速需求。以語音訂票系統為例，對於外籍人士或老年人來說，速度稍慢的語音，可供足夠的反應時間聽懂內容；而對於本國籍人士，提供稍快速度的語音可以節省使用者寶貴的時間。因此，為滿足這些實際的需求，即必須對不同語速的韻律做進一步的探討，以利開發多語速語音合成系統。

在自動語音辨識系統(Automatic Speech Recognition, ASR)方面，有很多先前的研究指出，一定程度慢速或快速的語音，會造成辨識效能大幅度下降，如果能夠在辨識系統中補償語速所造成的影響，則能提升極端語速語音的辨識效果。

1.2 文獻回顧

說話速度是一個很重要的韻律參數，其影響了很多語音現象，像是音節長度、停頓時長、基頻軌跡形狀、音素之間的 coarticulation 程度、停頓發生的機率等等。利用語音信號估計語速[1]，探索語速對於韻律及語言參數的影響[2-4]是一直被探討的議題。[2]採用階層式韻律架構對三種不同語速(快、中、慢)之平行語料庫做分析，實驗對於語速的測量分兩類，一類為 Speech

Rate(SR)，定義為每秒包含停頓長度的發音音節個數，另一類為 articulation rate(AR)，定義為每秒不包含停頓時長的發音音節個數。其實驗結果發現改變說話速度對各層韻律邊界的停頓時長為非線性的；語速的快慢會影響基頻軌跡(F0 contour)的分佈，快速語料的音高平均較高且變動範圍較小，而慢速語料的音高平均較低、變動範圍較大。此篇文章提出一些不錯的觀點，但語料庫的資料量不夠大，導致其分析結果不夠一致。[5]提出一階層式多短語韻律句群架構，並使用逐步迴歸(step-wise regreesion)來估算語料中語速對韻律階層單元的影響，共分析出三種不同語速之中文韻律詞、韻律短句及呼吸組層次的音長和音強 pattern；解析出各層次單元於音長及音強的貢獻。其實驗結果發現，音長的韻律詞 pattern 呈現一勺子狀曲線，以中速語料的延長/縮短效應最小，慢速最大；韻律短句在快、中速語料的短句有拉長現象而慢速卻沒有。音強方面，發現越長的韻律詞所需能量越大，慢速語料的平均音強為最大；韻律短語層次部份，發現越長的韻律短語需要越大的能量。此篇研究提出了不少新發現，但因其語料庫不為同一語者所發音，導致部份實驗結果有不一致現象。

在 ASR 及 TTS 的應用中，如何建立模型去考慮語速效應也是很重要的議題。在 ASR 方面，主要針對快速和慢速的語音做補償[6-9]，[6]提出對語速正規化方法，依據語速調整音框長度去求得動態頻譜特徵參數(dynamic spectral feature)，用此方式來補償語速對 ASR 的影響。[7]提出語速對於聲學模型的補償方法，依據每個音框所屬的語速去調整 HMM 的混合權重(mixture weight)和轉移機率。[8]發現語速快於某程度之語音會使 ASR 的辨識效能嚴重下降，並提出三種補償方法，分別為 Baum-Welch 碼本(codebook)的調適、HMM 轉移機率的調適、發音字典的修正，其中方法二使相對錯誤率(relative error rate)降低了 4-6%。[9]提出利用兩個平行特定語速的聲學模型去對快速與慢速語音做辨認，實驗結果提升了 1.9%的絕對正確率。

對於 TTS 來說，語速控制在人機介面的使用是必須的[10-12]，在一些特別的應用，例如快速的合成語音會較適合視障人士[13-14]。[10]進行大規模主觀測試三種語速控制的方法，分別為：(1)針對目標語速選取相近語速之語料來訓練 HMM 模型，(2)依比例去伸縮合成語句的發音長度，及(3)基於 ML 準則去決定狀態長度(state duration)，這些方法都是建立於 HMM-based 的語音合成系統，實驗結果發現方法(2)最適合用於快語速合成語音，而方法(1)較適合慢速語

音。[11]研究關於語速對於韻律參數造成的影響，進而將語速調整加入到中文 TTS 系統；其研究的重點在於韻律架構、音長、基頻分佈以及口音位置對於語速的變化。[12]提出控制音素時長(phoneme duration)的方法，以 HMM-based 合成系統為基礎，配合快速、正常、慢速音素時長模型之間的內插，達到目標語速的合成語音。

1.3 研究方向

本研究考慮語速對中文語音韻律的影響，提出的新方法將延續過去的研究。主要是基於[15]提出之非監督式中文語音韻律標記及韻律模式(unsupervised joint Prosody Labeling and Modeling, PLM)演算法，此演算法可用來對語音做韻律標記及韻律模擬，相關的說明如下：[3]利用 PLM 演算法對一女性語者的四種不同語速平行語料庫各別建立階層式韻律模型(Hierocical Prosodic Model, HPM)，利用此演算法模擬韻律中的音節基頻軌跡(syllable F0 contour)、音節長度(syllable duration)、音節能量位階(syllable energy level)、停頓時長(pause duration)，從信號的角度出發，自動對不同語速的語音做韻律標記和韻律結構分析，探討相異語速在各層韻律單元上的韻律參數變化，以及韻律斷點和語言參數之間的關係。

在過去的研究中，發現語速的確對於各個韻律層造成影響。本論文以 PLM 演算法為基礎，提出一語速韻律模型之建立方法，將語速當作一連續變數引入模型，修正原本的 PLM 演算法，使其與語速相依，最後對四個不同語速的平行語料庫建立一個語速相依的 HPM(SR-HPM)。最後結合於 HMM-based 語音合成器來實現一個可控制語速之 TTS 系統。

1.4 語料庫簡介

本論文所使用的實驗語料庫，是由一位專業女性播音員讀稿錄製之快速、正常、中速及慢速平行語料庫，總計 1478 個音檔，共有 203746 個音節，每個語料庫的平均音節長度分別為快語速的 0.181 秒、正常語速的 0.198 秒、中語速的 0.244 秒及慢語速的 0.264 秒，音檔均為 20kHz 的取樣頻率及 16-bit 之 PCM 格式，只有正常速度語料庫為 16kHz 取樣頻率，語料庫的錄製文

字為 Sinica Treebank 語料庫中選出的短篇文章，主要內容大多摘錄自新聞、網路文章，由數個句子所組成的段落，音檔的字數分佈如圖 1.1 所示，平均每個語句(utterance)音節數為 138。所有的音節切割標記和基頻的偵測均先自動由 Hidden Markov Model Tool Kit(HTK)[16]和 WaveSurfer[17]完成，明顯錯誤再以人工方式修正。本論文假設每個語句唸的速度是穩定的，以平均音節長度代表該語句的語速 SR(Speaking Rate)，四個語料庫的語句數目在 SR 之分佈如圖 1.2 所示，整體語料庫 SR 分佈範圍在 0.147-0.297 second/syllable (or 3.4-6.8 syllables/sec)之間，且語料庫間有嚴重的重疊部份。

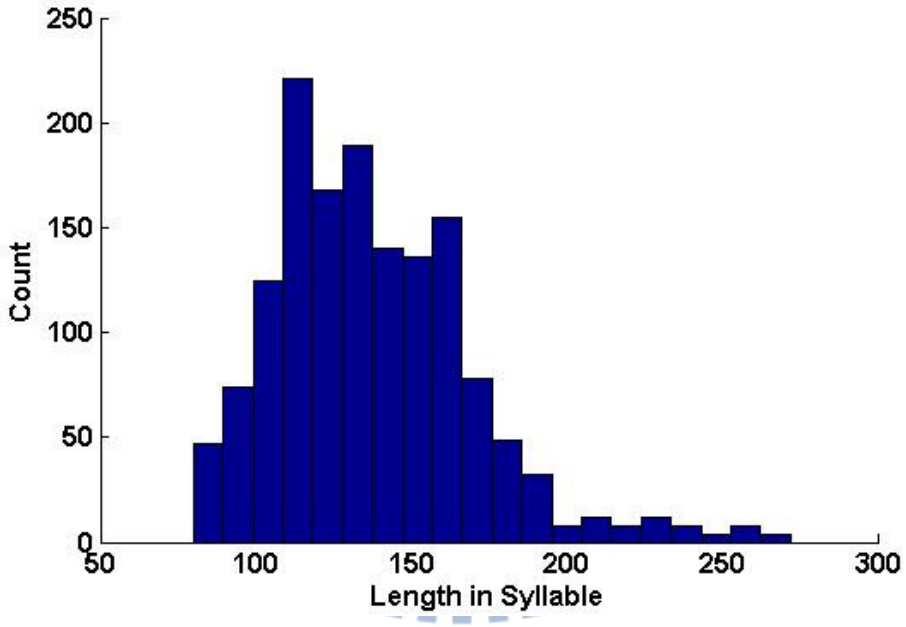


圖 1.1：所有語句的音節數目分佈圖

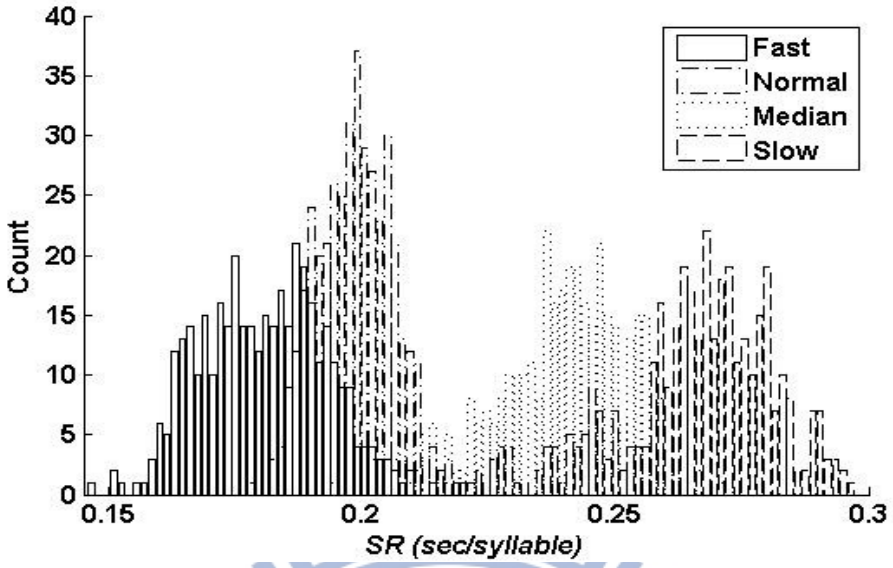


圖 1.2：語句數目在語速之分佈圖

1.5 章節概要說明

本論文一共分為五章，其各章節內容分配如下：

第一章：緒論

第二章：語速相依之階層式韻律模型建立

第三章：語速韻律模型訓練結果與分析

第四章：可控制語速之 TTS 應用

第五章：結論與未來展望

第二章 語速相依之階層式韻律模型建立

本章節以江振宇博士所提出之 HPM 為基礎[15]，引入語速當作新的影響因子，提出新的演算法來建立語速韻律模型。

2.1 漢語語音階層式韻律架構

依據語言學家的研究[18]，語音的韻律結構是呈階層式架構。[19]提出韻律標記的概念並定義了階層式多短語韻律句群(Hierarchical Prosodic Phrase Grouping, HPG)架構，如圖 2.1 所示，最底層為音節層次(SYLlable layer, SYL)，為漢語最基本的字義，其中聲調為最強烈的影響因素，不只影響音節基頻軌跡之走向，也影響了音節長度及能量位階；往上發展依序為韻律詞層次(Prosodic Word layer, PW)，由雙音節或多音節所構成的詞組，通常在句法和語意上關係緊密；韻律短語層次(Prosodic Phrase layer, PPh)，由一或多個韻律詞所組成，結尾常會帶有不明顯但可察覺之停頓；呼吸組層次(Breath Group, BG)，由單一或數個韻律短語組成的句子，其結尾通常帶有明顯停頓；最上層為韻律組句(Prosodic phrase Group, PG)，由一個或數個呼吸組構成。

停頓標記是用來區分韻律組成份子的邊界，*B0* 和 *B1* 區分了 SYL 的邊界，其中 *B0* 表示 reduced syllabic boundary，而 *B1* 表示 normal syllabic boundary，這兩種停頓類別通常都不具明顯停頓；*B2* 和 *B3* 分別是韻律詞和韻律短語的邊界；*B4* 則代表了呼吸組的邊界，和 *B2*、*B3* 比較起來會有較明顯的停頓；至於 *B5* 定義了韻律句組邊界，代表一個完整的段落結束，通常句尾會有音節長度拉長(final lengthening)及能量減弱等現象。

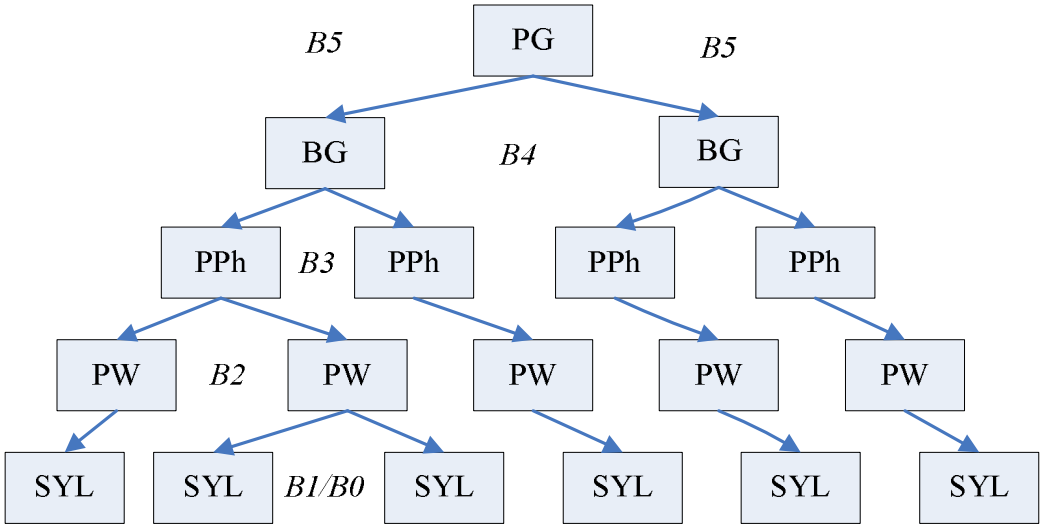


圖 2.1：中文語音韻律階層式架構概念 [19]

本研究使用之語料庫為大段落的語音，因此就以 HPG 架構為基礎，經過進一步的修改後，利用此韻律階層架構來建立本論文所提出之韻律模型。首先將 *B2* 再細分為 *B2-1*、*B2-2*、*B2-3*，分別代表明顯音高重置(pitch reset)、短停頓(short pause)及含有音節拉長效應(duration lengthening)之韻律詞邊界等不同現象。接著將 BG 和 PG 合併為同一層，因為這兩層所描述的韻律特性相近，*B4* 則和 *B5* 合成為 *B4*。整個架構從 5 層變成 4 層，如圖 2.2 所示。最後採用的 7 種韻律邊界停頓(break type)為 **B**={ *B*0, *B*1, *B2-1*, *B2-2*, *B2-3*, *B3*, *B4*}，以此來標記四種韻律單元：音節(SYL)、韻律詞(PW)、韻律短語(PPh)、呼吸組/韻律句組(BG/PG)。

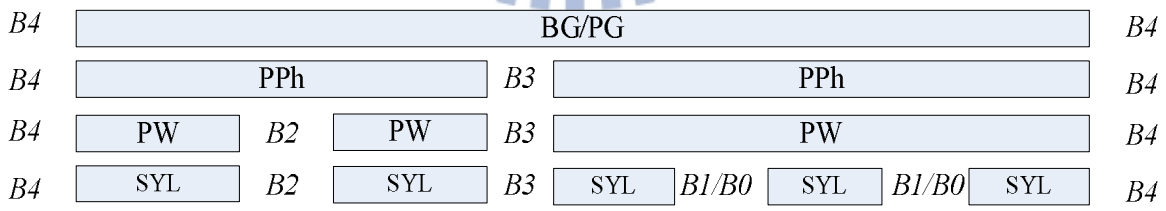


圖 2.2：本研究所採用之階層式韻律架構

2.2 語速韻律模型之建立方法

圖 2.3 為本研究所提出之語速韻律模型建立流程圖。首先，對語句 k 求得平均音節長度 μ_k^{sd} (不包含停頓時長)，以此當作該語句的語速量測 $SR(k)$；接著利用此量測值與語速正規化函

數，對該語句之韻律聲學特徵參數進行語速正規化，目的為補償語速對於韻律聲學特徵參數造成的影響；最後提出一修正型 PLM 演算法來訓練語速韻律模型，同時產生韻律標記。在此修正型 PLM 演算法中，將語速影響加入到 HPM 的兩個子模型，分別為停頓語法模型和韻律狀態模型，目的是補償語速對於韻律架構上層(PW、PPh、BG/PG)所造成的影響。

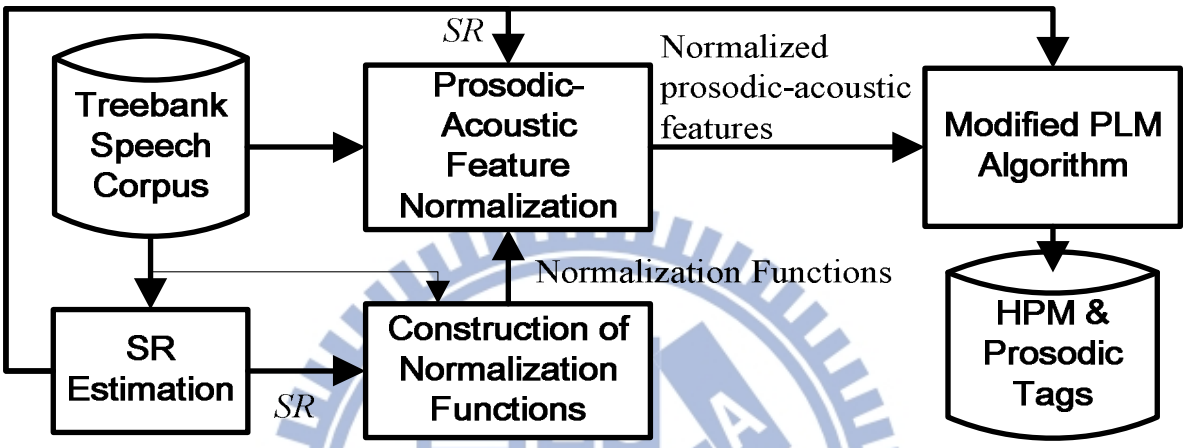


圖 2.3：本研究所提出之語速韻律模型設計流程圖

2.3 韻律聲學特徵參數之語速正規化

此節中，本研究提出語速正規化方法來消除語速對於韻律特徵聲學參數的影響，其中待正規化的聲學參數包括音節長度、停頓時長、音節基頻軌跡及音節能量位階，將在以下四個小節分別介紹其正規化方法。在先前的研究當中[20]，韻律聲學特徵參數被依每個語句去做正規化，以音節長度為例[20]，先對該語句估計音節長度的平均值和標準差，接著做高斯正規化。雖然此種方法簡易有效率，但亦可能造成過度正規化(over-normalization)，例如圖 2.4(b)，有些語句 *SR* 相近但標準差卻相異甚遠，若以語句為單位做正規化，可能將導致部份除語速外的影響因子被壓制，以本論文所採用語料為例，文章組成架構不同會使讀者閱讀方式有所差異。因此，本論文採取較為保守的方法，使用平滑的曲線來模擬每個語句正規化參數(例如音節長度的標準差)與語料庫中 *SR* 影響因素的關係；最後估算出平滑曲線參數來形成語速正規化函數(SR-specific normalization functions)，並用以補償韻律聲學特徵參數中的語速效應。

2.3.1 音節長度之語速正規化

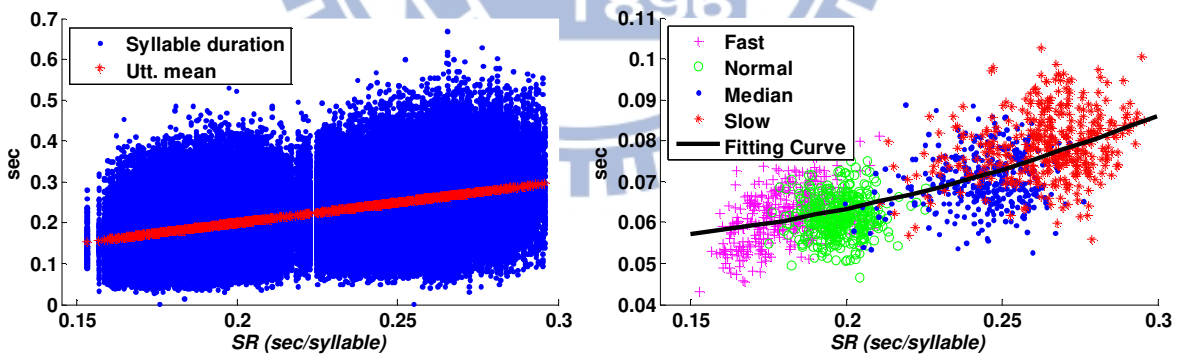
在韻律聲學特徵參數中，音節長度受語速影響最明顯，又漢語音節長度可近似於高斯分佈。因此，我們對語句 k 的音節長度採取高斯正規化法，使用之正規化參數平均值為 $\mu_k^{sd} = SR(k)$，標準差則是以已平滑化標準差取代原始估計的。圖 2.4(a)顯示音節長度對 SR 的分佈圖，(b)為語句標準差對 SR；由圖可發現音節長度之標準差會隨著 SR 增加而增加，故在此使用二階多項式曲線來模擬不同 SR 的音節長度標準差，其音節長度的正規化函數如下所示：

$$sd_n' = \left(sd_n - \mu_k^{sd}\right) \Big/ \tilde{\sigma}^{sd}(SR(k)) \times \sigma_g^{sd} + \mu_g^{sd} \tag{2.1}$$

其中

$$\tilde{\sigma}^{sd}(SR) = a_1 \left(SR\right)^2 + b_1 \cdot SR + c_1 \tag{2.2}$$

為平滑化後的標準差，sd_n 和 sd_n' 分別代表原始音節長度和語速正規化後的音節長度；μ_g^{sd} 和 σ_g^{sd} 為語料庫整體的音節平均值與標準差。



(a) (b)

