國 立 交 通 大 學

電信工程研究所

碩 士 論 文

以韻律輔助之中文語音辨認系統之實現

An Implementation of Prosody-Assisted Mandarin Speech Recognition System

研究生：劉銘傑

指導教授：陳信宏 博士

中 華 民 國 一百 年 七 月

以韻律輔助之中文語音辨認系統之實現

An Implementation of Prosody-Assisted Mandarin Speech Recognition System

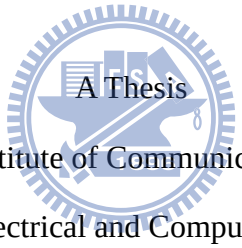
研 究 生：劉銘傑　　Student：Ming-Chieh Liu

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

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 2011

Hsinchu, Taiwan, Republic of China

中 華 民 國 一 百 年 七 月

以韻律輔助之中文語音辨認系統之實現

研 究 生：劉銘傑　　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

本研究提出一套新的整合韻律資訊於中文大辭彙連續語音辨認之方法。有別於以往只利用少數韻律資訊來幫助語音辨認，本研究利用先前已開發出的 PLM 演算法從大量未經人工標記的語料庫中自動產生訓練出 12 種韻律模型，並將其加入到 two-stage 自動語音辨認系統中，對系統中第一個 stage，也就是傳統 HMM 辨認器所產生的詞圖(word lattice)作重新評分的動作，如此可以得到更正確的詞辨認序列；此外，系統第二個 stage 還會同時解碼出更多資訊，包含詞性(POS)、詞後所接的標點符號(PM)以及用來建構測試語料之階層式韻律架構的兩種韻律標記。本研究實驗語料是利用包含朗讀式長句之 TCC300 語料庫，同時實驗中會引入一個 factored 語言模型，它是一個描繪詞、詞性及標點符號三者之間關係的模型，用以產生更好的 baseline 辨認效能。本研究在加入所有韻律資訊後之實驗結果對於詞(word)、字(character)、音節(syllable)的錯誤率分別為 20.1%、13.6%及 9.4%，與 baseline 結果比較起來則分別改善了 4.1%、4.0%及 2.4%的絕對錯誤率(16.9%、22.6%及 20.6%的相對錯誤率)。經由實驗結果分析，可以發現本系統能成功修正許多聲調及詞的錯誤辨認。

An Implementation of Prosody-Assisted Mandarin Speech Recognition System

Student：Ming-Chieh Liu　　Advisor：Dr. Sin-Horng Chen

Institute of Communication Engineering
National Chiao Tung University

Abstract

This thesis presents a new prosody-assisted ASR system for Mandarin speech. It differs from the conventional approach of using simple prosodic cues on employing a sophisticated prosody modeling approach to automatically generate 12 prosodic models from a large unlabeled speech database by the PLM algorithm proposed previously. By incorporating these 12 prosodic models into a two-stage ASR system to rescore the word lattice generated in the first stage by the conventional HMM recognizer, we can obtain a better recognized word string. Besides, some other information can also be decoded, including POS, PM, and two types of prosodic tags which can be used to construct the prosody hierarchical structure of the testing speech. Experimental results on the TCC300 database, which consists of long paragraphic utterances, showed that the proposed system significantly outperformed the baseline scheme using a factored LM to model word, POS, and PM. Performances of 20.1%, 13.6%, and 9.4% in word, character, and base-syllable error rates were obtained, which corresponds to 4.1%, 4.0%, and 2.4% absolute (16.9%, 22.6%, and 20.6% relative) error reductions. By error analysis, we found that many word segmentation errors and tone recognition errors were corrected.

致謝

由衷感謝陳信宏老師兩年來的教誨，在百忙之中仍然心繫著學生的研究，不時提點我在研究上的盲點。感謝王逸如老師教導我如何做一位真正的研究生，而不是交作業的大學生。感謝一路指導我和一起奮戰的智合學長、皓翔學長、超強的性獸學長、帥哥阿德學長、幽默的輝哥學長、跑去開公司的巴金叔、愛護地球的希群學長。感謝和我一起奮戰兩年的文良、啟全、豆腐、大胖、小蝦、智障、進竹、冠驛、佳緯以及帥氣優秀文武雙全的碩一學弟妹。最後感謝我的所有親人，祝大家心想事成。