圖 2.4：(a)音節長度 vs. SR，(b)音節長度之語句標準差 vs. SR

2.3.2 停頓時長之語速正規化

經由觀察停頓時長 pd 的分佈，我們發現伽瑪分佈(Gamma distribution)比高斯更適於模擬

pd 的分佈，它的表示式如下：

$$f(pd;\alpha,\beta)=\beta^{\alpha}\Gamma(\alpha)^{-1}(pd)^{\alpha-1}e^{-\beta\cdot pd},\ \text{for}\ x\geq 0\ \text{ and }\ \alpha,\beta>0 \tag{2.3}$$

由於語句 k 的伽瑪分佈參數 α_k^{pd} 和 β_k^{pd} 可用該語句的平均值 μ_k^{pd} 和標準差 σ_k^{pd} 來表示，因此我們可先求取平滑化的平均值 $\tilde{\mu}^{pd}(SR(k))$ 和標準差 $\tilde{\sigma}^{pd}(SR(k))$，以此形成語速正規化函數。圖 2.5(a)顯示平均值 μ_k^{pd} 對 SR 的分佈圖，(b)則是語句標準差 σ_k^{pd} 對 SR，我們可觀察到兩者皆隨著 SR 而增加。與 2.3.2 小節類似，使用二階多項式曲線來模擬不同 SR 的平均值和標準差，其數學式如下：

$$\tilde{\mu}^{pd}(SR)=a_2\left(SR\right)^2+b_2\cdot SR+c_2 \tag{2.4}$$

$$\tilde{\sigma}^{pd}(SR)=a_3\left(SR\right)^2+b_3\cdot SR+c_3 \tag{2.5}$$

接著以已平滑化 $\tilde{\mu}^{pd}(SR(k))$ 和 $\tilde{\sigma}^{pd}(SR(k))$ 去對 pd 的分佈正規化，其正規化方法如下：

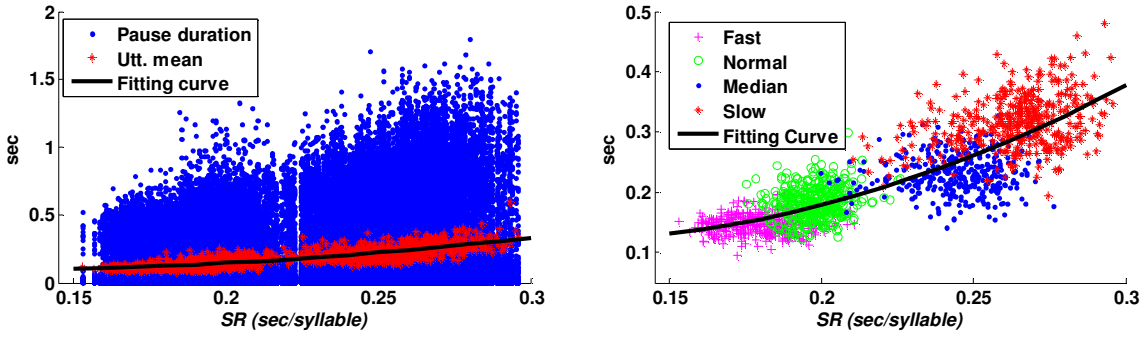
$$pd'=\mathrm{G}^{-1}(\mathrm{G}(pd,\tilde{\alpha}^{pd}(SR(k)),\tilde{\beta}^{pd}(SR(k))),\alpha_g^{pd},\beta_g^{pd}) \tag{2.6}$$

其中 $\mathrm{G}(pd,\alpha,\beta)$ 為伽碼分佈累積密度函數(Cumulative Density Function, CDF)；

$$\tilde{\alpha}^{pd}(SR(k))=(\tilde{\mu}^{pd}(SR(k)))^2/(\tilde{\sigma}^{pd}(SR(k)))^2 \tag{2.7}$$

$$\tilde{\beta}^{pd}(SR(k))=(\tilde{\sigma}^{pd}(SR(k)))^2/\tilde{\mu}^{pd}(SR(k)) \tag{2.8}$$

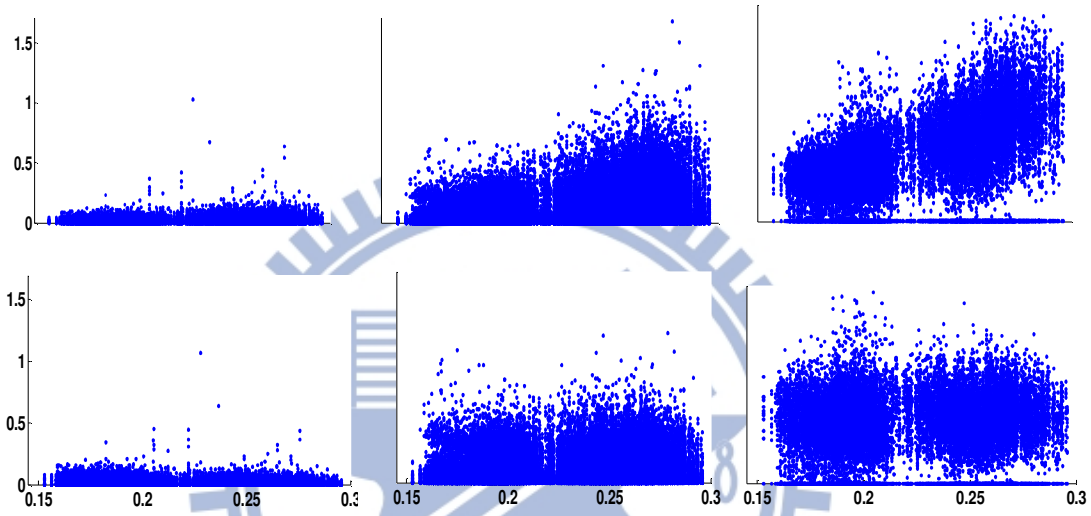
為平滑過的伽碼分佈參數，α_g^{pd} 和 β_g^{pd} 為語料庫整體的平均值和標準差。



(a) (b)

圖 2.5：(a)停頓時長與語句平均值 vs. SR，(b) 停頓時長之語句標準差 vs. SR

圖 2.6 比較了原始 *pd* 和正規後的 *pd*′，在此分三種音節邊界觀察：詞內邊界(intra-word)、非標點符號詞外邊界(non-PM, inter-word)及標點符號詞外邊界(PM, inter-word)。由圖觀察可得知所提出的方法適當地正規化了兩種 inter-word 邊界的 *pd*，而 intra-word 本身受 *SR* 影響不大，故正規化後並無太大差異，此結果符合原本的認知。



(a) (b) (c)

圖 2.6：音節邊界在(a) intra-word, (b) non-PM inter-word, (c) PM inter-word 的 *pd* (上)和 *pd*′ (下) vs. *SR*。(y-axis: pause duration(sec), x-axis: *SR*(sec/syllable))

2.3.3 音節基頻軌跡之語速正規化

本研究將音節基頻軌跡進行正交展開(orthogonal expansion) [21]，投影到四個 Legendre 多項式基底，以所得之四維正交參數表示基頻軌跡，即 $sp_n = \left[a_n^0\ a_n^1\ a_n^2\ a_n^3\right]^T$，四維正交參數分別代表軌跡的平均值、斜率、加速率和彎曲率。由於 sp_n 於漢語五個聲調的分佈差異極大，故依詞彙聲調(lexicon tone)對 sp_n 每一維做語速正規化，其數學式如下：

$$sp_n'(i) = \frac{sp_n(i) - \tilde{\mu}^{sp}(SR(k), t_n, i)}{\tilde{\sigma}^{sp}(SR(k), t_n, i)} \times \sigma_g^{sp}(t_n, i) + \mu_g^{sp}(t_n, i) \tag{2.9}$$

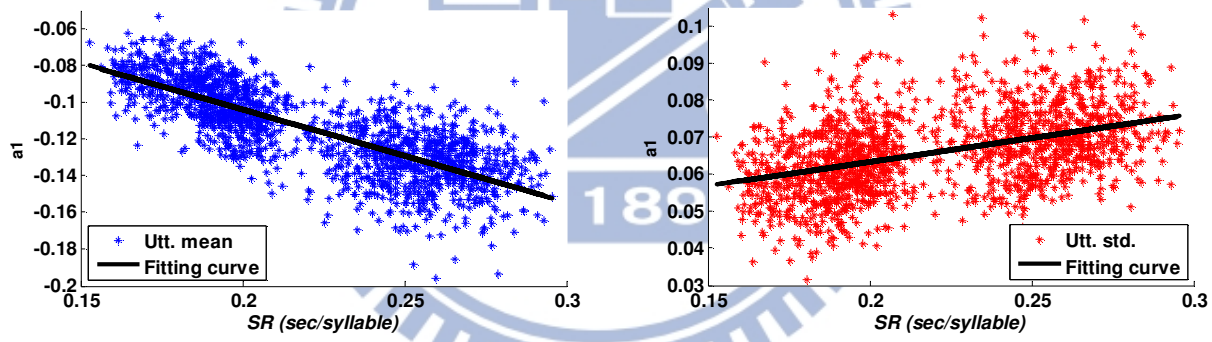
其中

$$\tilde{\mu}^{sp}(SR,t,i) = b_4(t,i)\cdot SR + c_4(t,i) \tag{2.10}$$

$$\tilde{\sigma}^{sp}(SR,t,i) = b_5(t,i)\cdot SR + c_5(t,i) \tag{2.11}$$

分別代表 sp_n 第 i 維、第 t 聲調平均值與標準差之語速正規化函數所表示，正規化函數由一階多項式所形成；$\mu_g^{sp}(t,i)$ 和 $\sigma_g^{sp}(t,i)$ 為整體語料庫的平均值與標準差。

圖 2.7 為一例子，為 $sp_n(2)$ (即軌跡斜率 a_n^1)在第四聲調的語句(a)平均值與(b)標準差對 *SR* 之分佈。由(a)發現 a1 平均值隨著 *SR* 增加而負增加，表示第四聲調在慢語速的基頻軌跡斜率較快語速來得陡峭；由(b)可看出 a1 標準差會隨著 *SR* 一起增加，表示第四聲調的基頻軌跡在慢語速的變動範圍較大。



(a) (b)

圖 2.7： $sp_n(2)$ 於第四聲調之語句(a)平均值 vs. *SR*，(b)標準差 vs. *SR*

2.3.4 音節能量位階之正規化

一般來說，音節能量與錄音條件相關性甚大，例如麥克風與語者距離、麥克風錄音品質、錄音環境等等因素，亦藉由觀察音節能量分佈，確實發現每個語句的能量位階受錄音條件影響，遠大於受語速之影響。因此，本研究的音節能量採語句為基礎的高斯正規化。

2.3 語速韻律模型之設計

本節提出一修正型 PLM 演算法來訓練一個語速相依韻律模型，訓練結果包括四個韻律子模型及兩種韻律標記結果，兩種標記分別為韻律狀態標記及停頓標記，四個韻律子模型主要為描述觀察到的韻律聲學特徵參數、語言參數及韻律階層架構之間的關係。本研究假設語速在韻律聲學特徵參數所造成的影響，已被 2.2 節所提出之正規化方法合理消除。因此，我們可使用修正型 PLM 演算法對這些不同語速的語句訓練一個 HPM，但仍需要使部份 HPM 模型參數與語速相依，以補償語速對於韻律高層次的影響。由於停頓發生的頻率與語速有極大相關性[3]，例如：語者說話速度快時，會產生詞邊界停頓易被遺漏之現象，而說話速度很慢時，停頓則容易被強調，因此本研究對 HPM 中的停頓語法子模型考慮語速影響，詳細將介紹於 2.3.4 小節；另外，在先前的研究[3]亦發現，韻律狀態的轉移受到語速影響，最明顯的例子為長停頓發生時，語速越快則韻律狀態的轉移範圍越大，反之語速越慢其轉移範圍越小，故同樣地對韻律狀態子模型進行修正，使其考慮語速影響因子，詳細將介紹於 2.3.3 小節。

修正型 PLM 演算法可視為一個韻律標記過程，並同時更新模型參數。在給定語料庫之韻律聲學特徵參數集合 **A**、相對應的語言參數集合 **L** 及語速 **SR** 之下，找出一組最佳韻律標記集合 **T**，整個過程可以看成一參數最佳化問題，即

$$\mathbf{T}^{*} = \arg\max_{\mathbf{T}} P(\mathbf{T} \mid \mathbf{A}, \mathbf{L}, \mathbf{SR}) = \arg\max_{\mathbf{T}} P(\mathbf{T}, \mathbf{A} \mid \mathbf{L}, \mathbf{SR}) \tag{2.12}$$

韻律標記集合 **T={B, PS}**包含兩種重要的韻律訊息，第一種為音節邊界的停頓標記(Break Type)，用來表示階層式架構的韻律組成份子邊界，本論文定義韻律邊界停頓標記集合為 **B**={*B*0, *B*1, *B*2-1, *B*2-2, *B*2-3, *B*3, *B*4}；第二種韻律標記為音節韻律狀態分為 **PS**={p, q, r}，其所代表意義分別為經過正規化和量化的音節基韻律態 p、音節長度韻律狀態 q 及音節能量韻律狀態 r，正規化後的韻律狀態扣除了音節層次的貢獻，以基頻韻律狀態來說，扣除聲調和連音的影響因素，音長或能量韻律狀態則扣除聲調、基本音節類型或韻母類型等影響因素。

本論文韻律聲學參數 **A={X, Y, Z}**分為兩類，第一類為音節本身的聲學參數 **X={sp, sd, se}**，分別為音節基頻軌跡 **sp**、音節長度 **sd** 及音節能量位階 **se**，本研究假設此類聲學參數與韻

律狀態標記有很大相關性，與音節邊界停頓標記相關性非常小，本論文稱 **X** 為音節韻律參數(syllable prosodic feature)；第二類為音節邊界的聲學參數**{Y, Z}={pd, ed, pj, dl, df}**，分別為音節邊界的停頓時長(pause duration, *pd*)、能量低點位階(energy-dip level, *ed*)、正規化基頻差(normalized pitch jump, *pj*)及兩種正規化長度拉長因子(normalized duration lengthening factor, *dl* and *df*)等，假設此類型的聲學參數與停頓標記有很大相關性，與韻律狀態標記的相關性很小，本論文稱 **Y={pd, ed}**為音節內韻律參數(inter-syllabic prosodic feature)、**Z={pj, dl, df}**為差分韻律參數(differential prosodic feature)；最後 **SR** 為本論文所定義的語速測量值，即語句的平均音節長度。

在語言參數方面，用 **L** 表示所有的語言參數集合。其中特別將音節聲調、基本音節類型與韻母類型從 **L** 中獨立出來，用意在於這三個語言參數對音節基頻軌跡、音節長度及音節能量位階有顯著的影響，把剩餘的語言參數統一定義為 **l**(reduced linguistic feature set)。完整的符號定義整理於表 2.1。

表 2.1：韻律標記、聲學參數及語言參數之表示符號

T : prosodic tag	**B** : break type={*B0*, *B1* ,*B2-1*, *B2-2*, *B2-3*, *B3*, *B4*}	
	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
SR : speaking rate		
L : linguistic feature	**l**: reduced linguistic feature set	
	t: syllable tone sequence **s**: base-syllable type sequence **f**: final type sequence	

綜合上述之討論，可將 $P(\mathbf{T},\mathbf{A}|\mathbf{L},\mathbf{SR})$ 改寫成以下形式：

$$\begin{aligned} P(\mathbf{T},\mathbf{A}|\mathbf{L},\mathbf{SR}) &= P(\mathbf{A}|\mathbf{T},\mathbf{L})P(\mathbf{T}|\mathbf{L},\mathbf{SR}) = P(\mathbf{X},\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{PS},\mathbf{L})P(\mathbf{B},\mathbf{PS}|\mathbf{L},\mathbf{SR}) \\ &\approx P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})P(\mathbf{PS}|\mathbf{B},\mathbf{SR})P(\mathbf{B}|\mathbf{L},\mathbf{SR}) \end{aligned} \tag{2.13}$$

其中 $P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})$ 稱為音節韻律模型，用來敘述音節韻律參數受到停頓標記 **B**、韻律狀態 **PS** 和語言參數 **L** 之間的影響而產生的變化； $P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$ 稱為停頓聲學模型，用以敘述在各個不同停頓標記 **B** 和語言參數 **L** 下，其韻律邊界的聲學特性；$P(\mathbf{PS}|\mathbf{B},\mathbf{SR})$ 稱為修正型韻律狀態模型，描述了韻律狀態在不同停頓標記 **B** 及不同語速 **SR** 下的轉移變化； $P(\mathbf{B}|\mathbf{L},\mathbf{SR})$ 稱為修正型停頓語法模型，描述在不同的語言參數 **L** 及不同語速下，各種停頓標記出現的頻率。以下將分四小節針對這四種韻律模型做更深入的探討。

2.3.1 音節韻律模型

音節韻律模型 $P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})$ 可進一步分解成三個獨立子模型，分別用來模擬音節基頻軌跡、音節長度及音節能量位階，其數學式如下：

$$\begin{aligned} p(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L}) &\approx p(\mathbf{sp}|\mathbf{B},\mathbf{p},\mathbf{t})\,p(\mathbf{sd}|\mathbf{q},\mathbf{t},\mathbf{s},\mathbf{u})\,p(\mathbf{se}|\mathbf{r},\mathbf{t},\mathbf{f},\mathbf{u}) \\ &\approx \prod_{n=1}^{N} p(\mathbf{sp}_n|B_{n-1}^{n},p_n,t_{n-1}^{n+1}) \prod_{n=1}^{N} p(sd_n|q_n,t_n,s_n,u_n) \prod_{n=1}^{N} p(se_n|r_n,t_n,f_n,u_n) \end{aligned} \tag{2.14}$$

$p(\mathbf{sp}_n|B_{n-1}^{n},p_n,t_{n-1}^{n+1})$ 用以模擬第 n 個音節基頻軌跡 $\mathbf{sp}_n$，在此假設所觀察到的 $\mathbf{sp}_n$ 受到的影響因素(Affecting Pattern, AP)為：目前的聲調 t_n、目前的基頻韻律狀態 p_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 所造成的連音影響，此處 $B_{n-1}^{n}=(B_{n-1},B_n)$，$t_{n-1}^{n+1}=(t_{n-1},t_n,t_{n+1})$。而 $\mathbf{sp}_n$ 是將音節基頻軌跡進行正交展開(orthogonal expansion)，投影到四個 Legendre 多項式基底所得到的四維正交參數[21]，依以上描述可將 $\mathbf{sp}_n$ 表示成

$$\mathbf{sp}_n = \mathbf{sp}_n^{r} + \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\beta}^{f}_{B_{n-1},t_{n-1}^{n}} + \boldsymbol{\beta}^{b}_{B_n,t_n^{n+1}} + \boldsymbol{\mu}_{sp} \tag{2.15}$$

在(2.15)式中，tp_n 是 tone pair $t_n^{n+1}=(t_n,t_{n+1})$，$\boldsymbol{\beta}_{t_n}$ 及 $\boldsymbol{\beta}_{p_n}$ 則分別為目前音節音調 t_n 及目前音節韻律

狀態 p_n 的 APs，其中韻律狀態的影響只限制對目前音節的 LogF0 level，故將 $\boldsymbol{\beta}_{p_n}$ 的四維正交係數，僅第一維設為非零值；$\boldsymbol{\beta}^{f}_{B_{n-1},t^{n}_{n-1}}$ 及 $\boldsymbol{\beta}^{b}_{B_n,t^{n+1}_{n}}$ 分別是第 n-1 個和第 n+1 個音節所貢獻的前後連音效應 APs；$\boldsymbol{\mu}_{sp}$ 為總體平均值(global mean)，僅第一維為非零值；$\mathbf{sp}^{r}_{n}$ 為正規化後的 $\mathbf{sp}_n$ ，即 $\mathbf{sp}_n$ 扣除 $\boldsymbol{\beta}_{t_n}$ 、$\boldsymbol{\beta}_{p_n}$ 、$\boldsymbol{\beta}^{f}_{B_{n-1},t^{n}_{n-1}}$ 、$\boldsymbol{\beta}^{b}_{B_n,t^{n+1}_{n}}$ 和 $\boldsymbol{\mu}_{sp}$ 的殘餘值(residual)，圖 2.8 顯示出 $\mathbf{sp}_n$ 與這些影響因子之間的關係，在此假設 $\mathbf{sp}^{r}_{n}$ 為一平均值為零的高斯分佈隨機變數，即 $N(\mathbf{sp}^{r}_{n};\mathbf{0},\mathbf{R}_{sp})$ ，因此得到

$$p(\mathbf{sp}_n \mid p_n, B^{n}_{n-1}, t^{t+1}_{t-1}) = N(\mathbf{sp}_n;\boldsymbol{\beta}_{t_n}+\boldsymbol{\beta}_{p_n}+\boldsymbol{\beta}^{f}_{B_{n-1},t^{n}_{n-1}}+\boldsymbol{\beta}^{b}_{B_n,t^{n+1}_{n}}+\boldsymbol{\mu}_{sp},\mathbf{R}_{sp}) \tag{2.16}$$

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

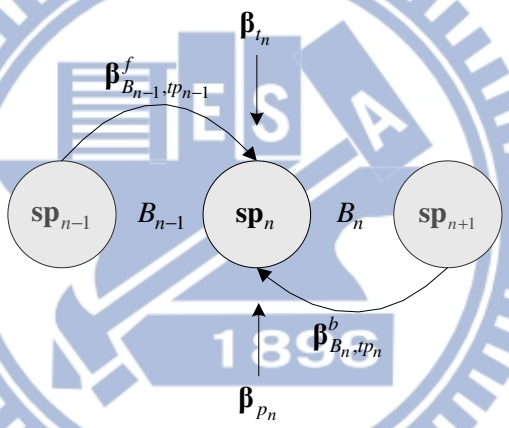


圖 2.8：音節基頻軌跡與其影響因素關係圖

依此類推，第二個模型和第三個模型可表示成：

$$p(sd_n \mid q_n, s_n, t_n) = N(sd_n;\gamma_{t_n}+\gamma_{s_n}+\gamma_{q_n}+\mu_{sd},R_{sd}) \tag{2.17}$$

$$p(se_n \mid r_n, f_n, t_n) = N(se_n;\omega_{t_n}+\omega_{f_n}+\omega_{r_n}+\mu_{se},R_{se}) \tag{2.18}$$

(2.17)式模擬了音節時長 sd_n ，其中 γ_{t_n} 、γ_{s_n} 和 γ_{q_n} 分別為聲調、基本音節類型和韻律狀態對 sd_n 的 APs，μ_{sd} 和 R_{sd} 分別為 sd_n 總體平均及其殘餘值之變異數；(2.18)式模擬了音節能量位階 se_n，其中 ω_{t_n} 、ω_{s_n} 和 ω_{q_n} 分別為聲調、聲母類型和韻律狀態對 se_n 的 APs，μ_{se} 和 R_{se} 則分別為 se_n 總體平均及其殘餘值之變異數。

2.3.2 停頓聲學模型

將停頓聲學模型 $P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$ 做進一步分解

$$P(\mathbf{Y},\mathbf{Z}\,|\,\mathbf{B},\mathbf{L}) \approx P(\mathbf{Y},\mathbf{Z}\,|\,\mathbf{B},\mathbf{l}) \approx \prod_{n=1}^{N} p(pd_n, ed_n, pj_n, dl_n, df_n \,|\, B_n, \mathbf{l}_n) \tag{2.19}$$

在此使用音節內及差分韻律參數**{Y, Z}={pd, ed, pj, dl, df}**描述韻律邊界的聲學特性，pd_n 為第 n 個音節跟隨的接合點(juncture n，之後以第 n 個接合點表示)停頓長度；ed_n 為第 n 個接合點的能量下降程度；pj_n 為跨越第 n 個接合點的正規化基頻差，其定義如下：

$$pj_n = (sp_{n+1}(1) - \beta_{t_{n+1}}(1)) - (sp_n(1) - \beta_{t_n}(1)) \tag{2.20}$$

而兩種正規化長度拉長因子 dl 和 df 定義如下：

$$dl_n = (sd_n - \gamma_{t_n} - \gamma_{s_n}) - (sd_{n-1} - \gamma_{t_{n-1}} - \gamma_{s_{n-1}}) \tag{2.21}$$

$$df_n = (sd_n - \gamma_{t_n} - \gamma_{s_n}) - (sd_{n+1} - \gamma_{t_{n+1}} - \gamma_{s_{n+1}}) \tag{2.22}$$

由於對韻律停頓而言 $\mathbf{l}_n$ 的空間仍太大，故本研究藉由分類樹與決策樹(Classification and Regreesion Tree, CART)演算法來估計 $p(pd_n, ed_n, pj_n, dl_n, df_n \,|\, B_n, \mathbf{l}_n)$，其節點分類標準依據最大概似函數增益(maximum likelihood gain)搭配一個事先設計好的問題集去實施 CART 演算法，依據不同的韻律邊界停頓將所有音節邊界的 pd_n、ed_n、pj_n、dl_n、df_n 做好分類，並於決策樹的每個終止節點(leaf node)統計參數分佈。在此我們將 pd_n 以伽瑪分佈(Gamma distribution)來模擬，而 ed_n、pj_n、dl_n、df_n 以高斯分佈模擬，假設五種聲學間彼此互相獨立，因此 $p(pd_n, ed_n, pj_n, dl_n, df_n \,|\, B_n, \mathbf{l}_n)$ 會是一個伽瑪分佈和四個高斯分佈的乘積，其數學式如下：

$$\begin{aligned} &\prod_{n=1}^{N} p(pd_n, ed_n, pj_n, dl_n, df_n \,|\, B_n, \mathbf{l}_n) \\ &\approx \prod_{n=1}^{N} \Big\{ g(pd_n; \alpha_{B_n,L_n}, \beta_{B_n,L_n}) N(ed_n; \mu_{ed,B_n,L_n}, \sigma^2_{ed,B_n,L_n}) \\ &\qquad N(pj_n; \mu_{pj,B_n,L_n}, \sigma^2_{pj,B_n,L_n}) N(dl_n; \mu_{dl,B_n,L_n}, \sigma^2_{dl,B_n,L_n}) \\ &\qquad N(df_n; \mu_{df,B_n,L_n}, \sigma^2_{df,B_n,L_n}) \Big\} \end{aligned} \tag{2.23}$$

2.3.3 修正型韻律狀態模型

韻律狀態模型 $P(\mathbf{PS}|\mathbf{B},\mathbf{SR})$ 進一步以三個子模型近似之，如下式所示：

$$P(\mathbf{PS}\,|\,\mathbf{B},\mathbf{SR}) \approx P(\mathbf{p}\,|\,\mathbf{B},\mathbf{SR})P(\mathbf{q}\,|\,\mathbf{B},\mathbf{SR})P(\mathbf{r}\,|\,\mathbf{B},\mathbf{SR}) \tag{2.24}$$

分別用來模擬音節基頻、長度及能量三種韻律狀態。本研究假設目前的韻律狀態僅與前一韻律狀態及前一停頓標記有關，使用一階馬可夫模型(1 order Markov Model)實現之，並以 bin 來區分不同語速所造成的影響，$P(\mathbf{p}\,|\,\mathbf{B},\mathbf{SR})$ 、 $P(\mathbf{q}\,|\,\mathbf{B},\mathbf{SR})$ 、 $P(\mathbf{r}\,|\,\mathbf{B},\mathbf{SR})$ 最後被表示如下：

$$P(\mathbf{p}\,|\,\mathbf{B},\mathbf{SR}) \approx p(p_1|\text{bin}(SR_1))\left[\prod_{n=2}^{N} p(p_n\,|\,p_{n-1}, B_{n-1}, \text{bin}(SR_n))\right], \tag{2.25}$$

$$P(\mathbf{q}\,|\,\mathbf{B},\mathbf{SR}) \approx p(q_1|\text{bin}(SR_1))\left[\prod_{n=2}^{N} p(q_n\,|\,q_{n-1}, B_{n-1}, \text{bin}(SR_n))\right], \tag{2.26}$$

和

$$P(\mathbf{r}\,|\,\mathbf{B},\mathbf{SR}) \approx p(r_1|\text{bin}(SR_1))\left[\prod_{n=2}^{N} p(r_n\,|\,r_{n-1}, B_{n-1}, \text{bin}(SR_n))\right] \tag{2.27}$$

其中 $p(p_1|\text{bin}(SR_1))$ 、 $p(q_1|\text{bin}(SR_1))$ 及 $p(r_1|\text{bin}(SR_1))$ 表示三種韻律狀態的初始機率(initial probability)； $p(p_n\,|\,p_{n-1}, B_{n-1}, \text{bin}(SR_n))$ 、 $p(q_n\,|\,q_{n-1}, B_{n-1}, \text{bin}(SR_n))$ 及 $p(r_n\,|\,r_{n-1}, B_{n-1}, \text{bin}(SR_n))$ 表示三種韻律狀態，在給定停頓標記 B_{n-1} 及語速 SR_n 之情況下，從第 n-1 個音節的韻律狀態到第 n 個音節韻律狀態之轉移機率(transition probability)；bin(.)為索引函數(index function)。

2.3.4 修正型停頓語法模型

首先，修正型停頓語法 $P(\mathbf{B}|\mathbf{L},\mathbf{SR})$ 模型可先簡化為 $P(\mathbf{B}|\mathbf{l},\mathbf{SR})$ ，並假設每個音節邊界可分開模擬，因此可表示成

$$P(\mathbf{B}\,|\,\mathbf{l},\mathbf{SR}) = \prod_{n=1}^{N-1} p(B_n\,|\,\mathbf{l}_n, SR_n) \tag{2.28}$$

其中 $p(B_n\,|\,\mathbf{l}_n, SR_n)$ 由兩個步驟建構而成，第一步：使用 CART 演算法去對標記結果 B_n 訓練一顆決策樹，並對決策樹所有終止節點估計 $p(B_n\,|\,\mathbf{l}_n)$ ；第二步：對步驟一所建構的決策樹所有終止

節點之七種停頓標記，使用一階多項式曲線來模擬停頓標記出現頻率對 *SR* 的關係，其數學式如下：

$$P(B_n = m \mid \mathbf{l}_n, SR_n) = \frac{P(B_n = m \mid \mathbf{l}_n, SR_n)}{\sum_{x \in \text{all break types}} P(B_n = x \mid \mathbf{l}_n, SR_n)} \approx \frac{c_{m,j} SR_n + d_{m,j}}{\sum_{x \in \text{all break type}} c_{x,j} SR_n + d_{x,j}} \tag{2.29}$$

其中 j 為語言參數向量 $\mathbf{l}_n$ 所對應到決策樹的終止節點索引值；$c_{m,j}$ 及 $d_{m,j}$ 為停頓標記 m、終止節點 j 的線性迴歸係數。

2.4 修正型 PLM 演算法之訓練過程

此章節將介紹如何使用修正型 PLM 演算法來訓練 2.3 節所提出之韻律模型，PLM 演算法是基於最大概似度法則(Maximum Likelihood, ML)，對所有語句找出最佳的韻律標記，並估計模型參數。首先，我們依 2.3 節所設計之 8 個模型定義一目標函數(objective function)如下：

$$Q = \left(\prod_{n=1}^{N} p(\mathbf{sp}_n \mid p_n, B_{n-1}^{n}, t_{n-1}^{n+1}) p(sd_n \mid q_n, t_n, s_n) p(se_n \mid r_n, t_n, f_n) \right) \left(p(p_1 \mid \text{bin}(SR_1)) p(q_1 \mid \text{bin}(SR_1)) p(r_1 \mid \text{bin}(SR_1)) \prod_{n=2}^{N} p(p_n \mid p_{n-1}, B_{n-1}, \text{bin}(SR_n)) p(q_n \mid q_{n-1}, B_{n-1}, \text{bin}(SR_n)) p(r_n \mid r_{n-1}, B_{n-1}, \text{bin}(SR_n)) \right) \left(\prod_{n=1}^{N-1} \left(p(pd_n, ed_n, pj_n, dl_n, df_n \mid B_n, \mathbf{l}_n) p(B_n \mid \mathbf{l}_n, SR_n) \right) \right) \tag{3.21}$$

接著再採取一連串的最佳化程序，逐項估計各個子模型的模型參數及標記韻律標記，重複執行此程序直到收斂。整個演算法的實現分成兩個部份：初始化及疊代訓練，我們將在 2.4.1 和 2.4.2 分別介紹。

2.4.1 初始化

初始化過程分兩部份：(a)標記所有音節邊界的初始停頓標記，(b)使用 ML 法則估計 8 個子模型的初始模型參數及標記每個音節的初始韻律狀態。

(a) 標記所有音節邊界的初始停頓標記

初始停頓標記方法是採取[22]所提出之決策樹來實現，如圖 2.9 所示。建構決策樹主要是由音節邊界之聲學特性配合一般所認知的語言特性所設計出來的，其中 Tr1~Tr8 是由韻律邊界的聲學特性所決定，詳細決定 Tr1~Tr8 的演算法如附錄一所示。我們期望以此決策樹來使初始停頓標記符合以下之定義：首先，對於大多的 PM 音節邊界會有長的停頓時長，容易被標為主要停頓(major break)，對應於階層式韻律架構 *B3* 及 *B4*；其次，在 non-PM, inter-word 的音節邊節中，具有中等停頓時長定義為 *B2-2*、具有中等基頻跳躍定義為 *B2-1*、具有中等音節時長拉長為 *B2-3*，這些停頓類別皆屬次要停頓(minor break)；最後，大多數 intra-word 的邊界停頓時長都很短，屬非停頓(non-break)，對應到韻律架構中的 *B0* 及 *B1*，其中 *B0* 屬於緊密連接(tightly coupling)的韻律邊界，相對於 *B1* 有較小的基頻停頓(pitch pause)和較大的能量低點位階(energy-dip level)。

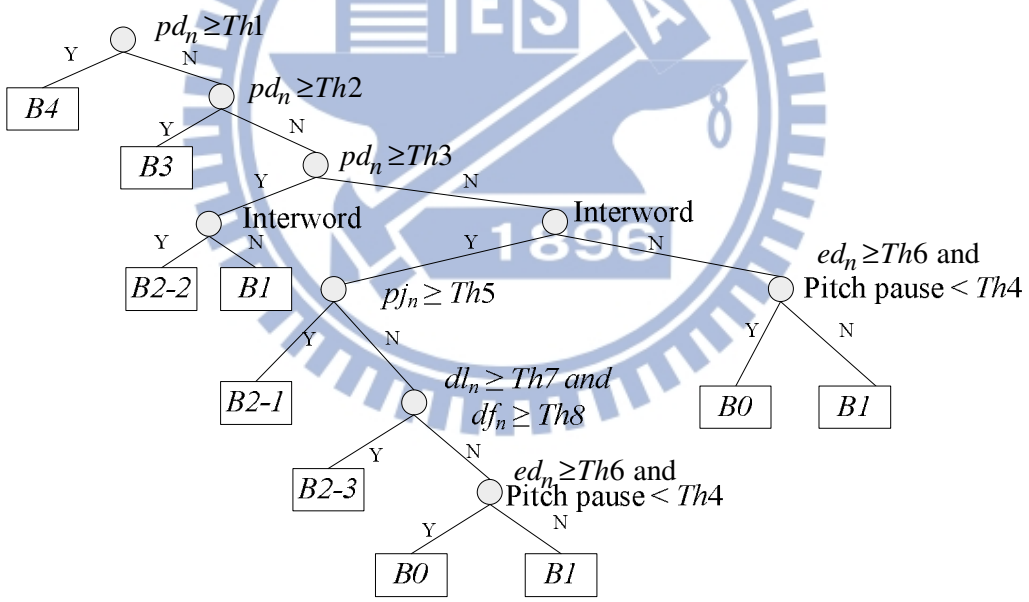


圖 2.9：初始化停頓標記決策樹

(b) 估計 8 個子模型的初始模型參數及標記每個音節的初始韻律狀態

有了初始化停頓標記後，我們使用 CART 演算法來建構停頓聲學模型及停頓語法模型，再使用(2.29)式修正停頓語法模型，其中 Θ 為 CART 所使用之問題集，詳細內容如附錄二所示。

至於音節韻律模型則是用一個漸進式的估測程序，首先估測總體平均值(global mean)的 APs{$\boldsymbol{\mu}, \mu_d, \mu_e$}，接下來依序估測聲調 APs{$\boldsymbol{\beta}_t, \gamma_t, \alpha_t$}、基本音節類型與韻母類型 APs{$\gamma_s, \alpha_f$}、連音效應 APs{$\boldsymbol{\beta}^f_{B,tp}, \boldsymbol{\beta}^b_{B,tp}$}和韻律狀態 APs{$\boldsymbol{\beta}_p, \gamma_q, \alpha_r$}。其中初始的韻律狀態則是將音節基頻軌跡、長度及能量位階各別扣除其它 APs 的殘餘值做向量量化(vector quantization, VQ)，將量化之後的碼字(codeword)當作初始韻律狀態。最後，修正型韻律狀態模型 $P(\mathbf{p}|\mathbf{B},\mathbf{SR})$、$P(\mathbf{q}|\mathbf{B},\mathbf{SR})$ 和 $P(\mathbf{r}|\mathbf{B},\mathbf{SR})$ 則是利用已初始化停頓標記及韻律狀態估計而成。

2.4.2 疊代訓練

經初始化後，我們使用一疊代過程來訓練模型，其步驟如下：

步驟 1：固定其它 APs，更新聲調的 APs{$\boldsymbol{\beta}_t, \gamma_t, \alpha_t$}。

步驟 2：固定其它 APs，更新連音效應的 APs{$\boldsymbol{\beta}^f_{B,tp}, \boldsymbol{\beta}^b_{B,tp}$}，接著更新共變數矩陣 $\mathbf{R}$。

步驟 3：固定其它 APs，更新基本音節類型及韻母類型的 APs{γ_s, α_f}，接著更新變異數 R_d 和 R_e。

步驟 4：利用維特比(Viterbi)演算法重新標記所有語句之韻律狀態序列，使得目標函數 Q 達到最大值，然後更新韻律狀態的 ARs{$\boldsymbol{\beta}_p, \gamma_q, \alpha_r$}，最後更新修正型韻律狀態模型 $P(\mathbf{p}|\mathbf{B},\mathbf{SR})$、$P(\mathbf{q}|\mathbf{B},\mathbf{SR})$ 和 $P(\mathbf{r}|\mathbf{B},\mathbf{SR})$ 以及共變數矩陣 $\mathbf{R}$、變異數 R_d 和 R_e。

步驟 5：利用維特比(Viterbi)演算法重新標記所有語句之停頓標記序列，使得目標函數 Q 達到最大值，接著更新修正型韻律狀態模型 $P(\mathbf{p}|\mathbf{B},\mathbf{SR})$、$P(\mathbf{q}|\mathbf{B},\mathbf{SR})$ 與 $P(\mathbf{r}|\mathbf{B},\mathbf{SR})$ 以及共變數矩陣 $\mathbf{R}$、變異數 R_d 和 R_e。

步驟 6：利用 CART 演算法和 Θ 重新建構決策樹，分別更新停頓聲學模型 $p(pd_n, ed_n, pj_n, dl_n, df_n | B_n, \mathbf{l}_n)$ 及修正型停頓語法模型 $P(B_n|\mathbf{l}_n, SR_n)$。

步驟 7：重複步驟 1 到 7 的過程直到收斂為止。

第三章 語速韻律模型訓練結果與分析

本實驗所使用訓練語料為 SR-Treebank 語料庫，共 1478 句，總音節數為 203746 個，快、正常、中等和慢語速四個語料庫韻律聲學參數的統計資料列於表 3.1。由表可觀察到音節 Log-F0 軌跡變異數向量的第二、三、四維都隨著語速變慢而上升，表示在語速慢時，基頻軌跡的變化幅度較為劇烈，而語速快時，常發出不完整基頻軌跡，動態變化被侷限在較小範圍；音節長度與語速相關性最大，其平均值或變異數皆隨著語速變慢而增加；至於音節能量位階，由於正常語速語料庫的錄音條件與其它三者不同，故統計結果有所差異。

表 3.1：SR-Treeabk 韻律聲學參數之統計資訊

	音節 Log-F0 平均值	音節 Log-F0 軌跡共變數矩陣	音節長度平均值	音節長度變異數	音節能量位階平均值	音節能量位階變異數
Fast	5.28	$[422, 51, 11, 3]\times10^{-4}$	0.18	400×10^{-5}	52.5	22.3
Normal	5.31	$[546, 90, 17, 5]\times10^{-4}$	0.20	380×10^{-5}	60.0	48.6
Median	5.25	$[407, 89, 17, 5]\times10^{-4}$	0.24	510×10^{-5}	53.1	22.7
Slow	5.25	$[433, 94, 18, 4]\times10^{-4}$	0.26	650×10^{-5}	53.0	23.5

將以上韻律聲學參數和停頓時長經過語速正規化後，採取修正型 PLM 演算法疊代訓練至 71 次達到收斂，其對應的目標總概似度(total likelihood of objective function)如圖 3.1 所示。接下來的章節將對模型訓練結果及韻律標記結果進行分析。

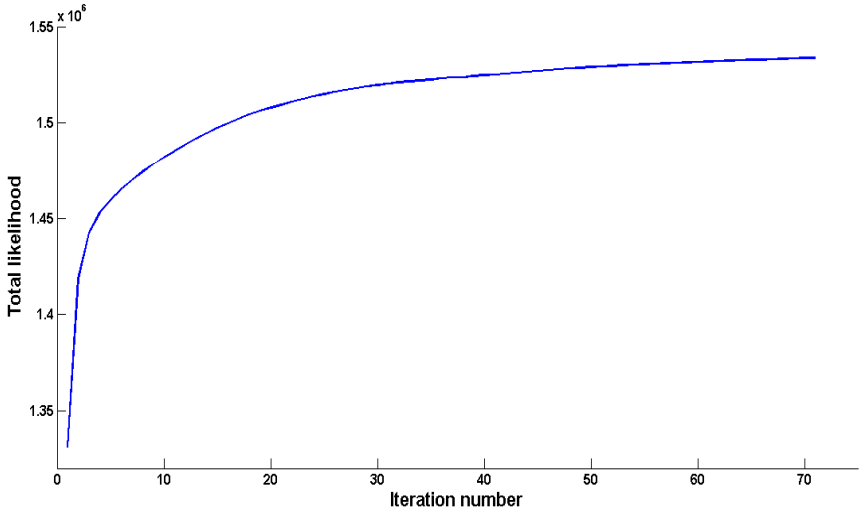


圖 3.1：疊代次數與目標總概似度

3.1 韻律模型參數之分析

此節針對四個韻律子模型的訓練結果進行探討與分析，並與語料庫為基礎之 HPM[3] (Corpus-based HPM)做比較。

3.1.1 音節韻律模型

音節韻律模型可分成三個子模型，分別用以模擬音節基頻軌跡、音節時長及音節能量位階，本節將探討各種 APs 對於音節韻律所造成的影響，以及模型參數與語速間的關係變化。

首先，由音節基頻軌跡韻律模型開始，影響因子包含聲調、連音效應和韻律狀態。圖 3.2 顯示基頻軌跡的聲調 APs，此結果與過去研究[15]所得之基頻軌跡相符合。

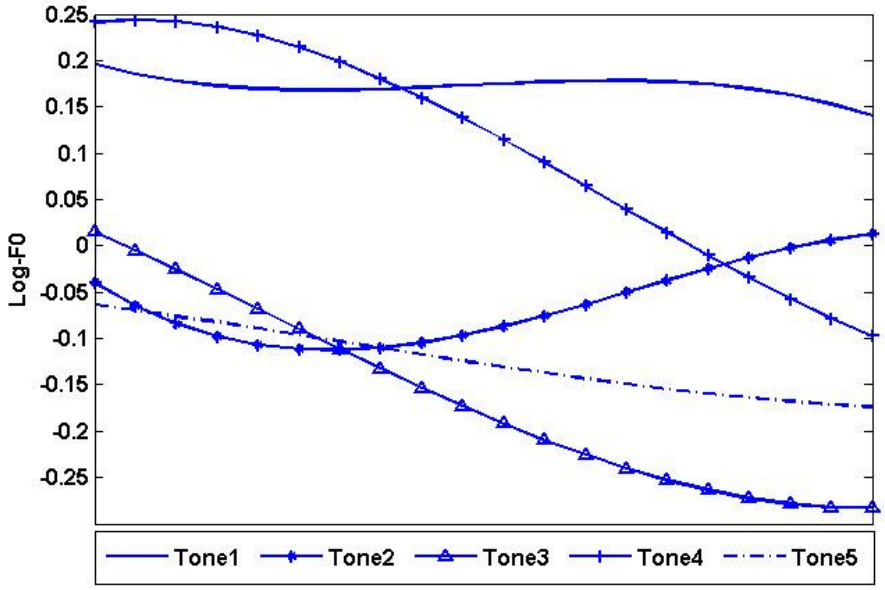


圖 3.2：基頻軌跡聲調 APs

圖 3.3 顯示基頻軌跡在停頓標記 *B0*、*B1* 和 *B4* 時的前音節連音效應 APs，橫軸 *i* 表示目前的聲調，縱軸 *j* 表示前一音節之聲調。在此選擇 B0、B1 和 B4 為連音效應最極端的例子，由圖可清礎發現 *B0* 的連音效應最嚴重、*B1* 次之，*B4* 影響最小，對於聲調組合為(1, 2)、(1, 3)、(2, 2)、(2, 3)、(1, 5)等有 hight-low mismatch 現象，$\boldsymbol{\beta}^{f}_{B,tp}$ 會產生向下彎曲的基頻軌跡來補償其連音效應；另外聲調組合為(3, 1)、(3, 4)、(5, 1)、(5, 4)、(4, 1)、(4, 4)等有 low-hight mismatch 現象，$\boldsymbol{\beta}^{f}_{B,tp}$ 則會產生向上彎曲的基頻軌跡來補償。