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究動機

使用韻律訊息於語音辨認是近幾年來非常熱門的研究議題。所謂韻律就是指在連續語音當中具有跨越區段(suprasegmental)的特徵現象，像是重音表現、聲調、停頓、語調及節奏等；如果將韻律現象以物理特性表現出，通常會出現在語音中音高軌跡的變化、能量強度、語音長度及停頓當中。韻律與各層級的語言參數都有高度的相關性，從音素(phone)、音節(syllable)、詞(word)、片語(phrase)到句子(sentence)甚至是更高層次的語言參數，也因為有著重要的相關性，韻律資訊對於提升語音辨認的準確度是會有幫助的。一般來說，為了使用韻律資訊於語音辨認，我們會先找出語音中韻律聲學參數與語言參數之間的關係，並且將其關係特性訓練成韻律模型，最後將這些模型加入到語音辨認中，達到運用韻律資訊的效果。

1.2 文獻回顧

過去已經有許多關於使用韻律訊息於語音辨認的研究被發表[1]-[10]。Ananthakrishnan等人 [1] - [3]提出以加入韻律語言模型和韻律聲學模型對於傳統的基於HMM的語音辨認器所產生的N-best詞串或是word lattice作重新評分的動作，其中韻律聲學模型採用GMM/MLP來模式化(model)詞的二元音高重音(binary pitch accent)標記與韻律聲學參數(來自於語料中音高軌跡、能量及音長)之間的關係；語言韻律模型則是使用trigram語言模型並將其中每一個詞建立複合標記(compound tokens)以及二元音高重音標記。此外，由於兩種韻律模型是使用少量人工韻律標記語料所訓練而成，所以研究中使用一非監督式的方法來對兩種韻律模型進行調適[1]，用以解決因缺乏大量人工韻律標記的語料庫所造成的問題，研究結果對於在Boston University Radio News Corpus (BU-RNC)的詞錯誤率(WER)相對改善了1.2-3.1%。Chen等人[4]使用了兩種韻律資訊，分別為語調

短語邊界(intonational phrase boundary)和音高重音(pitch accent)，用以在語音辨認器中建立韻律相依(prosody-dependent)的詞及音素模型，研究結果對於在BU-RNC的WER相對改善了6.9%。Milone等人[5]提出將重音資訊加入到語音辨認，這種方法是利用語音信號中的音高及能量來建立一個詞的重音結構序列，完成後將其加入到語音辨認，其中音高及能量的取得是利用一個基於HMM的分類器或是類神經樹狀網路(neural tree networks)分類器；辨認系統中的語言模型之建立也會利用到片語(phrase)中的詞屬重音結構。研究結果對於在medium-vocabulary Spanish continuous-speech recognition task的WER相對改善了28.91%。Vergyri等人[6]提出整合多樣的韻律資訊語音辨認，研究中所使用的韻律模型包含詞長(word duration)模型、停頓(pause)語言模型以及一些隱藏事件（如句子邊界和語音不流暢性）的韻律模型，研究結果對於在Switchboard database的WER相對改善了2.6-3.1%。Ostendorf等人[7]提出了一種統計模型框架對於使用韻律資訊在語音辨認，其中幾個議題將在此進行討論，包括在不同時間刻度(time scale)中抽取韻律特徵參數和參數正規化、比較兩種建立韻律模型的方式，一種是使用一個intermediate symbol representation來作韻律模式(prosody modeling)；另一種是直接給定聲學相關(acoustic correlates)的條件下作韻律模式、如何設計問題集關於對韻律結構中的聲學模型作分類、在給定韻律聲學特徵之條件下如何建立動態發音模型。在文獻[8]中，將在中文語音辨認裡採用word-dependent聲調模式(tone modeling)，此方法所使用的韻律參數包含音節長度、三個F0 values並搭配兩種back-off策略，研究結果對於在Mandarin broadcast news ASR task的字元錯誤率(CER)有少量的改善。在文獻[9]中，利用韻律特徵(音高軌跡)參數和語言參數所建立的聲調模型於中文語音辨認，研究結果對於在Project-863 database的CER相對改善了3.65%。

除了上述關於使用韻律資訊幫助提升語音辨認的準確率外，還有許多跟韻律模式結合語音辨認的相關研究議題被提出，Liu等人[11]提出了一套豐富語音辨認作法，它能夠自動偵測出一般電話中對話及新聞廣播(NIST RT-04F)中的句子邊界及語音不流暢性。Shriberg等人[12]使用了決策樹的方法將語音中的節奏(rhythmic)及旋律(melodic)等特徵給模式化，並利用至多項研究議題，如句子切割、語音不流暢性偵測、廣播新聞中的主