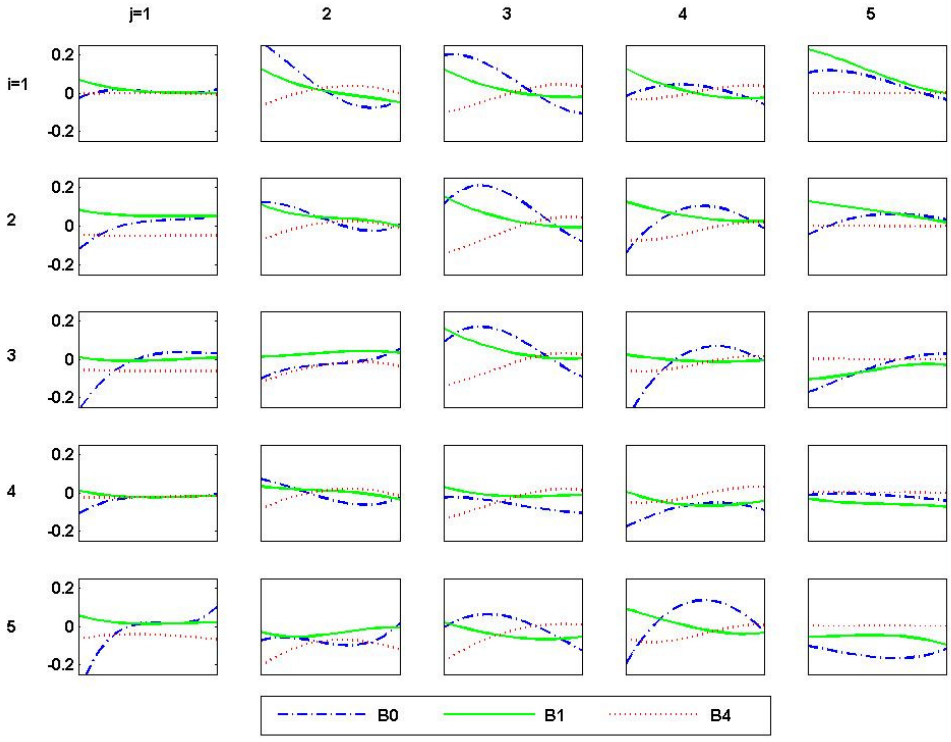


圖 3.3：基頻軌跡在停頓標記 *B0*、*B1* 和 *B4* 時前音節連音效應 APs，在此 $tp = (i, j)$

圖 3.4 顯示基頻軌跡在停頓標記 *B0*、*B1* 和 *B4* 時的前音節連音效應 APs，橫軸 i 表示目前的聲調，縱軸 j 表示下一音節聲調。由圖觀察到，比較於前音節連音效應 $\boldsymbol{\beta}^{f}_{B,tp}$，後音節連音效應 $\boldsymbol{\beta}^{b}_{B,tp}$ 變化範圍明顯小了很多，表示後音節連音效應的影響程度不如前音節，此結果與先前研究[24]符合。其中，特別注意到聲調組合為(3, 3)時，$\boldsymbol{\beta}^{b}_{B,tp}$ 產生劇烈上揚的曲線，這是因為此聲調組合的第一個三聲會被發音為二聲，即變調規則(tone sandhi rule)。

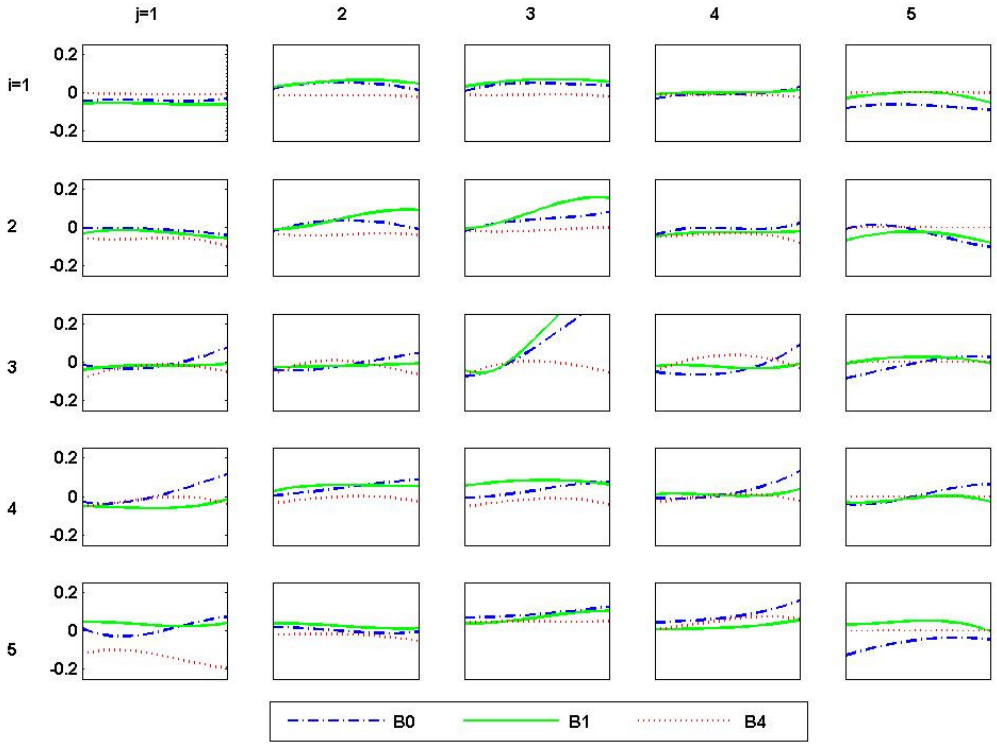
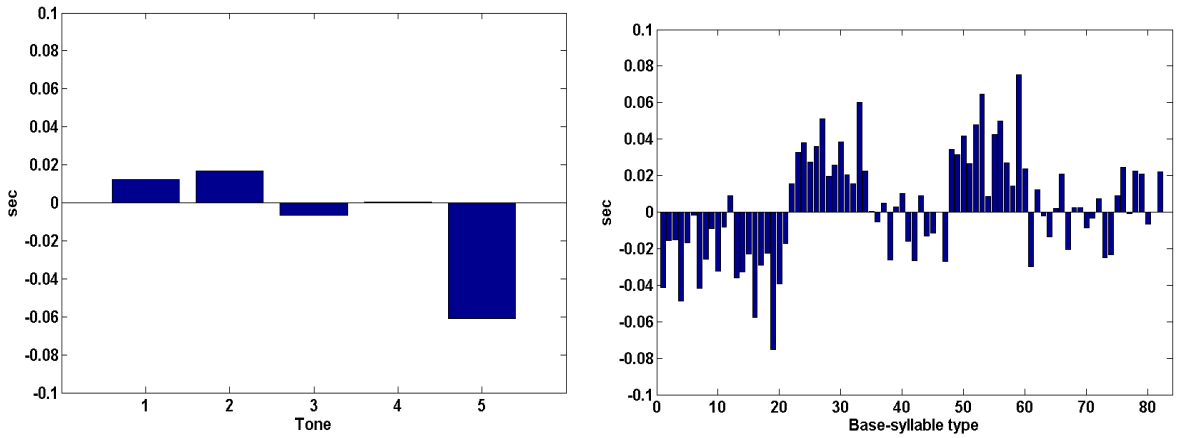


圖 3.4：基頻軌跡在停頓標記 *B0*、*B1* 和 *B4* 時後音節連音效應 APs，在此 $tp = (i, j)$

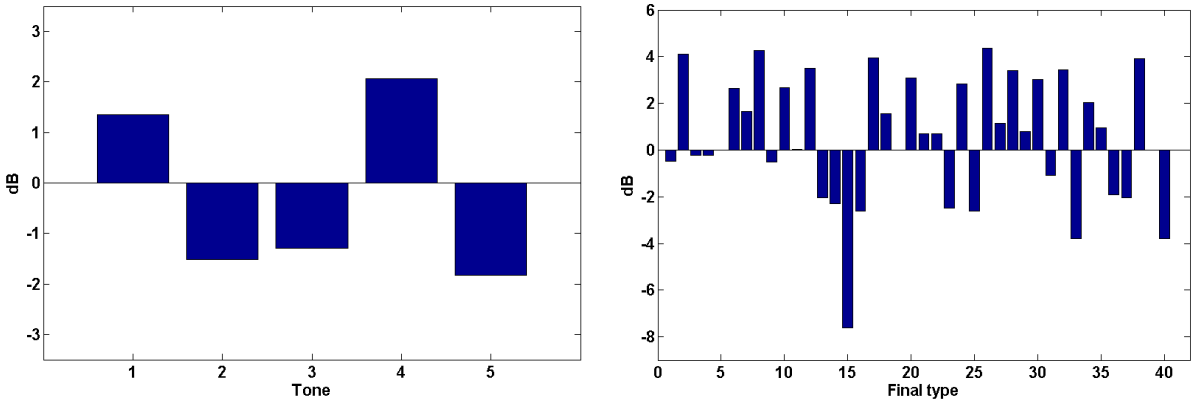
接下來為音節長度韻律模型分析，影響因子包含聲調、基本音節類型和韻律狀態。圖3.5(a)顯示音節長度的聲調APs，其中漢語一、二聲的音節長度都較長，五聲特別短，圖3.5(b)顯示音節長度的基本音節類型APs，此基本音節類型是把漢語411基本音節類型依發音特性分成82類，其中第19類的音節發音最長，此類對應到411音節類型包括”qu”、 ”du”和”bu”；第59類的音節發音為最短，對應到411音節類型包括”quan”、 ”qun”和”qiong”。



(a) (b)

圖 3.5：音節長度之(a)聲調 APs，(b)基本音節類型 APs

最後是音節能量位階韻律模型，影響因子包含聲調、韻母類型及韻律狀態。圖3.6(a)顯示音節能量位階的聲調APs，其中漢語以一、四聲音節能量位階最大，二、三和五聲則較小，圖3.6(b)顯示音節能量位階的韻母類型APs，在此韻母類型有40類，其中第19類的”wu”音節能量位階最小，此韻母類型對應到411音節類型如”su”、 ”tu”等；第26類的”wa”音節能量位階最大，此韻母類型對應到411音節類型如”zhua”、 ”gua” 等。



(a) (b)

圖 3.6：音節能量位階之(a)聲調 APs，(b)韻母類型 APs

利用修正型PLM演算法所標記出來的**{B, PS}**，搭配其所對應之語言參數{**t, s, f**}，可以圖3.7的方式重建不同語速的韻律聲學參數。藉由音節韻律模型模擬韻律聲學特徵參數$\hat{\mathbf{sp}}$, $\hat{\mathbf{sd}}$, $\hat{\mathbf{se}}$，

再利用語速正規化參數將$\hat{\mathbf{sp}}$, $\hat{\mathbf{sd}}$, $\hat{\mathbf{se}}$還原回各自原本的語速，得到最後的$\mathbf{sp}'$, $\mathbf{sd}'$, $\mathbf{se}'$。

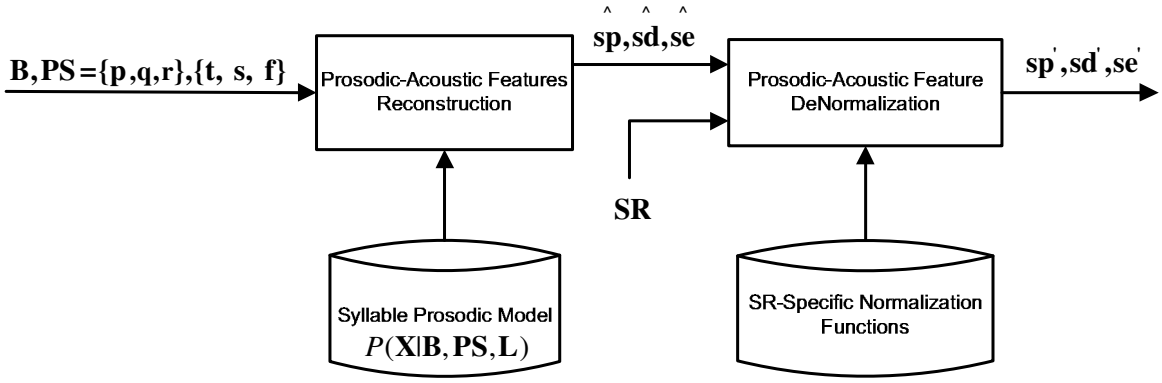


圖 3.7：以音節韻律模型及語速正規化參數來重建韻律聲學特徵參數之流程圖

表3.2列出以不同APs組合下，各韻律參數重建之總殘餘誤差值(Total Residual Error, TRE)，即扣除各種APs組合後，韻律參數殘餘值變異數與原始韻律參數變異數之比值。其中，加入韻律狀態APs後，各韻律參數之TRE都變得非常小，為所有影響因素中最重要的APs。

表 3.2：使用音節韻律模型不同 APs 組合下音節韻律參數之 TREs

Log-F0		**Duration**		**Energy level**	
APs	TRE	APs	TRE	APs	TRE
+Tone	66.9%	+Tone	70.2%	+Tone	61.2%
+Coarticulation	60.1%	+Base-syllable	51.1%	+Final	47.7%
+Prosodic state	0.7%	+Prosodic state	1.1%	+Prosodic state	1.4%

圖3.8利用$\boldsymbol{\beta}_t$和γ_t來模疑快語速與慢語速之五種聲調的音節基頻軌跡。從圖可發現不管快速或慢速，聲調五的長度皆為最短，其餘四個聲調長度則差異不大。基頻軌跡整體來說，快語速的基頻軌跡動態範圍(dynamic range)較慢語速小，因說話速度快使得基頻軌跡不完整，而說話慢時會產生較完整的五種聲調基頻軌跡，故快語速的基頻軌跡可視為慢語速中間的一部份[3]。接下來各別分析快慢語速的五種聲調：(1)一聲的基頻軌跡形狀大致無差異，(2)二聲的基頻軌跡會上揚，在慢速上揚程度最大，(3)三聲在慢速的基頻軌跡尾端會往下走，而快速較為平坦，(4)四聲基頻軌跡在慢速時的起始點較高、斜率較陡且動態範圍較大，快速則反之，(5)

輕聲的基頻軌跡在快速時較為平坦，在慢速時像低階的三聲。此模擬結果與[3]一致。

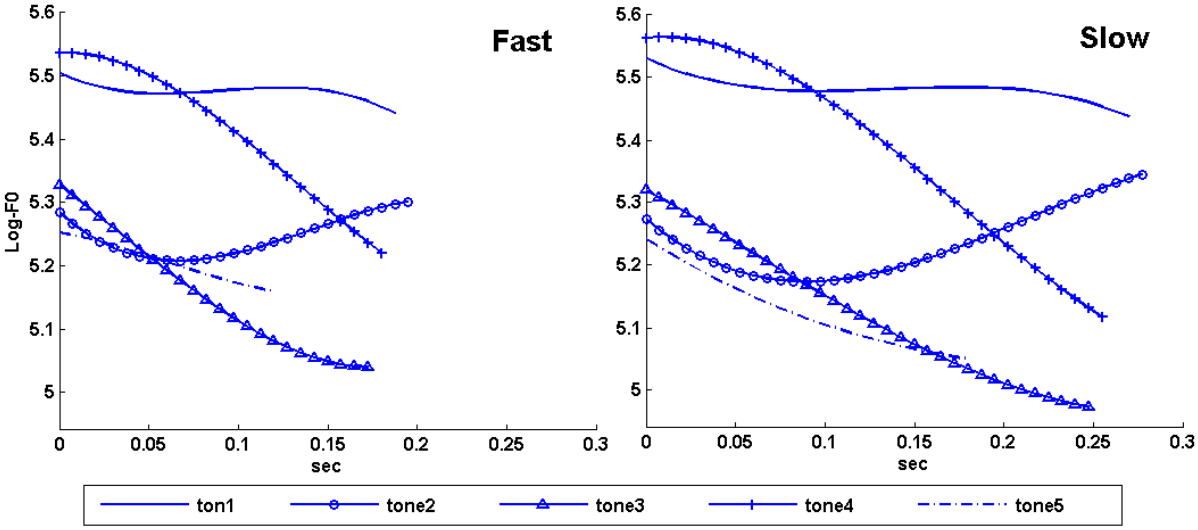


圖 3.8：快語速與慢語速之五種聲調基頻軌跡模擬圖

3.1.2 停頓聲學模型

停頓聲學模型由CART演算法建構而成，用以描述七種停頓標記**B**、語言參數**l**以及音節間韻律參數**{Y}={pd, ed}**和音節差韻律參數**{Z}={pj, dl, df}**之間的關係。圖3.9顯示在不同停頓標記下，決策樹根節點(root node)五種韻律參數的機率分佈。由圖可發現越上層韻律架構的停頓標記如*B3*、*B4*，擁有較長的停頓時長、較低的能量低點、較明顯的基頻跳躍及音節拉長因子；而*B0*、*B1*的停頓時長都非常的短，但*B0*的能量低點較大，表示*B0*為兩音節緊密連接的邊界；*B2-2*則有中等的停頓時長；*B2-1*和*B2-3*的能量低點與停頓時長分佈與*B1*相似，但*B2-1*擁有較明顯的基頻跳躍，*B2-3*則是音節拉長因子較為明顯。這些韻律參數的特性分佈符合本研究最初所定義之停頓標記特性。

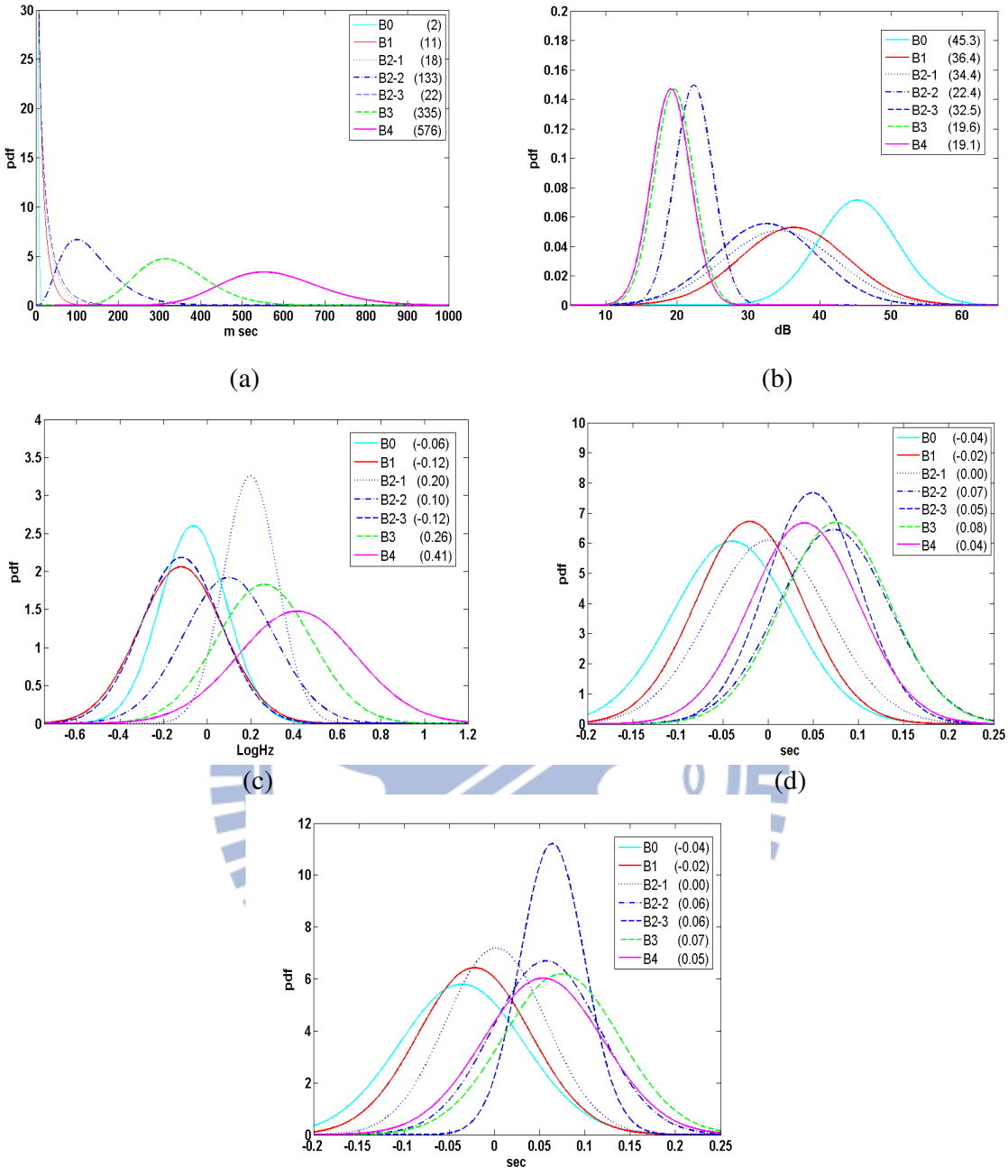


圖 3.9：(a)停頓音節長度，(b)音節能量低點，(c)正規化基頻跳躍值，(d)正規化音節拉長因子 1，(e)正規化音節拉長因子 2 之決策樹根節點機率分佈，其中括號中數值為分佈平均值

停頓時長為停頓標記中最重要的聲學參數，圖3.10顯示了七種停頓標記的平均停頓時長 vs. *SR*，圖上標出的值，為corpus-based HPM訓練結果[3]。此結果符合預期，*B0*、*B1*、*B2-1*和 *B2-1*等不具明顯停頓時長的停頓類別幾乎不受*SR*影響；而*B2-2*、*B3*和*B4*等具明顯停頓時長的