題偵測及追蹤、口語對話中的語音辨認、對話行為標記等；雖然韻律模式對於以上這些研究議題有幫助，但對於提升詞辨認率而言仍屬微量。

從上述關於使用韻律資訊於語音辨認的相關研究中，我們將重點歸類在韻律模型的建立，圖1.1便是這些相關研究如何做韻律模式的流程圖，這些研究的共同點都是先找尋一些關鍵且重要的韻律資訊，然後在利用人工做好標記的少量語料來建立韻律模型，用以描述韻律資訊與不同階層的語言參數甚至是韻律聲學參數之間的關係，通常人工韻律標記是基於ToBI標記系統[13]之下。這樣作法的主要缺點是缺乏一個大量且優質標記的語料庫，因此只有少數明顯的韻律資訊被利用，像是音高重音(pitch accent)及語調片語邊界(intonational phrase boundary)等，如此一來，想要靠使用韻律資訊來提升語音辨認的準確性，其提升幅度便會大受影響。

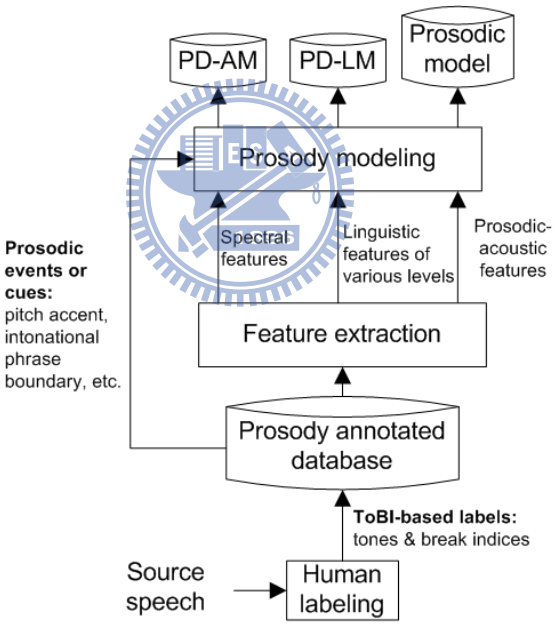


圖 1.1：傳統韻律模型之設計流程圖

1.3 研究方向

本研究將使用新的韻律模型，其產生方式如圖1.2所示，這也是延續了實驗室之前的研究[14]，關於使用未經人工標記的語料來做非監督式語音韻律標記及韻律模型建立。韻律模型是建立在四階層式韻律架構之下，並且使用韻律邊界停頓、音節韻律狀態這兩種韻律標記來表達這階層式的韻律架構，本研究中所使用的多種韻律模型就是描述這兩種韻律標記與語言參數及韻律聲學參數之間的關係。至於訓練模型的方法，是使用韻律標記及韻律模式(joint prosody labeling and modeling，簡稱PLM)演算法[14]從大量未經標記的語料中訓練各種韻律模型。也由於現在所使用的韻律模型內包含了更多、更完整的韻律資訊，若是將它與語音辨認結合，可以期望其效能會超越以往的相關研究；此外，本研究方法最終不只會解碼出詞(word)序列，同時會包含更多資訊，像是詞性(POS)、標點符號(PM)等語言參數序列及代表韻律架構的兩種韻律標記序列，屬於豐富語音辨認系統(enriched speech recognizer)。

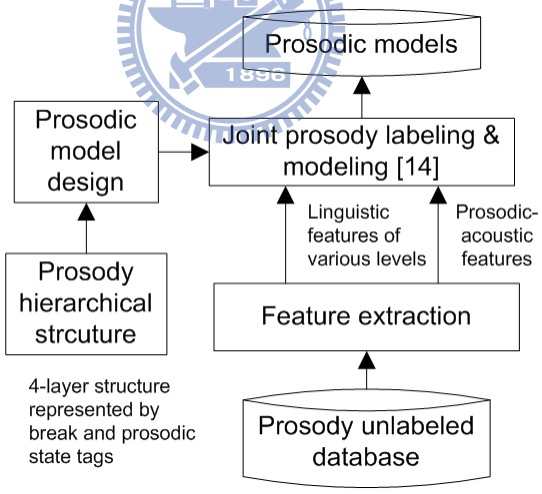


圖 1.2：本研究所使用之韻律模型設計流程圖

1.4 章節概要說明

本論文一共分為六章，其各章節內容分配如下：

第一章：緒論。

第二章：基本語音辨認系統。

第三章：階層式語音韻律模型

第四章：使用階層式語音韻律模型於中文大詞彙語音辨認系統

第五章：實驗結果及分析。

第六章：結論與未來展望。

第二章 基本語音辨認系統

本章是介紹本研究中所使用的基本語音辨認系統，辨認器中的包含聲學模型及語言模型，其中聲學模型是使用 TCC-300 語料庫建立，以隱藏式馬可夫模型呈現，用以描述發音過程的狀態轉移現象和輸出結果，並加入由一大量文字語料庫訓練出的語言模型來提升詞(word)辨認率。