類別，其停頓時長隨著*SR*呈非線性增加，尤其*B3*、*B4*更是如此；另外，由圖得出此結果與corpus-based HPM訓練結果一致。

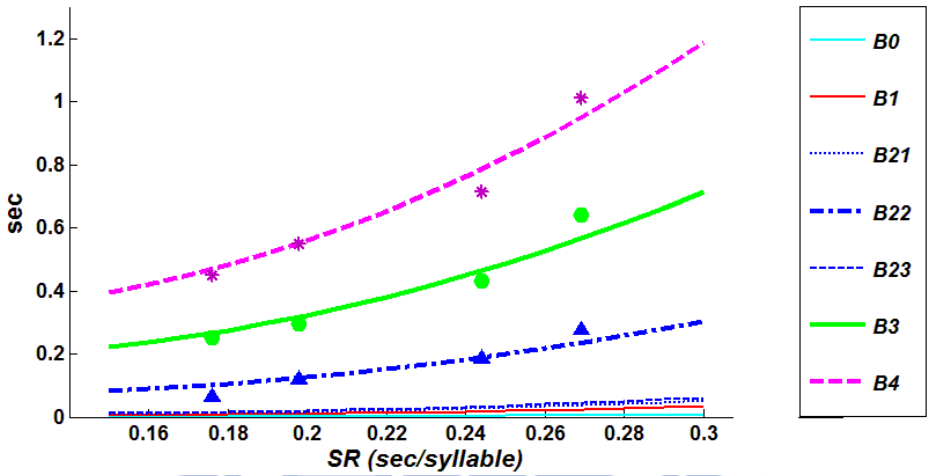


圖 3.10：平均停頓時長 vs. *SR*；標出值為四個語速 corpus-HPM 訓練結果

表3.3列出七種停頓類別之重建停頓時長的根均方差(Root Mean Square Errors, RMSEs)。其中只有*B2-2*、*B3*和*B4*誤差較大，因為這些停頓類別都為major break或minor break，故此結果尚可接受。

表 3.3：重建停頓時長之 RMSEs

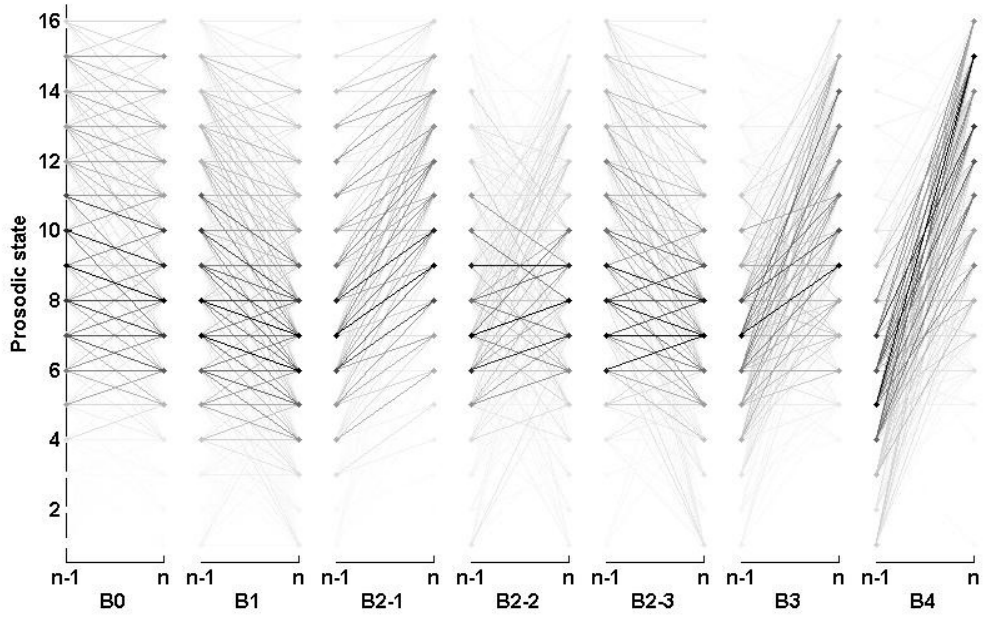
Break Type	B0	B1	B2-1	B2-2	B2-3	B3	B4
RMSE	2.4 ms	18.5 ms	24.9 ms	86.3 ms	30.8 ms	100.6 ms	147.8 ms

3.1.3 修正型韻律狀態模型

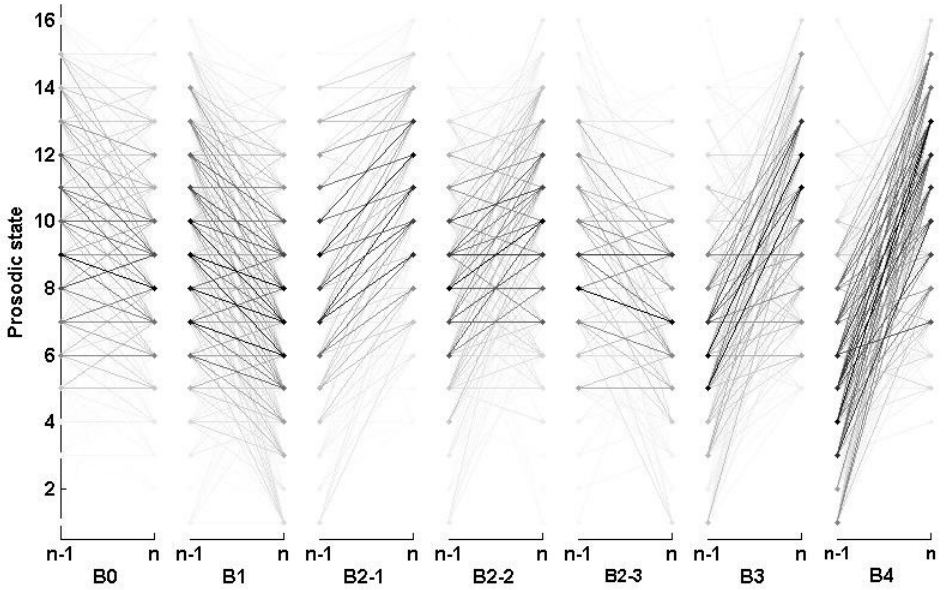
韻律狀態模型描述了韻律狀態於各停頓邊界的轉移情形，本論文所提出之修正型韻律態模型對語速分bin估計轉移機率值，藉此區分不同語速的狀態轉移情形。

圖3.11顯示音節基頻韻律狀態轉移情形，(a)為第一個bin(即最快語速)韻律狀態轉移情形，

(b)為最後一個bin(即最慢語速)韻律狀態轉移情形，圖上顏色越深的線表示其轉移情形越為重要。由圖發現無論語速快慢，*B0*與*B1*都以下降一或二階情形居多，表示在一個韻律詞之內，基頻韻律變化是由高緩慢至低的；*B2-1*、*B3*及*B4*的韻律狀態都有明顯low-to-high情形，顯示這些韻律邊界容易產生音高重置現象；而*B2-3*的轉移情形相似於*B0*/*B1*，表示*B2-3*並不以音高重置現象來代表韻律詞邊界。接著比較快語速及慢語速的韻律狀態轉移情形可發現，慢語速在*B4*的轉移變動較大，快語速在*B4*的轉移模式較為集中，*B4*的基頻韻律狀態轉移為不同語速差異最大的地方。



(a)

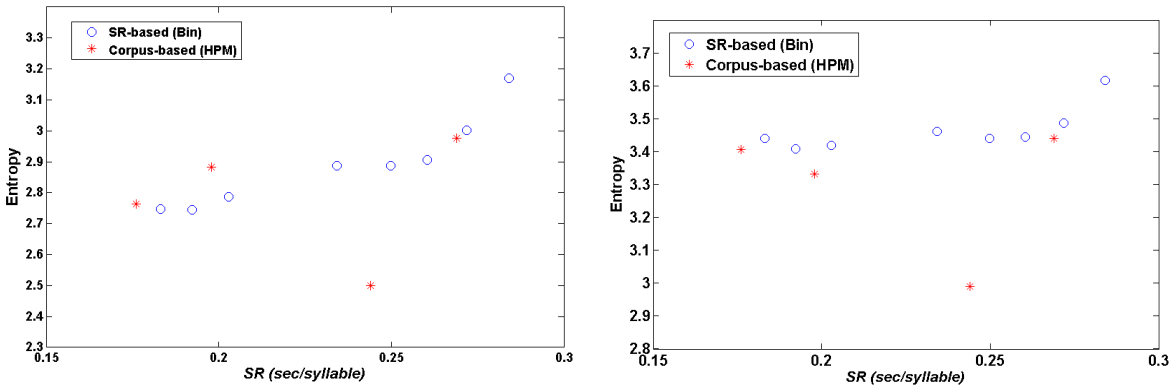


(b)

圖 3.11：(a)快語速，(b)慢語速於不同停頓標記下基頻韻律狀態的轉移情形。顏色越深表示轉移情形越重要

圖3.12是以條件熵(conditional entropy) $H(p_n \mid p_{n-1}, B_n = b)$ 量化基頻韻律狀態轉移對*SR*的關係，在此僅顯示*B0*和*B4*等最極端的韻律邊界。由圖發現無論*B0*或*B4*，entropy皆隨著*SR*增加

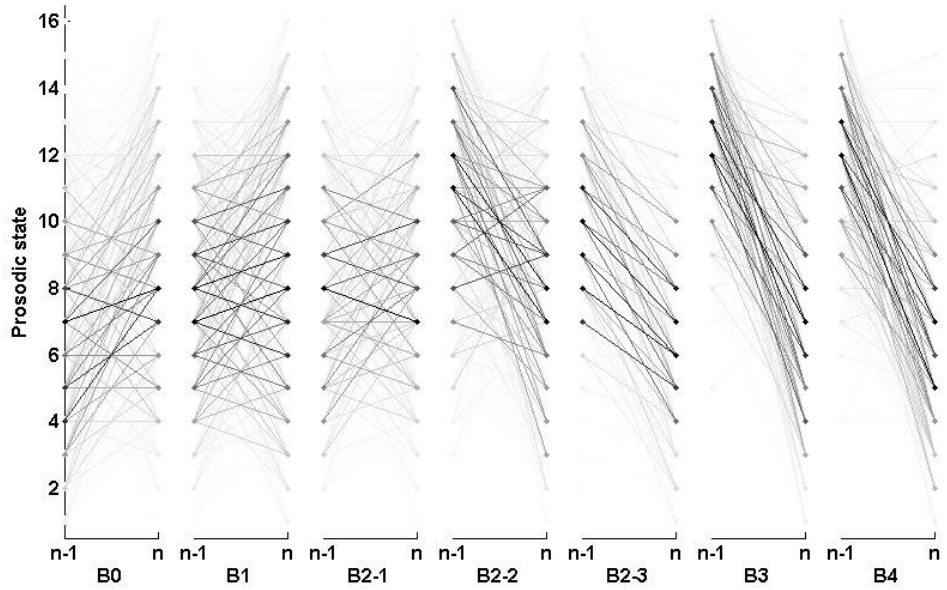
而有升高的趨勢，又以*B0*邊界尤其明顯。這代表語速越慢，基頻韻律狀態轉移越不一致，此結果更確定了韻律狀態轉移和語速相關聯。另外圖亦標上corpus-based HPM訓練結果，其結果僅中速語料較為不一致。



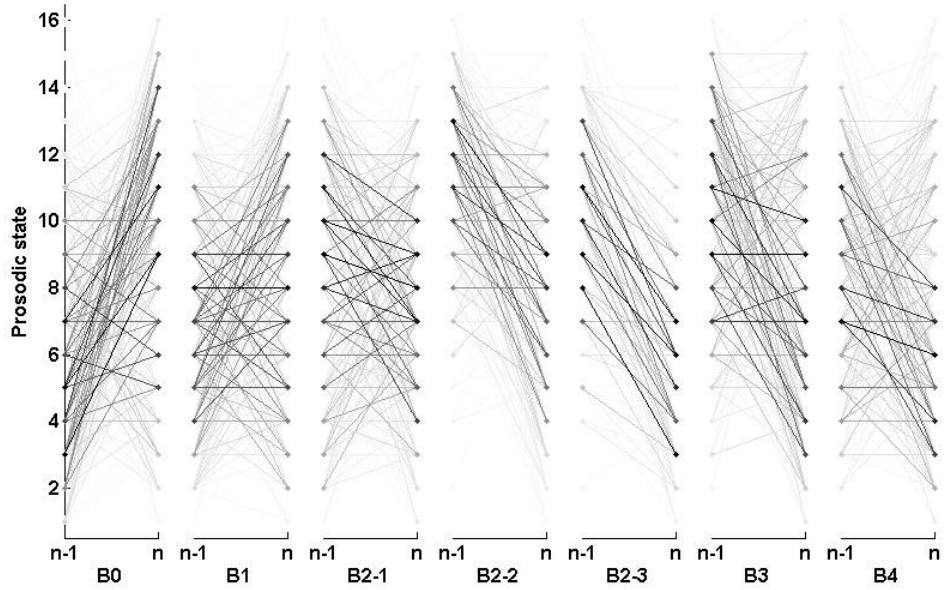
(a) (b)

圖 3.12：韻律標記為(a)B0，(b)B4 時基頻韻律狀態轉移 entropy vs. *SR*

圖3.13顯示音節長度韻律狀態轉移情形，*B3*、*B4*擁有最大範圍的high-to-low狀態轉移變遷，代表在PPh和BG/PG等大韻律單元邊界容易產生final lengthening effect；*B2-2*則為較小的high-to-low狀態轉移變遷，final lengthening effect不如*B3*、*B4*強烈，*B2-3*亦是如此，表示即使無明顯停頓時長，仍可以音節拉長現象來反應此為一韻律詞邊界。在不同語速比較中，可發現快語速在*B3*、*B4*擁有大範圍的high-to-low狀態轉移變遷，且其轉移模式較固定，而慢速則轉移範圍較小，轉移模式為散亂不固定；另外*B0*、*B1*部份，慢語速有較快語速更明顯的low-to-high狀態轉移變遷，表示一個韻律詞內慢語速的韻律變化是由低至高的。



(a)

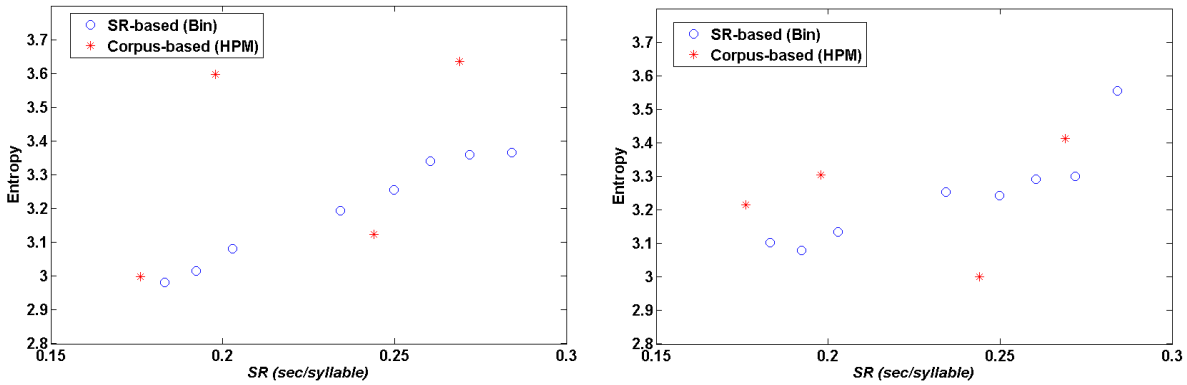


(b)

圖 3.13：(a)快語速，(b)慢語速於不同停頓標記下音長韻律狀態的轉移情形。顏色越深表示轉移情形越重要

圖3.14顯示音節長度韻律狀態轉移的conditional entropy $H(q_n \mid q_{n-1}, B_n = b)$ 對*SR*的關係；此結果與基頻韻律狀態的類似，都是*SR*越大時entropy越大。在corpus-based HPM只有快速語料較

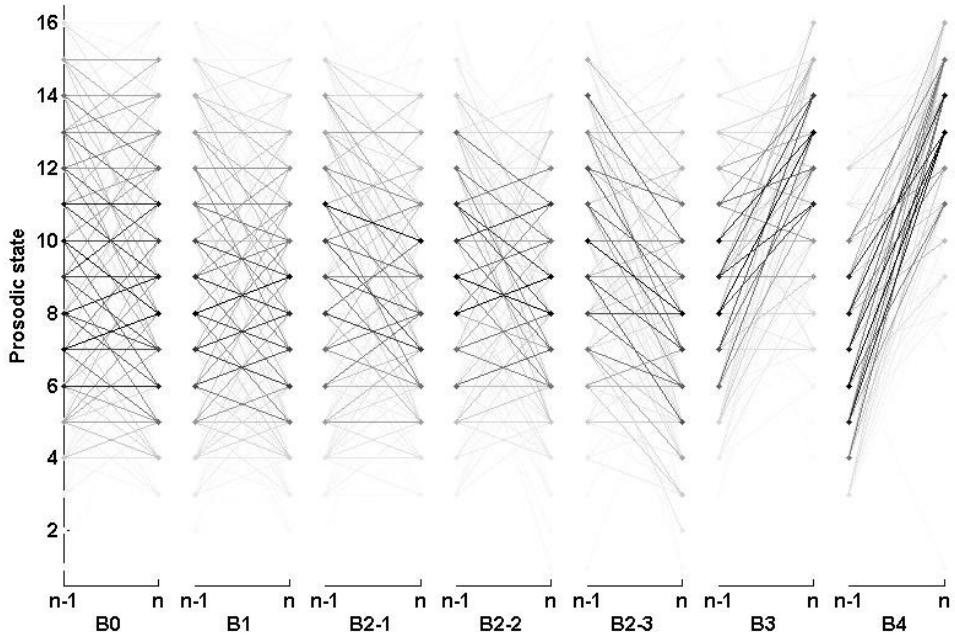
為一致，正常語速和慢速語料的entropy皆偏高，可能是其語料有極端語速的語句所造成之結果。



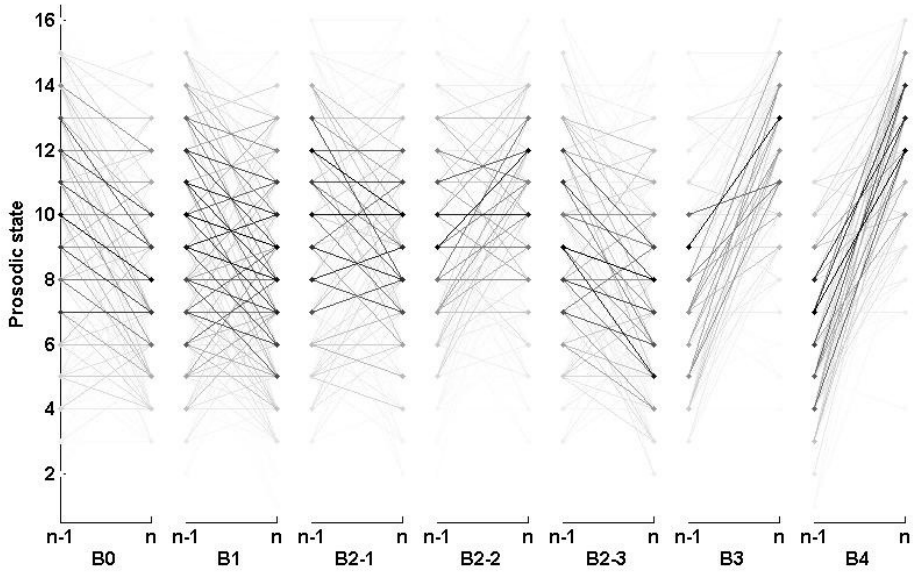
(a) (b)

圖 3.14：韻律標記為(a)B0，(b)B4 時音長韻律狀態轉移 entropy vs. *SR*

圖3.15顯示了音節能量韻律狀態的轉移情形，在*B3*、*B4*有大範圍low-to-high轉移情形，表示PPh和BG/PG等後邊界的音節能量會降至很低，再由新的韻律單元起始將能量重新提高，此結果驗證了在PPh和BG/PG裡，音節能量的趨勢是由高衰減至低的，接著再進行能量重置。而不同語速能量韻律狀態的轉移情形差異不大，表示人說話的能量變化和*SR*相關性是很小的，如2.3.4所討論。



(a)



(b)

圖 3.15：(a)快語速，(b)慢語速於不同停頓標記下能量韻律狀態的轉移情形。顏色越深表示轉移情形越重要

3.1.4 修正型停頓語法模型

修正型停頓語法模型之建構是由兩個步驟完成，第一步是以CART演算法訓練一顆決策樹，根據語言參數對不同類型的停頓標記分類，再對所有終止節點估計 $p(B_n|\mathbf{l}_n)$。其中本論文在CART演算法的兩個分裂停止條件如下：

1. 決策樹分裂出之子節點，其最小樣本數必須大於750。
2. 決策樹訓練過程中，其相對相似度增益(relative likelihood gain)必須大於0.01。

採用這兩個設定值是為了控制決策樹不要長過深，由於建構修正型停頓語法模型的第二步為將所有終止節點的統計機率對*SR*展開，需要足夠的樣本數來完成估計 $p(B_n|\mathbf{l}_n, SR_n)$。決策樹訓練出的結果如圖3.16所示，每個節點都有對應的編號及問題，編號1即代表根節點，節點內的直方圖為該節點停頓標記之機率分佈，由左至由右分別為*B0*、*B1*、*B2-1*、*B2-2*、*B2-3*、*B3*、*B4*，節點內數值為該節點的總樣本數，另外，實線表示父節點問題為“是”，虛線則為“否”。在根節點所問之第一個問題為PM，由節點2之直方圖可得知大部份樣本數都集中於*B3*和*B4*，由此顯示大部份的*B3*和*B4*都產生於標點符號處；由節點4直方圖分佈可確信在標點符號為逗號時，其對應的停頓標記幾乎皆為*B3*或*B4*，代表語者經常以逗號來當作一個的PPh和BG/PG結尾，若標點符號不為逗號，則很有可能是句號，故節點5中的*B4*機率很高；從節點2往下長，問題集都偏向句子層次的語言參數居多，例如LPS>=7(前一個句子的長度是否大於等於7)，表示這邊的韻律組成份子邊界大多屬於major break。接下來分析韻律邊界為non-PM的部份，節點3以是否為inter-word邊界來分裂節點，若為”否”，即intra-word邊界，如節點7所示，幾乎都屬於*B0*、*B1*等non-break類別；若為inter-word邊界，則如節點6所示，此類之停頓時長可能短至*B0*、*B1*等non-break類別，長至接近major break的*B3*，造成此類型邊界不容易得到一致性的標記結果，推測原因可能是語速不同而造成的混淆，此節點往下長都為語速影響顯著的節點，也是最需要補償的地方。