2.1 TCC-300 基本辨認系統架構

下圖 2.1 就是基本語音辨認系統架構，從輸入音檔中抽取聲學特徵參數序列 $\mathbf{X}_a$，經辨認過程後輸出辨認詞串 $\mathbf{W}^*$，其基本原理數學式如下：

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} P(\mathbf{W} \mid \mathbf{X}_a) = \arg\max_{\mathbf{W}} P(\mathbf{W}) P(\mathbf{X}_a \mid \mathbf{W}) \quad (2.1)，$$

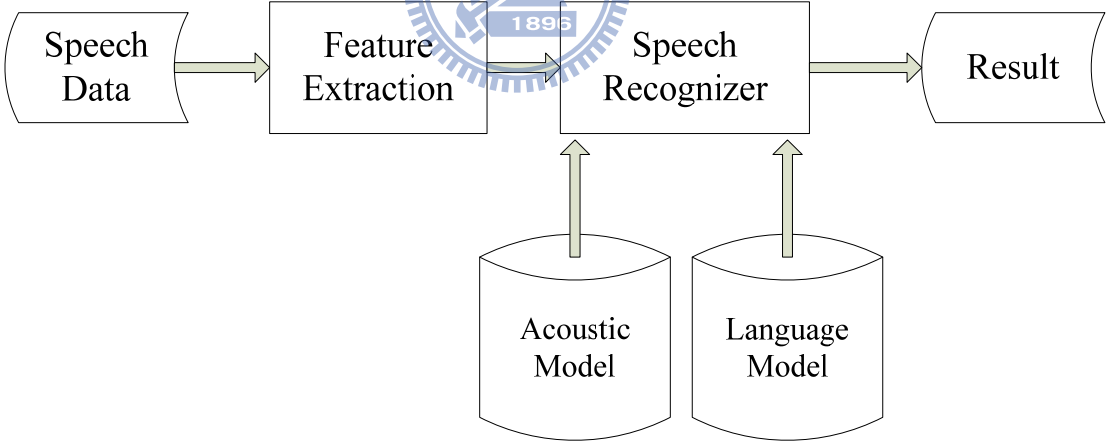


圖 2.1：基本語音辨認流程圖

(2.1)式中 $\mathbf{W}$ 代表詞，$\mathbf{X}_a$ 為聲學特徵參數，並經過最大事後機率法則將數學式化減成兩部分，其中 $P(\mathbf{W})$ 分數由語言模型(language model，簡稱 LM)計算得到；$P(\mathbf{X}_a \mid \mathbf{W})$ 分數則由聲學模型(acoustic model，簡稱 AM)計算得到。

2.1.1 語料庫簡介

本研究是使用 TCC-300 麥克風語音資料庫，它由國立台灣大學、國立成功大學及國立交通大學所共同錄製，此語料庫屬於麥克風朗讀語音，檔案統計資料如表 2.1 所示。每個學校之語句取樣頻率皆為 16000 赫茲（Hertz），取樣位元數為 16 位元。音檔檔頭為 4096 位元組（byte），副檔名為*.vat。

表 2.1：TCC-300 語料庫統計表

學校名稱	文章屬性	語者總數		總音節數		音檔總數	
台灣大學	短文	男	50	男	27541	男	3425
		女	50	女	24677	女	3084
		總數	100	總數	52218	總數	6590
交通大學	長文	男	50	男	75059	男	622
		女	50	女	73555	女	616
		總數	100	總數	148614	總數	1238
成功大學	長文	男	50	男	63127	男	588
		女	50	女	68749	女	582
		總數	100	總數	131876	總數	1170