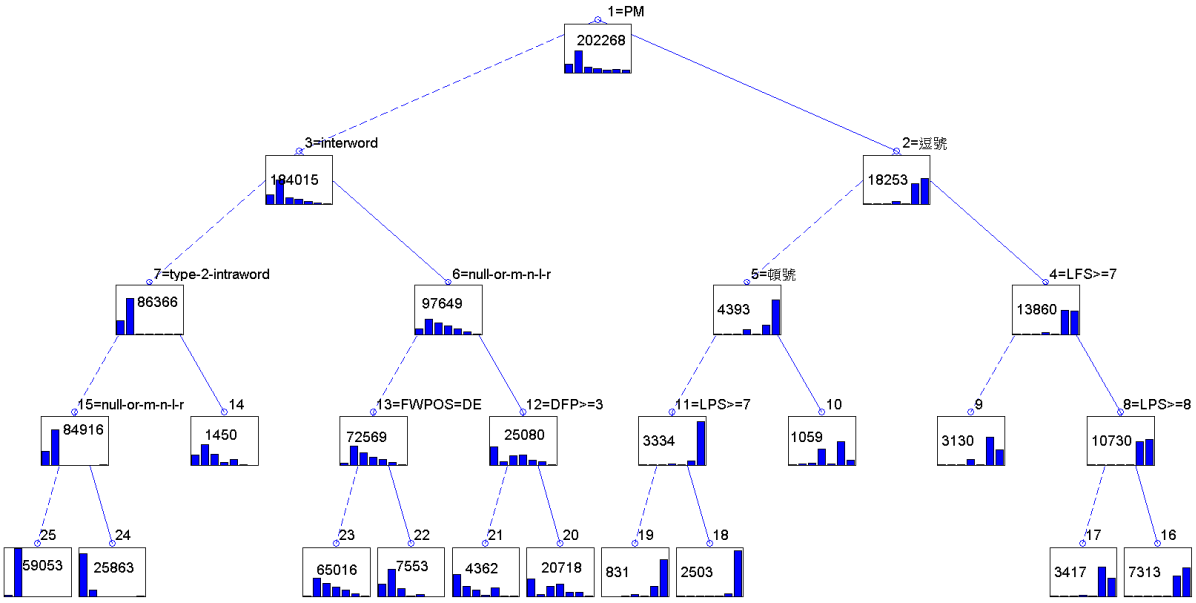


圖 3.16：停頓語法模型決策樹，節點中直方圖為各停頓標記的發生機率，由左至右分別是 *B0*, *B1*, *B2-1*, *B2-2*, *B2-3*, *B3*, *B4*，數值為該節點的總樣本數

接下來進行建構修正型語法模型第二步，在決策樹每一終止節點考慮語速的影響，藉由線性迴歸的方式得到 $p(B_n | 1_n, SR_n)$。圖3.17顯示了三個例子，分別為：(a)屬於major break的*B4*在PM node(即節點2)，(b) minor break中擁有短停頓的*B2-2*在non-PM inter-word node(即節點6)，以及(c)屬於non-break的B1在intra-word node(即節點7)。由圖可觀察到例子(b)*B2-2*在快語速的發生頻率很低，隨著*SR*增加其頻率呈線性增加；例子(a)*B4*擁有和例子(b)*B2-2*類似的趨勢，但斜率較不明顯；例子(c)*B0*在的情形則是和上述兩例相反，發生頻率在低*SR*時較高而高*SR*時較高。綜合以上觀察總結：在non-PM, interword的韻律邊界minor break受語速影響最嚴重的地方；而在標點符號的韻律邊界，無論語速快慢都容易出現major break，*SR*在此的影響不大；在intra-word韻律邊界亦是如此，non-break出現的頻率並不因*SR*而有明顯變化。

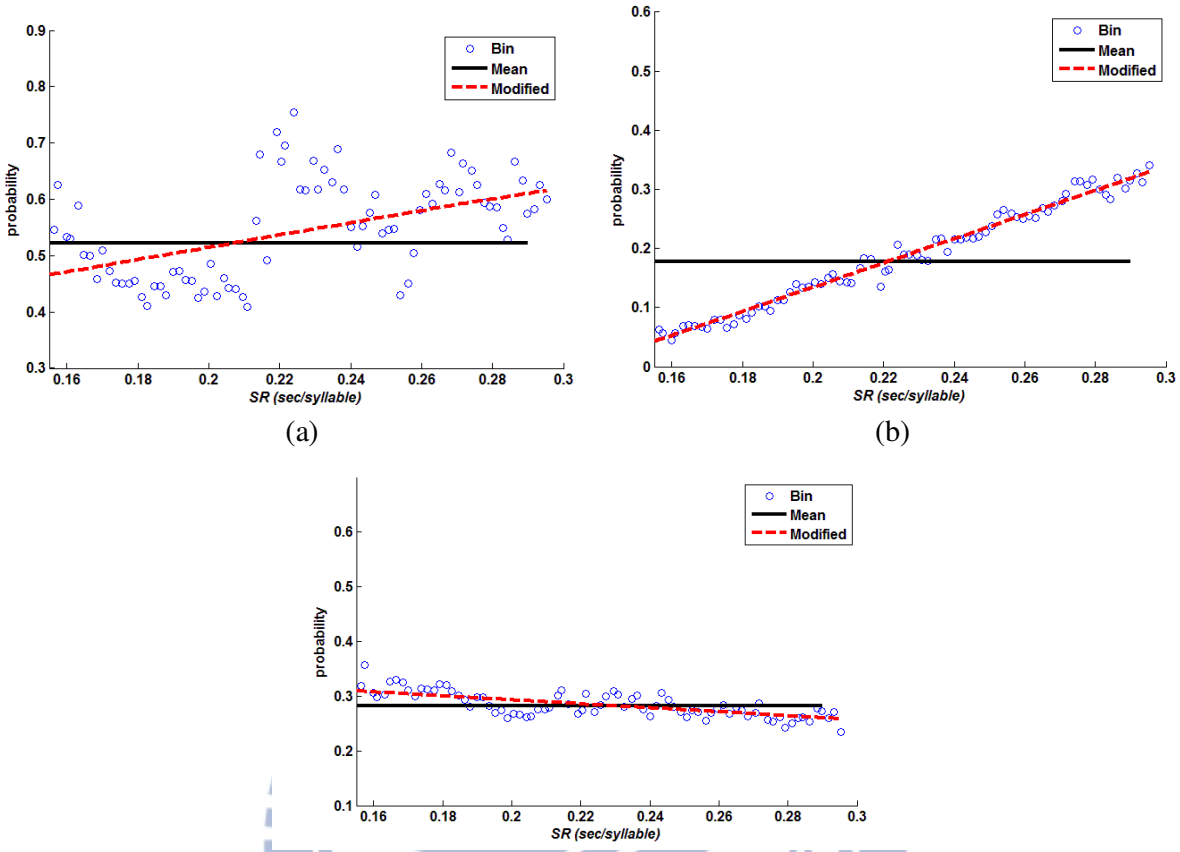


圖 3.17：(a) *B4* 於 PM 節點，(b) *B2-2* 於 non-PM, inter-word 節點，(c) *B0* 於 intra-word 節點之發生頻率 vs. *SR*

接下來由圖3.18觀察語速對於整個語法決策樹的影響，圖中節點的直方圖表示七種停頓類別發生機率對*SR*之斜率。在此分三個部份來討論，(1) intra-word node往下長的部份之停頓類別大多為*B0*或*B1*，大部份的斜率值都很低，表示停頓機率與*SR*相關性很小，除了type 2 intra-word node屬於較不緊密的intra-word邊界，其受*SR*的影響類似於non-PM, inter-word。(2) non-PM, inter-word node往下長的部份為*SR*影響最明顯之處，*B2-2*的斜率都為明顯正值，*B0*和*B1*大多為負值，表示語速由慢轉快時，部份*B2-2*開始轉為*B0*或*B1*；相反地，語速由快轉慢時，*B2-2*開始增加，*B0*、*B1*相對減少。(3)最後為PM node往下長的部份，此處停頓類別大多為major break的*B3*和*B4*，有中等的斜率值，其中*B4*有較*B3*稍大的絕對斜率，表示語速由快轉慢時，有部份的*B3*會轉成*B4*。

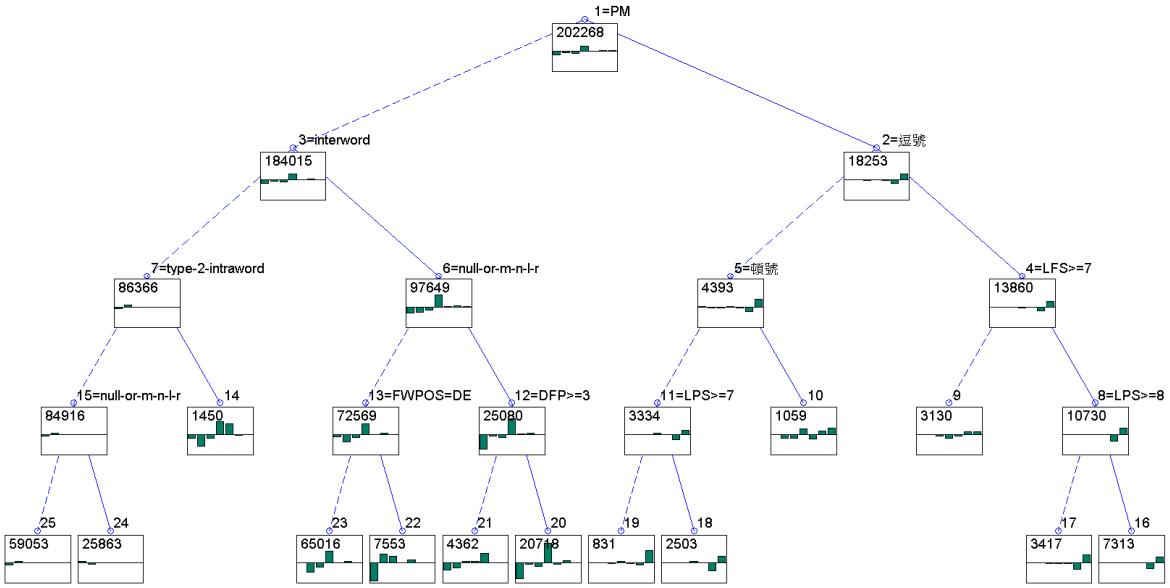


圖 3.18：停頓語法模型決策樹，節點中直方圖為各停頓標記機率對 *SR* 之斜率，由左至右分別是 *B0*, *B1*, *B2-1*, *B2-2*, *B2-3*, *B3*, *B4*，數值為該節點總樣本數

表3.4比較了停頓語法模型與修正型停頓語法模型的conditional entropy，即 $H(B|L)$ 與 $H(B|L,AR)$。從表中可看到，停頓語法模型的entropy在加入語速考慮後降低了約0.027，而主要的改善正是來自於non-PM, inter-word的部份，如本章節前一段所討論之。

表 3.4：停頓語法模型修正前後 entropy 之比較

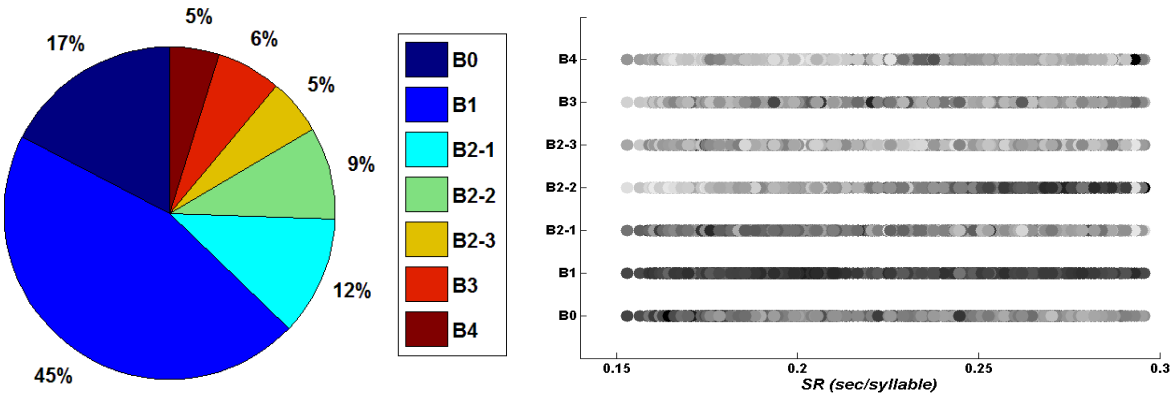
	All	PM part	Non-PM, inter-word part	Intra-word part
Break-Syntax Model (BSM)	**1.1012**	0.0970	**0.8976**	0.1066
Modified BSM	**1.0746**	0.0958	**0.8741**	0.1048

3.2 韻律標記結果之分析

此節分別對韻律狀態、停頓類別的標記結果進行分析，並探討韻律標記與語速之間的關係。

3.2.1 停頓類別標記

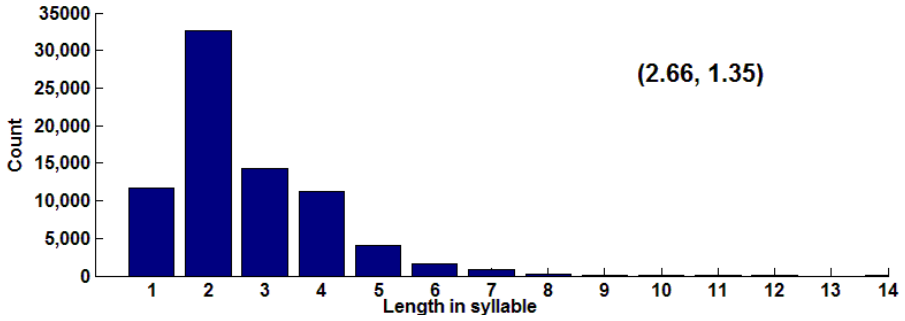
圖 3.19(a)顯示七種停頓類別在語料庫中所佔百分比；(b)則是各個停頓類別在語速的分佈情形，顏色越深代表比例越高。由圖可觀察到 *B2-2* 和 *B4* 在慢語速所佔比例有較高趨勢，以 *B2-2* 尤其明顯；而 *B0* 在極快語速時佔有最高比例，表示語速快到一定程度時，很容易發生 tightly coupling 的韻律邊界。



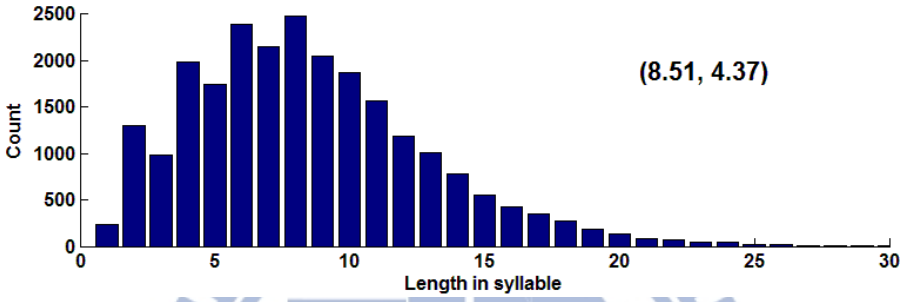
(a) (b)

圖 3.19：(a)各停頓標記在語料庫所佔百分比，(b)各停頓標記在 *SR* 之分佈情形(顏色越深代表比例越高)

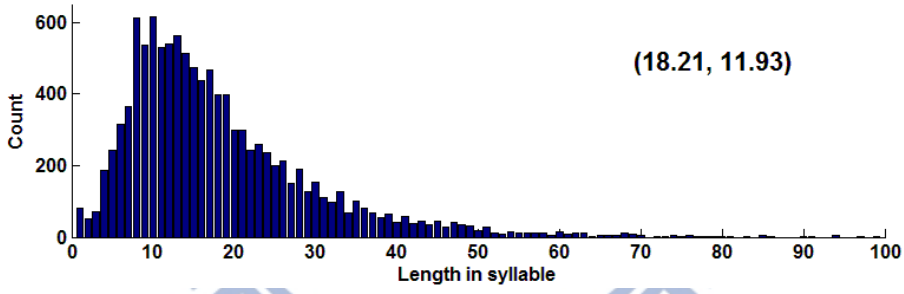
在標記完停頓類別後，可將語句區分為各種韻律組成份子，即韻律 PW，(b)PPh 和(c)BG/PG。圖 3.20 顯示各種韻律成份子的音節數目分佈，PW 通常由單字或 2 到 7 個字所組成；PPh 由 2 到 15 個字所組成；BG/PG 則是 5 到 30 個字組成，代表一個大型段落；可得知越上層的韻律組成份子，音節長度之平均值及標準差就越大。



(a)



(b)



(c)

圖 3.20：(a)PW，(b)PPh，(c)BG/PG 之音節個數直方圖，括弧內數字分別為其平均值及標準差

圖 3.21(a)、(c)、(e)分別顯示 PW、PPh 以及 BG/PG 韻律組成份子之音節平均數目與 *SR* 的關係，其中 PW 受語速影響最強烈、PPh 次之，音節數目平均值隨著 *SR* 增加而減少；BG/PG 則幾乎不受語速影響。圖 4.13(b)、(d)、(f) 分別顯示 PW、PPh 和 BG/PG 韻律組成份子之音節數目標準差與 *SR* 的關係，PW 的音節數目標準差隨著 *SR* 增加而下降；PPh 和 BG/PG 的標準差都偏大，與 *SR* 幾乎無相關性。

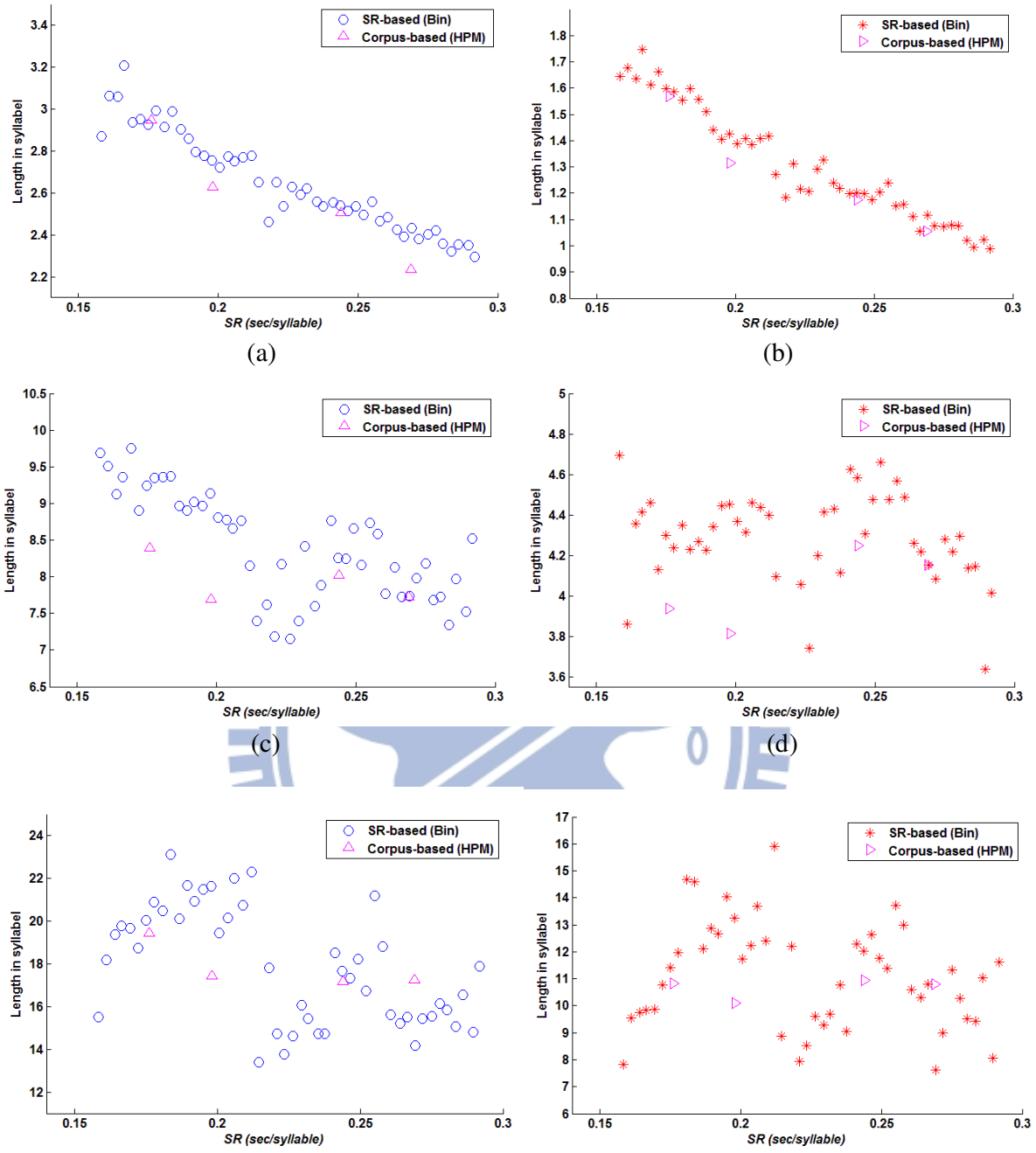


圖 3.21：(a)PW，(c)PPh 和(e)BG/PG 音節個數平均值 vs. *SR*；(b)PW，(d)PPh，(f)BG/PG 音節個數標準差 vs. *SR*

圖 3.22 為一停頓標記結果的範例，圖中列出一組語料庫平行語句，分別為快速、正常語速、中速和慢速，在此只標示具明顯停頓時長之停頓類別，即 *B2-2*、*B3* 和 *B4*。由範例可看出，在 PM 出現的音節邊界，四個語速的標記一致，皆為 *B3* 或 *B4*；而在 non-PM, inter-word 部份，

四種語速標記結果開始為不一致，如以快語速範例為基準，正常語速 *B2-2* 的比例即開始增加，中、慢語速則是連 *B3* 和 *B4* 的比例也相繼增加；在快語速中，為使語句快速流暢而忽略了許多詞邊界的停頓，只在標點符號出現時才有停頓；反之，在慢語速中，除了部分緊密結合的詞邊界，在標點符號和一般的詞邊界幾乎都有停頓發生。