依據上表 2.1，本研究所使用之語料庫共包含兩個群組，群組 A 一共包含 100 位語者，內容以短句為主；群組 B 一共包含 200 位語者，內容以長句為主。其中群組 A 的設計是用來平衡中文語音中的語音均衡(phonetic balance)；群組 B 的設計則會額外考慮到韻律學習(prosody study)。本研究會對上述語料庫分為訓練語料及測試語料，訓練語料的部分大約占 90%，共包含 274 位語者，長度一共約 23 小時；測試語料的部分大約占 10%，共包含 29 位語者，長度一共約 2.43 小時。然而，實際用來測試我們的辨認系統時，所使用的測試語料將全數包含在群組 B 的測試語料中，它包含 19 位語者共 226 個長句音檔，總長度約 2 小時，詞總數量為 14993，每個句子平均含有 117.2 個音節。

同時為了訓練本研究所使用的韻律模型，將從群組 B 的訓練語料中挑選 164 位語者的音檔共約 8.3 小時的語料來作訓練。

2.1.2 語言模型之架構及建立

2.1.2.1 語言模型架構

任何語言都有所屬之文法規則，利用文法規則所建立出的機率模型稱為語言模型。一般在建立語言模型時，是以詞(word)做為基本單位，現假設有一個句子共有 N 個詞，也就是「$w_1,w_2,\ldots,w_N$」，其中「w_i」代表句子中的第 i 個詞，則產生這個句子所對應的機率，可以拆解成一連串的條件機率之連乘：

$$\begin{aligned} & P(w_1, w_2, \ldots w_N) \\ & = P(w_1)P(w_2 \mid w_1)\cdots P(w_n \mid w_1, w_2, \ldots, w_{n-1}) \\ & = \prod_{i=1}^{N} P(w_i \mid w_1, w_2, \ldots, w_{i-1}) \end{aligned} \qquad (2.2)。$$

然而，要在有限的記憶體容量下求取所有詞的條件機率是難以達成的，所以我們利用 n-gram 的機率形式去近似(2.2)式，如下所示：

$$P(w_1, w_2, \ldots w_N) \cong \prod_{i=1}^{N} P(w_i \mid w_{i-n+1}, \ldots, w_{i-1}) \qquad (2.3)，$$

$$P(w_i \mid w_{i-n+1}, \ldots, w_{i-1}) = \frac{Count(w_{i-n+1}, \ldots, w_i)}{Count(w_{i-n+1}, \ldots, w_{i-1})} \qquad (2.4)，$$

在(2.3)式中每個 n-gram 的機率是在大量文章中詞串所累積的出現次數決定；而在(2.4)式中，$Count(\cdot)$ 表示詞串的出現次數，如果在分子項的 $Count(\cdot)$ 的值為 0 時，則此 n-gram 的機率會等於 0，在消息理論上來看機率 0 會使得資訊量無窮大，而造成錯誤的估計，此外一個詞串即使在訓練文字資料中沒有出現，並不代表在往後的辨認結果答案中不會出現，因此給定 0 的機率值並不合理，所以在此還必須對(2.4)式所計算出的機率做平滑化，使語言模型中所有的 n-gram 機率均能被良好的估計，一般常見的平滑方式如下所示：

$$P(w_i \mid w_{i-n+1},...,w_{i-1})$$

$$= \begin{cases} a(w_{i-n+1},...,w_{i-1})P(w_i \mid w_{i-n+2},...,w_{i-1}) & , Count(w_{i-n+1},...,w_i) = 0 \\ d_a \cdot \dfrac{Count(w_{i-n+1},...,w_i)}{Count(w_{i-n+1},...,w_{i-1})} & , 1 \le Count(w_{i-n+1},...,w_i) \le k \\ \dfrac{Count(w_{i-n+1},...,w_i)}{Count(w_{i-n+1},...,w_{i-1})} & , Count(w_{i-n+1},...,w_i) > k \end{cases} \quad (2.5)，$$

(2.5)式中 $a(w_{i-n+1},...,w_{i-1})$ 為經過正規化(normalization)的 back-off 係數，且需滿足下條件式：

$$\sum_{w \in V} P(w_i = w \mid w_{i-n+1},...,w_{i-1}) = 1 \quad (2.6)。$$

觀察(2.5)式，當計算 n-gram 機率所用的詞串出現次數為 0 時，利用其(n-1)-gram 的機率並乘上一個 back-off 係數，用以產生一個適當的機率值取代機率 0 的出現。另外如果 $Count(\cdot)$ 的值很小時，會造成計算出的 n-gram 機率不準確，這裡的解決方法是假設當一詞串的出現次數小於某特定次數時，將原始的 n-gram 機率乘上一個小於 1 的值 d_a (Discount Coefficient Factor)，它是依據 Good-Turning discounting 所計算出的，以降低其機率值，並將扣除的機率值分給詞串沒有出現的 n-gram 機率使用。