快速語料之範例：

依據行政院主計處的統計 @，十月份 * 一到二十日 / ，我國出口及進口金額 / 比起去年同期 * 均有增加 @，

正常語速語料之範例：

依據行政院主計處的統計 @，十月份 * 一到二十日 /，我國出口 * 及進口金額 / 比起去年同期 * 均有增加@，

中速語料之範例：

依據 * 行政院主計處的統計 @，十月份 / 一到 * 二十日 /，我國出口 * 及進口金額 / 比起去年同期 * 均有增加 @，

慢速語料之範例：

依據 / 行政院 * 主計處的統計 @，十月份 / 一 * 到 * 二十日 @，我國出口 * 及進口金額 / 比起去年同期 * 均有增加 @，

圖 3.22：語料庫平行語句之停頓標記範例，只標示具明顯停頓時長之類別（*B2-2*(*)、*B3*(/)和 *B4*(@)）

3.2.2 韻律狀態標記

圖 3.23 為韻律狀態被自動化標記的一個範例。由圖中可觀察到(global mean + prosodic state)有相對於 observed 較平滑的曲線，即表示韻律上的趨勢變化；在 *B4* 韻律邊界可觀察到大範圍的基頻重置現象，*B2-1*、*B2-2* 和 *B3* 則是小範圍的基頻重置現象；在 *B3*、*B4* 可觀察到音長 final lengthening 的現象，這些都符合本論文對停頓類別之定義。

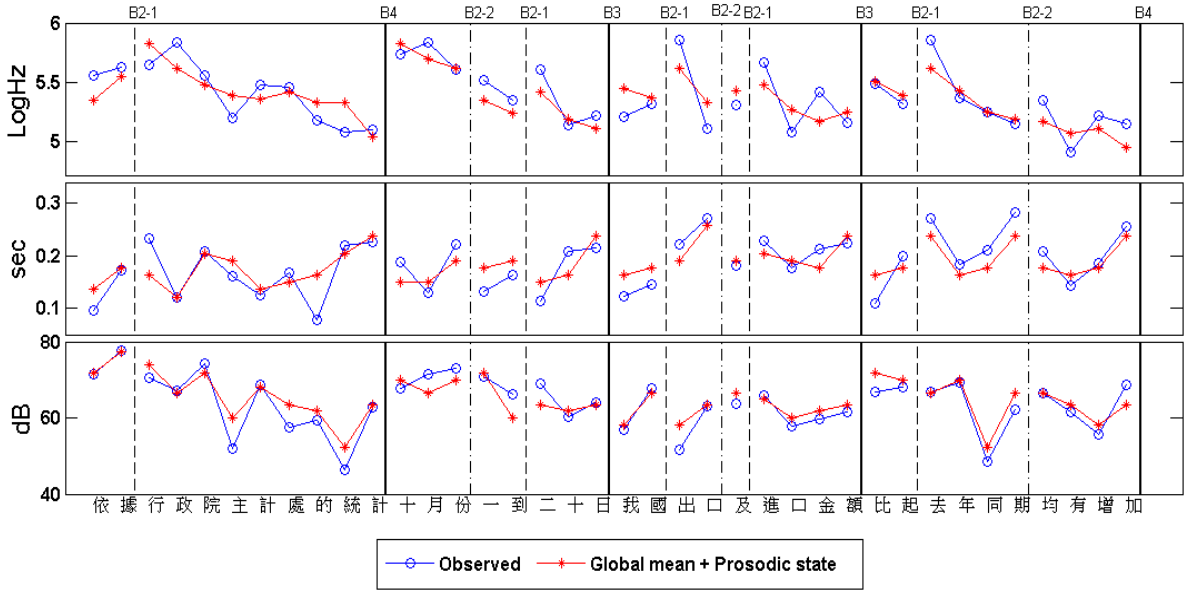


圖 3.23：韻律狀態標記範例，由上至下分別為音節基頻平均值、音節長度和音節能量位階的(global mean + prosodic state)和 observed

第四章 可控制語速之 TTS 應用

在語音合成中，語速控制使應用上更有彈性，如 1.1 節所討論。本章討論基於 SR-HPM 來實現一可控制語速之 TTS 系統(SR-TTS)；作法為依據給定的 *SR*，由 SR-HPM 來產生對應之韻律參數，最後藉結合 HMM-based 語音合成器完成之，其合成流程可參考圖 4.1。

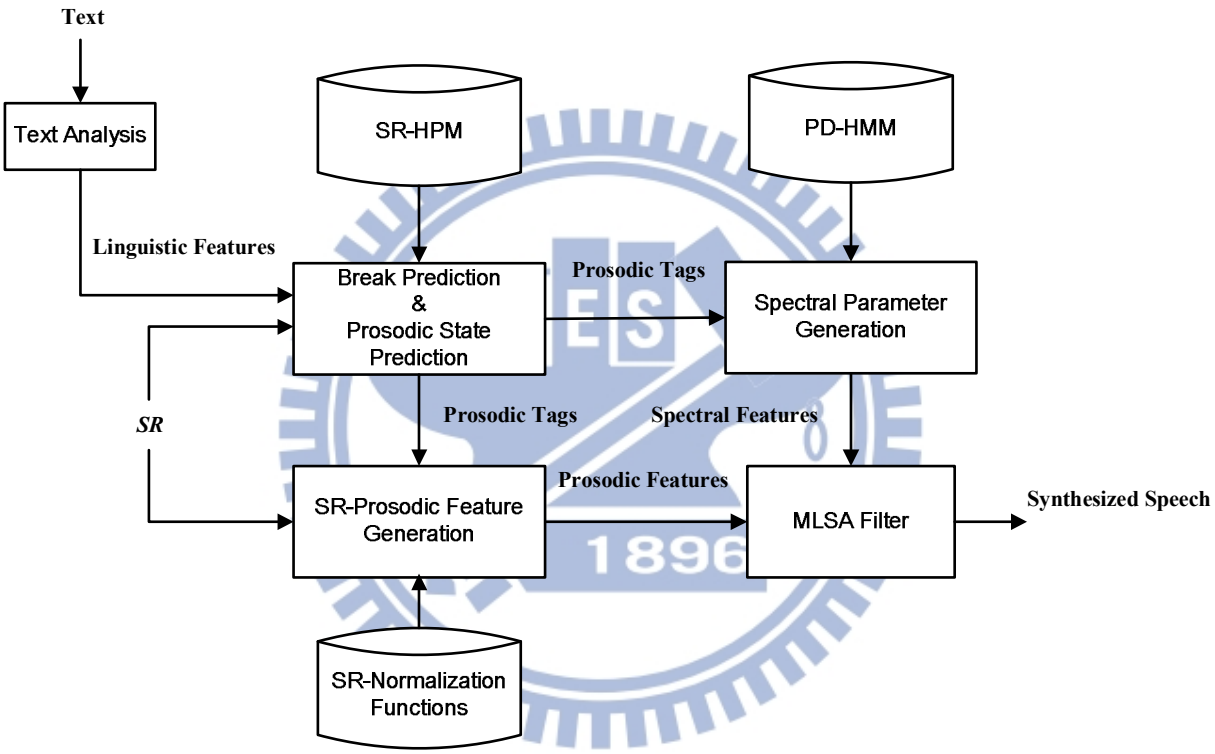


圖 4.1：可控制語速之 TTS 系統架構圖

本研究定義停頓標記和韻律狀態標記來描述韻律階層架構，訓練階段給定了聲學參數與語言參數，以修正型 PLM 演算法完成韻律標記過程；在 TTS 合成階段，因輸入僅有文字部份，因此只能由其產生的語言參數去預估韻律標記，本章重點在於如何以語言參數預估不同語速的韻律標記。4.1 節介紹停頓標記的預估方法，4.2 節介紹韻律狀態標記的預估方法，4.3 節介紹韻律參數產生方法，4.4 節介紹本論文所使用的頻譜模型，4.5 節為語音合成實驗結果與分析。

4.1 停頓標記預估

停頓預估的好壞將影響合成語音之流暢度，對於不同語速的語音，停頓發生的情形會有所不同。在此，使用SR-HPM訓練出的修正型停頓語法模型 $P(\mathbf{B}|\mathbf{l}, \mathbf{SR})$ 來預估停頓標記。此模型描述了語言參數、語速和七種停頓類別的關係，其預估方法如下：

$$B_n^* = \arg\max_{B_n} p(B_n | \mathbf{l}_n, SR_n) \tag{4.1}$$

表 4.1 顯示訓練語料停頓標記的預估結果，由表可得知(1)non-break 類別的 *B0*、*B1* 擁有最高的正確率，而 *B0* 容易預估成 *B1*，但因此二類皆被定義為韻律詞之詞內邊界，此錯誤結果對於整體的韻律表現影響不大；(2)minor break 中的 *B2-2* 預估效果較佳，*B2-1* 和 *B2-3* 的正確率皆偏低，大部份都被預估成 *B1*，可能是因 *B1* 在 non-PM, inter-word 的邊界數量過多，導致決策樹同一個終端節點內，*B1* 的機率相對要大的多；另外也可能 *B2-1*、*B2-3* 和語言參數的相關性不大，使得決策樹無法產生出鑑別此二類的終端節點。不過 *B2-1*、*B2-3* 都為不具明顯停頓長度的類別，因此錯預估為 *B1* 對於語音合成影響並不大；(3) major break 方面，*B4* 容易和 *B3* 混淆，而 *B3* 又容易和 *B2-2* 混淆，原因是 *B2-2*、*B3*、*B4* 皆以停頓時長來區分韻律邊界，聲學特性過於相近導致對應的語言參數也相類似。表 4.2 為測試語料之停頓標記預估結果，其預估清況與訓練語料一致。

表 4.1：訓練語料之停頓標記預估辨識率

Tar\Pre	B0	B1	B2-1	B2-2	B2-3	B3	B4	Total
B0	**85%**	9%	3%	2%	0%	0%	0%	32272
B1	5%	**87%**	5%	2%	0%	0%	0%	82235
B2-1	9%	34%	**44%**	12%	1%	1%	0%	21305
B2-2	6%	12%	18%	**53%**	1%	10%	0%	16211
B2-3	12%	40%	23%	20%	**4%**	2%	0%	9897
B3	2%	4%	4%	19%	0%	**44%**	27%	11561
B4	0%	0%	0%	1%	0%	20%	**79%**	8956

表 4..2：測試語料之停頓標記預估辨識率

Tar\Pre	B0	B1	B2-1	B2-2	B2-3	B3	B4	Total
B0	**86%**	9%	2%	2%	0%	0%	0%	3034
B1	5%	**87%**	5%	3%	1%	0%	0%	9506
B2-1	7%	34%	**41%**	16%	2%	1%	0%	2258
B2-2	6%	9%	16%	**56%**	1%	13%	0%	1985
B2-3	7%	39%	22%	25%	**4%**	2%	0%	1076
B3	2%	2%	4%	18%	0%	**39%**	36%	1218
B4	0%	0%	1%	1%	0%	13%	**86%**	754

4.2 韻律狀態預估

韻律狀態描述基頻、音長的韻律變化，其預估的好壞直接影響音節單元的「抑揚頓挫」，甚至是語意的傳達。在TTS中，只能由語言參數來預估韻律狀態，故引入模型 $p(p_n \mid \mathbf{l}_n)$、$p(q_n \mid \mathbf{l}_n)$ 來描述韻律狀態和語言參數的關係，搭配SR-HPM訓練出來的修正型韻律狀態模型 $P(\mathbf{PS} \mid \mathbf{B}, SR)$，最後以句子為單位，採取維特比演算法來預估韻律狀態，如4.2式。

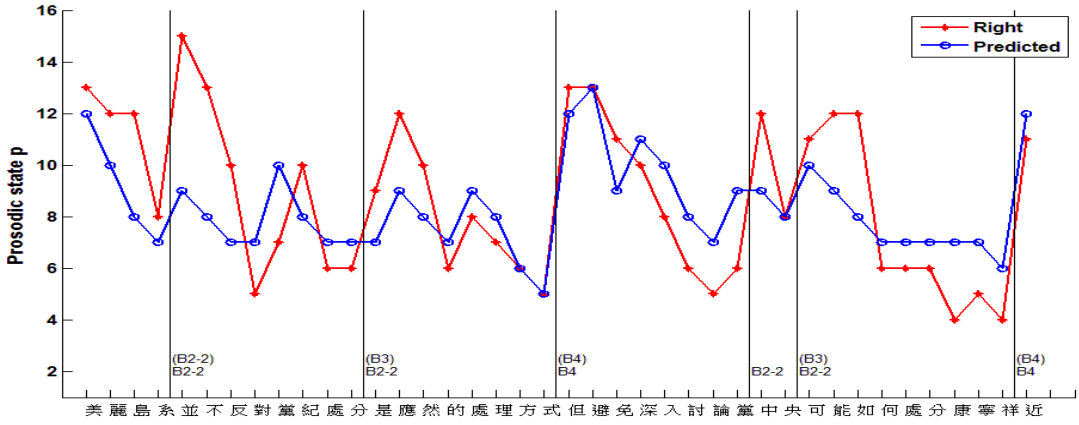
$$\mathbf{p}^*,\ \mathbf{q}^* = \arg\max_{\mathbf{p},\mathbf{q}} \left(\frac{p(p_1|\text{bin}(SR_1))\,p(q_1|\text{bin}(SR_1)) \prod_{n=2}^{N} p(p_n \mid p_{n-1}, B^*_{n-1}, \text{bin}(SR_n))\,p(q_n \mid q_{n-1}, B^*_{n-1}, \text{bin}(SR_n))}{\left(\prod_{n=1}^{N} p(p_n \mid \mathbf{l}_n)\,p(q_n \mid \mathbf{l}_n) \right)} \right) \tag{4.2}$$

其中 $p(p_n \mid \mathbf{l}_n)$、$p(q_n \mid \mathbf{l}_n)$ 以CART實現之，所使用的語言參數如表4.3，此模型可視為韻律狀態之靜態模型，可確保韻律狀態在語言參數的定位點；$P(\mathbf{PS} \mid \mathbf{B}, \mathbf{SR})$ 用以限制韻律狀態的轉移關係，可避免不合理的狀態轉移情形發生，此處也包含了語速之考量；B^*_{n-1} 為4.1節的停頓標記預估結果。

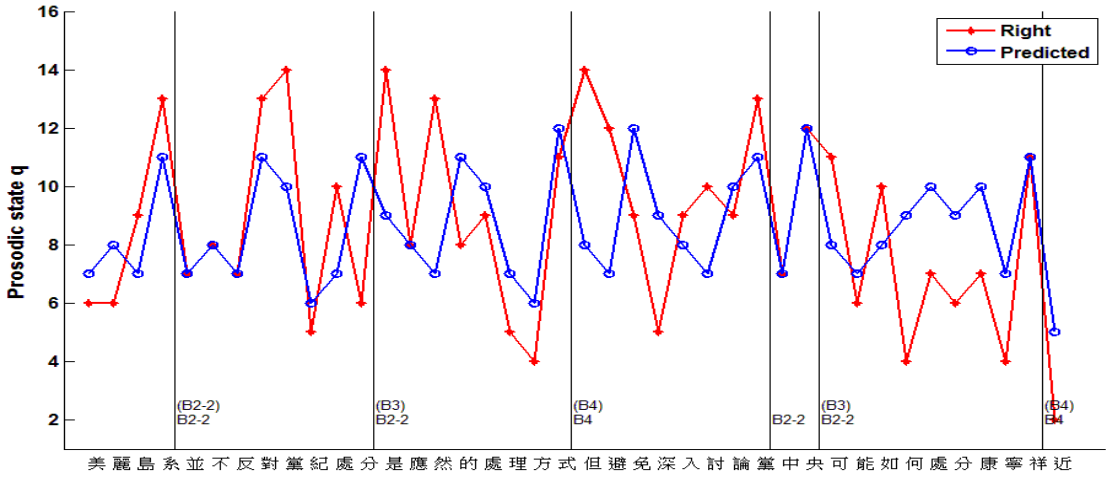
表 4..3：韻律狀態靜態模型之語言參數列表

Current word length in syllable: {1, 2, 3, 4, >4}.
Current syllable position in word: {1st, intermediate, last, mono-syllable word}.
Sentence length in syllable: {1, [2,5], [6,10], [11,15], [16,20], >20}.
Current syllable position in sentence: {1st, 2nd, 3rd, [4th, 5th], [6th, 7th], [8th, 11th], last, 2nd last, 3rd last, [5th last, 4th last], [7th last, 6th last], [11th last, 8th last], others}; Smaller count number from the beginning or end wins.
PM after the current syllable (five types).
POS3: 47-types POS.
Type of PM: comma, period, question mark, dun hao and others
Length of sentence in syllable

圖4.2為一個*SR*=1.97 的(a)基頻韻律狀態和(b)音長韻律狀態預估結果與正確標記的比較範例，其中括號內為正確的停頓標記，無括號為預估的標記。由圖可看出預估結果與正確標記仍有些差距，但整體趨勢尚為一致。進一步觀察預估結果，停頓標記如有發生預估錯誤，可能造成韻律狀態的預估受連帶影響，例如圖(a)的{黨紀處分(*B3*)是應用}預估成{黨紀處分(*B2-2*)是應用}，使得預期的基頻重置現象消失了；另外，發現音長韻律狀態在*B4*的下一個音節會特別低，例如{處理方式(*B4*)但還}的“但”，推測是因為韻律狀態模型只考慮了前一個停頓標記，而音長韻律狀態在*B4*是大範圍的higt-to-low，才會造成此結果。



(a)



(b)

圖 4.2：(a)基頻，(b)音長韻律狀態預估結果。括號內為正確停頓標記，無括號為預估出來的停頓標記

4.3 語速相依之韻律參數產生法

SR-TTS所需的韻律參數包括有音節基頻軌跡、音節長度及停頓時長，此研究分別使用SR-HPM的音節韻律模型 $P(\mathbf{PS}|\mathbf{B},\ \mathbf{L})$ 和停頓聲學模型 $P(\mathbf{X},\mathbf{Y}|\mathbf{B},\mathbf{L})$ 預估單語速韻律參數，再藉由語速正規化參數產生不同語速之韻律參數。在語速相依的停頓時長方面，產生方法如下：

$$pd_n' = \mathrm{G}^{-1}(\mathrm{G}(pd_n^*, \alpha_g^{pd}, \beta_g^{pd}), \tilde{\alpha}^{pd}(SR_n), \tilde{\beta}^{pd}(SR_n)) \tag{4.3}$$

其中 α_g^{pd} 、 β_g^{pd} 、 $\tilde{\alpha}^{pd}(SR_n)$ 和 $\tilde{\beta}^{pd}(SR_n)$ 的求法如2.3.2節所介紹；單語速停頓時長 pd_n^* 的預估如式子4.4。

$$pd_n^* = \arg\max_{pd_n} p(pd_n \mid B_n^*, \mathbf{l}_n) \tag{4.4}$$

語速相依的音節基頻軌跡產生方法如下：

$$sp_n'(i) = \frac{sp_n^*(i) - \mu_g^{sp}(t_n,i)}{\sigma_g^{sp}(t_n,i)} \times \tilde{\sigma}^{sp}(SR_n,t_n,i) + \tilde{\mu}^{sp}(SR_n,t_n,i) \quad , \ i = 1 \sim 4 \tag{4.5}$$

其中 $\mu_g^{sp}(t_n,i)$ 、 $\sigma_g^{sp}(t_n,i)$ 、 $\tilde{\sigma}^{sp}(SR_n,t_n,i)$ 和 $\tilde{\mu}^{sp}(SR_n,t_n,i)$ 的求法如 2.3.3 節所介紹；單語速基頻軌跡 $\mathbf{sp}_n^*$ 的預估如 4.6 式，以預估之韻律標記和聲調語言參數，挑選對應的相關音節層次 APs，疊

加產生。

$$\mathbf{sp}_n^* = \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n^*} + \boldsymbol{\beta}^f_{B^*_{n-1},t^n_{n-1}} + \boldsymbol{\beta}^b_{B^*_n,t^{n+1}_n} + \boldsymbol{\mu}_{sp} \tag{4.6}$$

最後為語速相依的音節長度產生方法：

$$sd'_n = \left(sd_n^* - \mu_g^{sd}\right) / \sigma_g^{sd} \times \tilde{\sigma}^{sd}(SR_n) + \mu_k^{sd} \tag{4.7}$$

其中 σ_g^{sd} 、μ_g^{sd} 、μ_k^{sd} 和 $\tilde{\sigma}^{sd}$ 的求法如2.3.1節所介紹；單語速音節長度 sd_n^* 的預估方法與音節基頻軌跡 $\mathbf{sp}_n^*$ 類似，如式子4.7。

$$sd_n^* = \gamma_{t_n} + \gamma_{s_n} + \gamma_{q_n^*} + \mu_{sd} \tag{4.8}$$

圖4.3為一個*SR*=1.97的韻律產生範例，分有無給予正確停頓標記兩種預估比較。由圖可看出大部份音節韻律參數的預估結果尚為理想，少部份音節長度的嚴重錯誤發在*B3*和*B4*出現的下一個音節，原因推測如上一小節所討論之；停頓時長的嚴重錯誤是來自於不正確的停頓標記預估。

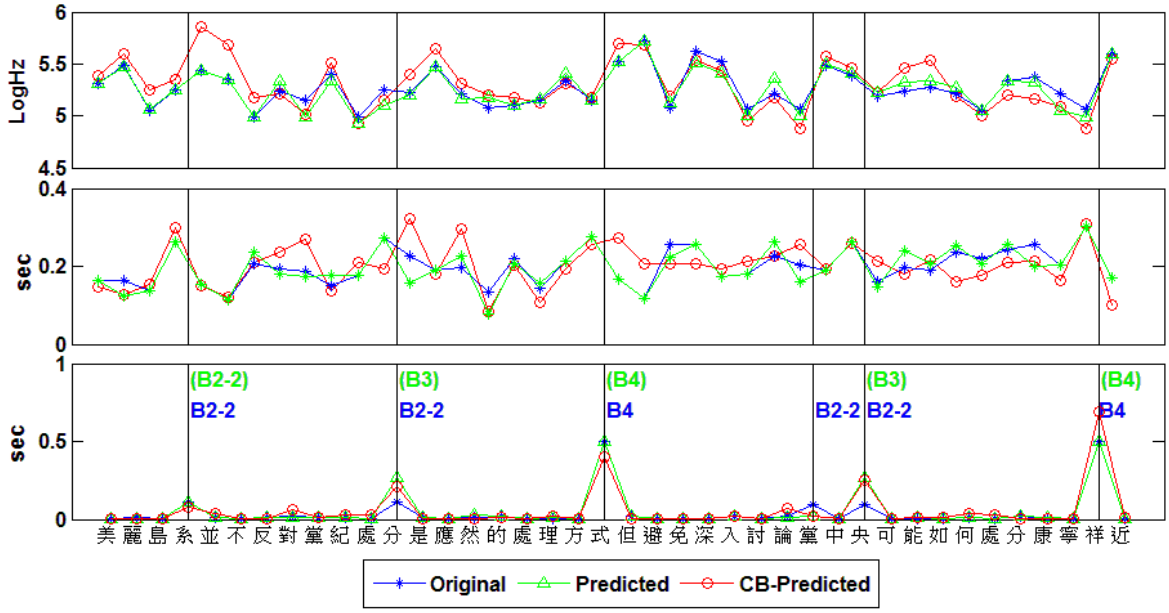


圖 4.3：韻律參數產生範列。由上至下分別為音節基頻軌跡平均、音節長度及停頓時長；括號內為正確停頓標記，無括號為預估出來的停頓標記

表 4.4 列出音節基頻軌跡、音節長度及停頓時長預估結果之 TREs，其中亦包含給定正確停頓標記所預估的結果。由表可得知後者的預估表現較佳，顯然停頓標記預估在韻律參數產生中扮演重要角色。因此，將來若要改進韻律參數的產生，停頓標記預估是必須且值得研究的議題。

表 4..4：韻律參數預估結果之 TREs

	With Predicted Break		With Correct Break	
	Training Set	Testing Set	Training Set	Testing Set
sp	53.0%	46.9%	44.6%	40.2%
sd	43.8%	40.9%	42.0%	39.8%
pd	18.3%	20.6%	6.7%	7.9%

4.4 HMM 頻譜模型

本研究使用HTS 2.1(HMM-based Speech Synthesis System, version 2.1)[23]建立頻譜模型，模型採用中文的聲母(22類)、韻母(40類)為基本音素單元，每個音素單元以五個狀態描述之；模型訓練時，加入韻律標記協助訓練，完整的文脈相關參數如表4.5，用此方式訓練出之模型稱為韻律相依HMM(PD-HMM)。此研究假設音素與語速無關，模型訓練僅採用SR-Treebank中的正常語速的部份，此部份的語料切割較為精準，共376個語句，包括51,868個音節。

表 4.5：HMM 頻譜模型之文脈相關資訊

level	ID	Description
Phonetic Feature	Pr_Phn	Previous initial/final
	-Cur_Phn	Current initial/final
	+Fol_Phn	Following initial/final
	^Phn_in_Syl	Initial/final position in a syllable

Prosodic Feature	p: Current LogF0 prosodic state tag q: Current duration prosodic state tag r: Current energy level prosodic state tag pb: Previous break type tag nb: Following break type tag

4.5 語音合成實驗結果與分析

本小節將ML為基礎的語速控制方法與我們所提出的方法進行主觀測試評估。ML為基礎的語速控制方法建立於傳統HTS合成系統，在給定語句總長度限制下，依ML準則決定每一音素的state duration；為了與我們的方法做比較，此HTS的訓練語速亦只採用正常語速的部份，模型的文脈相關資訊包括{Phone, POS, PM, Word length}。此實驗由兩種方法合成出三種語速(正常、慢、快)各10句，由15位受試者進行平均主觀值分數(Mean Opinion Scource, MOS)和偏好(Preference)兩種測驗。

(1) MOS測試結果

MOS的評分標準如表4.6所示，圖4.4顯示了測試結果，在正常語速的部份，SR-HPM based的語速控制方法略優於傳統ML based方法；快、慢語速方面，ML based方法造成聽覺自然度嚴重下降，而SR-HPM based方法依然有不錯的聽覺品質。

表 4.6：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

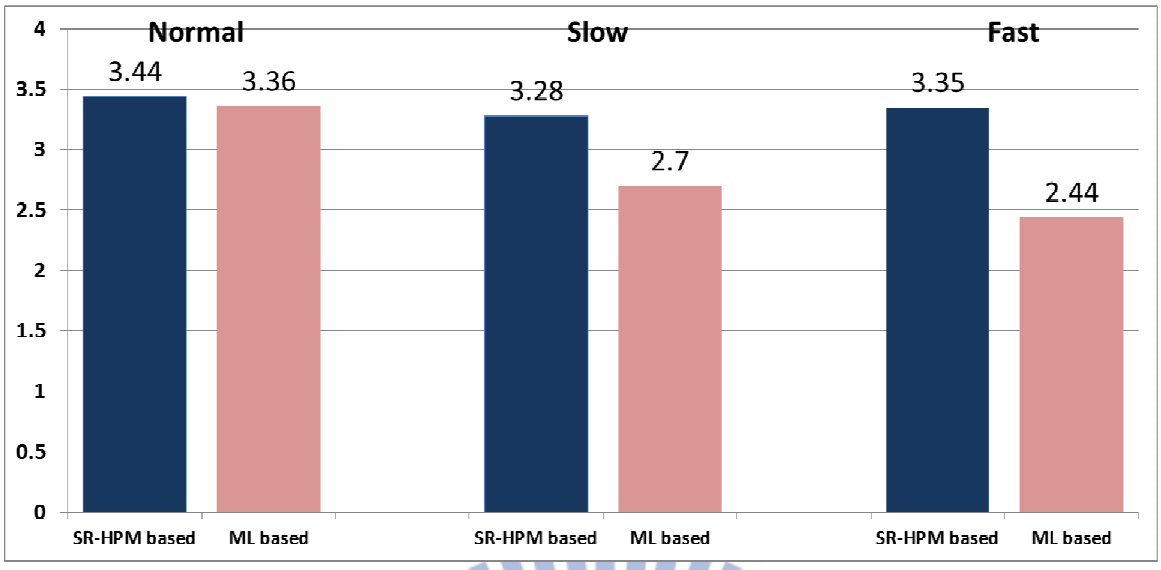


圖 4.4：MOS 測試結果

(2) Preference測試結果

將兩種方法的合成結果交由受試者選取較喜愛的一方。圖4.5顯示了測試結果，在正常語速的部份，兩種方法的喜愛比例是不相上下；而快、慢語速部份，SR-HPM based明顯勝於ML based方法，與MOS測試結果一致。

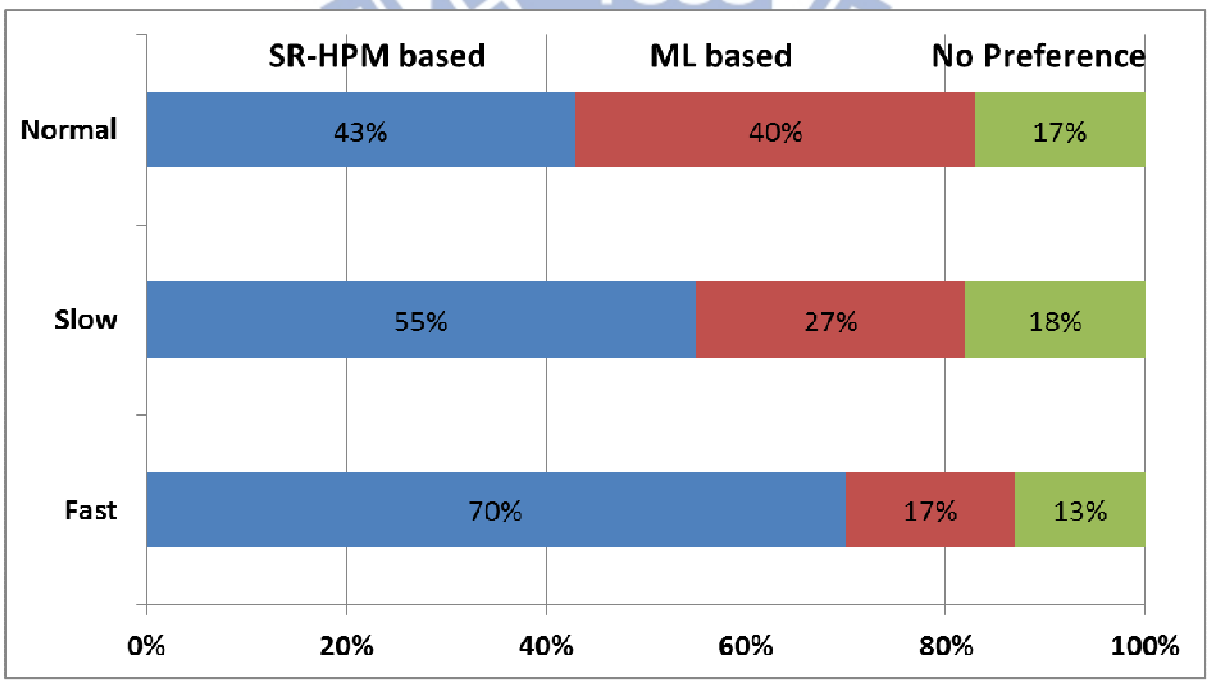


圖 4.5：Preference 測試結果

綜合以上的主觀測試結果，可得知SR-HPM based方法比起ML based方法更能掌握語速對於韻律全方面的影響，尤其是在停頓方面，SR-HPM考慮了不同語速的停頓出現時機及停頓呈度，如圖4.6範例，這是傳統方法很難做到的部份。

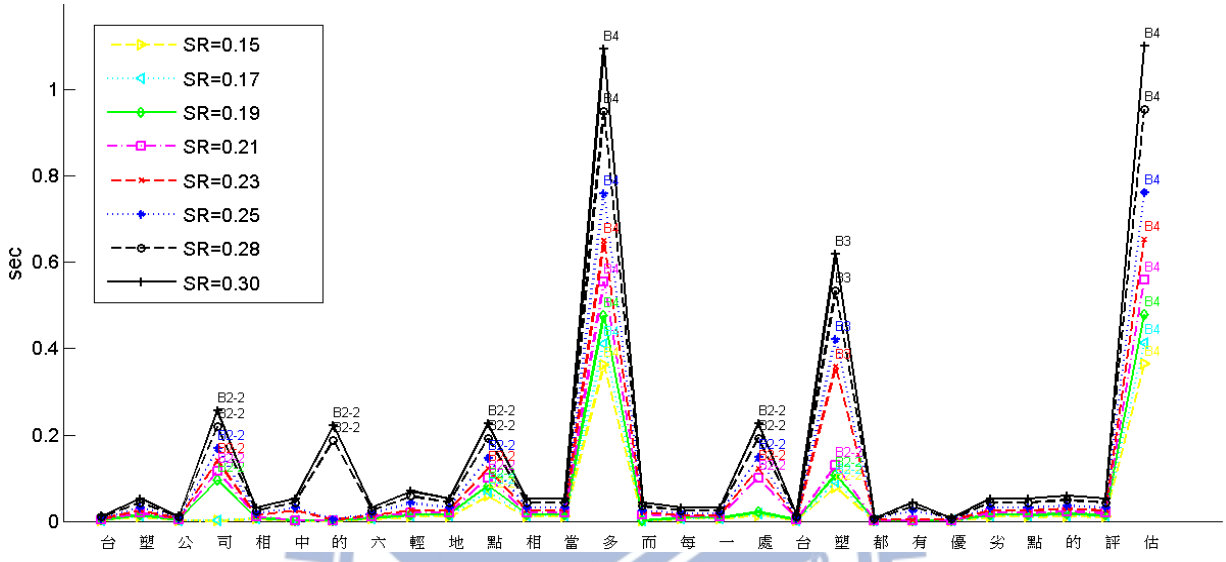


圖 4.6：不同語速的停頓預估結果

第五章 結論與未來展望

5.1 結論

本論文提出一新方法考慮漢語韻律中的語速效應，以修正型PLM演算法來建構語速相依之漢語階層式韻律模型，並自動完成韻律標記，其訓練出的模型參數和韻律標記符合我們所認知的語言學知識。經過研究與分析，本論文在語速對韻律之影響得出以下結論：(1)基頻軌跡形狀在漢語五種聲調中，各自受語速的影響程度不盡相同，但整體上慢語速的基頻軌跡形狀較為完整、dynamic range較大，快語速則反之；(2)音長與語速關係最為密切，為語速的一個重要測量值；另外，語速較慢時音長的變化會較大；(3)音節能量與語速無相關性，與發音環境較有關係；(4)停頓時長在停頓標記為*B2-2*、*B3*及*B4*時與語速呈非線性關係，在*B0*、*B1*、*B2-1*及*B2-3*則與語速相關性很小；(5)停頓發生機率方面，minor break在non-PM inter-word邊界是語速影響最明顯之處，其出現機率隨語速變慢而增加，而其它停頓類別受到之影響皆較輕微。

基於 SR-HPM，本研究提出語速相依之韻律參數產生方法，最後結合 HTS 來完成一個可控制語速之語音合成系統；由主觀測試結果可得知，本論文的語音合成方法，在正常、快、慢速合成語音皆有很好的聽覺品質。

5.2 未來展望

本論文所採用之語速測量值僅包含韻律中的音長部份，但實際語速的考慮因素還包括其它韻律參數，如停頓時長；另外，本論文以一段語句的平均音節長度來衡量該語句的語速，若語句太長或太短，則無法精確描述語速變化。未來希望能藉由 SR-HPM 來幫助語速之量測，引入停頓時長、基頻軌跡等等韻律參數，使得測量結果更為可靠。

在第四章所建構的 SR-TTS 系統中，並無考慮語速對音素之效應，HMM 的頻譜模型訓練僅採用正常語速的部份。實際上，語速應當會對音素造成影響，像是音素間的 coarticulation

程度或音素內 state duration 比例等等。因此，未來希望能加入其它速度的語料訓練之，並在文脈相關資訊中考慮語速之影響。

另外，在 SR-TTS 的停頓預估中，發現 *B2-2* 有過度預估(insertion)的現象，這會造成聽覺上不佳的感覺，此亦為未來待加強之部份。

參考文獻

[1] C. Heinrich and F. Schiel, “Estimating Speaking Rate by Means of Rhythmicity Parameters,” in Proc. INTERSPEECH-2011, Aug. 2011, pp. 1873-1876.

[2] A.-J. Li and Y. Zu, “Speaking Rate Effects on Discourse Prosody in Standard Chinese,” in Proc. Speech Prosody 2008, May 2008, pp. 449-452.

[3] C.-Y. Chiang, C.-C. Tang, H.-M. Yu, Y.-R. Wang and S. H. Chen, “An Investigation on the Mandarin Prosody of a Parallel Multi-Speaking Rate Speech Corpus,” in Proc. Oriental COCOSDA 2009, Aug. 2009, pp. 148-153.

[4] S.-A. Jun, “The effect of phrase length and speech rate on prosodic phrasing,” in Proc. of ICPhS, Barcelona, Spain, 2003, pp. 483-486.

[5] C.-Y. Tseng, “Corpus Phonetic Investigations of Discourse Prosody and Higher Level Information,” LANGUAGE AND LINGUISTICS, Insitute of Linguistics, Vol. 9,No. 3, 2008.

[6] T. Pfau, R. Faltlhauser and G. Ruske, “A Combination of Speaker Normalization and Speech Rate Normalization for Automatic Speech Recognition,” in Proc. ICSLP 2000, Oct. 2000, pp. 362-365.

[7] T. Shinozaki and S. Furui, “Hidden Mode HMM Using Bayesian Network for Modeling Speaking Rate Fluctuation,” in Proc. ASRU 2003, Nov. 2003, pp. 417-422.

[8] M. A. Siegler and R. M. Stem, “On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems,” In Proc. ICASSP'95, May 1995, pp. 612-615.

[9] J. Zheng, H. Franco and A. Stolcke, “Rate-of-Speech Modeling for Large Vocabulary Conversational Speech Recognition,” in Proc. ASRU 2000, Sept. 2002.

[10] T. Kato, M. Yamada, N. Nishizawa, K. Oura and K. Tokuda, “Large-scale Subjective Evaluations of Speech Rate Control Methods for HMM-based Speech Synthesizers,” in Proc. INTERSPEECH-2011, Aug. 2011, pp. 1845-1848.

[11] Y. Zu, A. Li and Y. Li, “Speech Rate Effects on Prosodic Features,” Report of Phonetic Research 2006, Institute of Linguistics, Chinese Academy of Social Sciences, pp. 141-144.

[12] K. Iwano, M. Yamada, T. Togawa, and S. Furui, “Speech-ratevariable HMM-based Japanese TTS system,” in Proc. TTS2002, Sept. 2002.

[13] T. Nishimoto, S. Sako, S. Sagayama, K. Ohshima, K. Oda and T. Watanabe, “Effect of Learning on Listening to Ultra-Fast Synthesized Speech,” in Proc. EMBC2006, Sept. 2006, pp. 5691-5694.

[14] M. Pucher, D. Schabus and J. Yamagishi, “Synthesis of fast speech with interpolation of adapted HSMMs and its evaluation by blind and sighted listeners,” in Proc. INTERSPEECH-2010, Sept. 2010, pp. 2186-2189.

[15] Chen-Yu Chiang, Sin-Horng Chen, Hsiu-Min and Yu, Yih-Ru Wang, “Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech,” J. Acoust. Soc. Am., vol. 125, No. 2, pp. 1164-1183, Feb, 2009.

[16] The HTK Book (for HTK version 3.4)

[17] WaveSurfer Homepage：www.speech.kth.se/wavesurfer/

[18] Z. Sheng, J.-H. Tao, and D.-L. Jiang, “Chinese prosodic phrasing with extended features,” Proceedings of the IEEE ICASSP 2003, Vol. 1, pp. 492–495.

[19] 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).

[20] Sin-Horng Chen, Jyh-Her Yang, Chen-Yu Chiang, Ming-Chieh Liu and Yih-Ru Wang, "A New Prosody-Assisted Mandarin ASR System", to be appeared in IEEE Trans. on Audio, Speech and Language Processing, vol. 20, no. 5, Jul. 2012.

[21] S.-H. Chen and Y.-R. Wang, “Vector quantization of pitch information in Mandarin speech,” IEEE Trans. Commun., vol. 38, no. 9, pp. 1317-1320, Sept. 1990.

[22] X. Shen and B. Xu, “A CART-based hierarchical stochastic model for prosodic phrasing in Chinese,” Proceedings of the ISCSLP 2000, pp. 105–109.

[23] 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/

[24] Y. Xu, “Contextual tonal variations in Mandarin,” J. Phonetics 25, 61-83, 1997.

附錄一

(1) *Th*1、*Th*2 和 *Th*3 的定義

*Th*1、*Th*2 和 *Th*3 是用來界定 B4、B3、B2-2 和 B0/B1 停頓時長的 threshold。由於 B3 和 B4 經常為標點符號的邊界，擁有較長的停頓時長，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 的停頓時長通常不明顯，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 non-PM, inter-word 邊界的停頓時長收集起來，同樣使用 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 non-PM, inter-word 過濾，並將不滿足此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 在停頓時長的特性差異不大，但 B2-1 定義為具有明顯基頻重置現象。將 intra-word 邊界的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 non-PM, 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 具明顯的基頻重置，再加上條件 $f_{\text{PM}}(\xi) > f_{\text{intra}}(\xi)$，滿足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 $f_{B0}(Pe)$，而 mean 較小的那群為 B1，ngery-dip 機率分佈為 $f_{B1}(Pe)$，則 *Th*6 即為此二高斯機率分佈 $f_{B0}(Pe)$ 和 $f_{B1}(Pe)$ 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 non-PM, nter-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 non-PM, 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。

附錄二

問題集Θ使用於停頓聲學模型 $p(pd_n, ed_n, pj_n, dl_n, df_n | B_n, \mathbf{l}_n)$ 和停頓語法模型 $P(B_n | \mathbf{l}_n)$ 之決策樹建立，如下：

1. Syllable Level

$Q_1$1.1: Is the initial of the following syllable a null one or in {*m, n, l, r*}?

$Q_1$1.2: Is the initial of the following syllable a null one?

$Q_1$1.3: Is the initial of the following syllable in {*b, d, g*}?

$Q_1$1.4: Is the initial of the following syllable in {*f, s, sh, shi, h*}?

$Q_1$1.5: Is the initial of the following syllable in {*m, n, l, r*}?

$Q_1$1.6: Is the initial of the following syllable in {*ts, ch, chi*}?

$Q_1$1.7: Is the initial of the following syllable in {*p, t, k*}?

$Q_1$1.8: Is the initial of the following syllable in {*tz, j, ji*}?

$Q_1$1.9: Is the inter-syllable location an inter-word?

$Q_1$1.10: Is the inter-syllable location a Type-1 intra-word?

$Q_1$1.11: Is the inter-syllable location a Type-2 intra-word?

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 = {period, exclamation mark, semicolon, question mark} and minor PMs={comma, dun hao(a mark in Chinese punctuation used to set off items in a series), colon}.

$Q_1$2.1: Does a PMs exist at the inter-syllable location?

$Q_1$2.2: Does a major PM exist at the inter-syllable location?

$Q_1$2.3: Does a minor PM exist at the inter-syllable location?

$Q_1$2.4 : Does a comma exist at the inter-syllable location?

$Q_1$2.5 : Does a dot or colon exist at the inter-syllable location?

2.2 Word length

$Q_2$2.2.1 ~ 4 : Is the preceding word an $n \in \{1, 2, 3, 4\}$-syllable word?

$Q_2$2.2.5 ~ 8 : Is the following word an $n \in \{1, 2, 3, 4\}$-syllable word?

$Q_2$2.2.9 : Is the length of the preceding word in syllable greater than 4?

$Q_2$2.2.10 : Is the length of the following word in syllable greater than 4?

2.3 Substantive/function words

$Q_2$2.3.1 ~ 2 : Is the preceding word a substantive word/function words?

$Q_2$2.3.3 ~ 4 : Is the following word a substantive word/function words?

2.4 Level-1 POS and special tags

$Q_2$2.4.1 ~ 11 : Is the POS of the preceding word A/C/D/N/I/P/T/V/DE/SHI/DM?

$Q_2$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_2$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_2$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_2$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_2$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_2$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_2$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}?

3. Questions related to sentence level features

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

3.1 Length of sentence

$Q_3$3.1.1~30: Is the length of the current sentence greater or equal to 1~30?

$Q_3$3.1.31~60: Is the length of the previous sentence greater or equal to 1~30?

$Q_3$3.1.61~90: Is the length of the following sentence greater or equal to 1~30?

3.2 Distances to PM

$Q_3$3.2.1~15: Is the distance to the nearest previous PM in syllable greater or equal to 1~15?

$Q_3$3.2.16~30: Is the distance to the nearest following PM in syllable greater or equal to 1~15?