2.1.2.2 語言模型之建立

訓練語言模型必須具備大量的文字資料庫，本研究使用的文字資料庫共有三個來源：

來源一：光華雜誌(Sinorama)，其內容為一般雜誌的文章，蒐集的資料年代範圍介於 1976 年到 2000 年之間。

來源二：NTCIR，它是一個建立資訊檢索系統的標竿測試集，其內容由數種不同學科領域文章構成。

來源三：中研院平衡語料庫(Sinica)，它是一套由中研院錄製，內容包含多種主題，以語言分析研究為目的的資料庫。

本研究將以上三個來源語料庫簡稱 NSS，此外，對於建立一個完善的語言模型而言，另一項重要關鍵便是詞典的選擇，由於受限於記憶體大小，我們僅能將較常出現、

較重要的詞整理在詞典內提供建立訓練語言模型使用，下圖 2.2 為語言模型訓練流程圖。

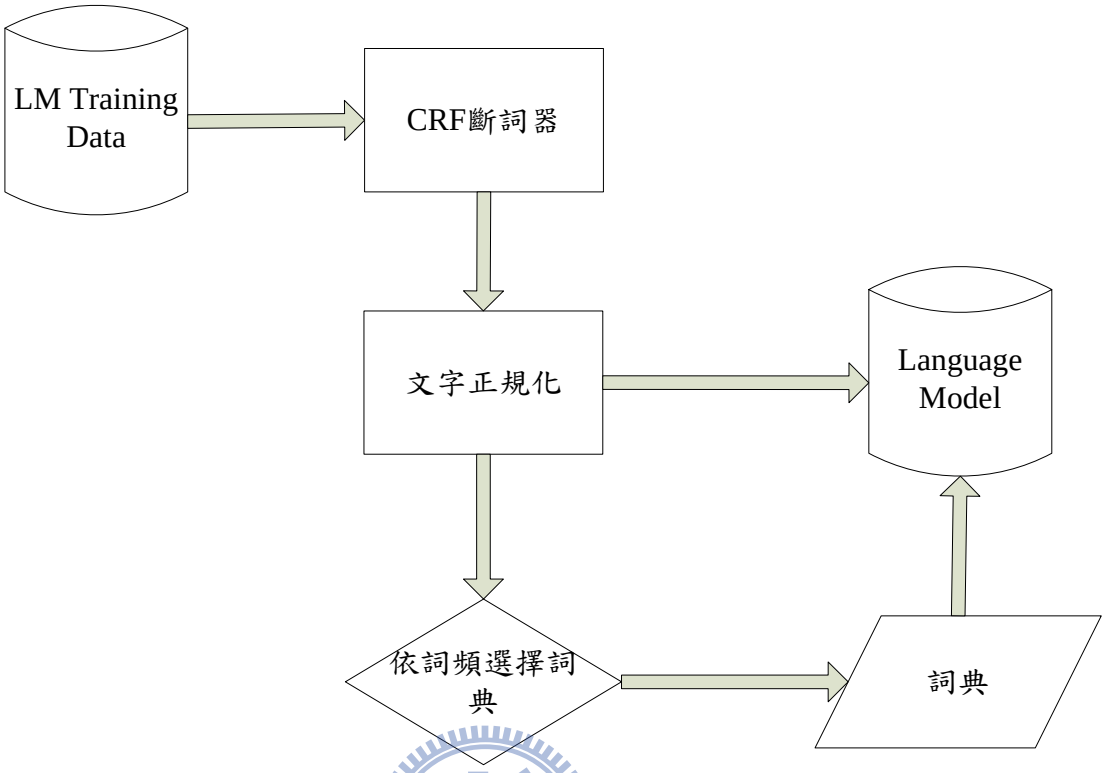


圖 2.2：語言模型訓練流程圖

如上圖 2.2，語言模型訓練流程中，NSS 在經過 CRF[30]斷詞器斷詞以及文字正規化[31]的處理後，得到詞的總數量為 122,541,303 個，字數為 231,225,705。至於本研究詞典的選擇方式是由斷詞結果中統計出各詞彙的詞頻，並依據詞頻大小來決定詞彙的重要性，這裡一共納入了 60,000 個常用詞彙，其平均詞長為 1.73 個字，最終便是利用這整理過的詞典來訓練語言模型。

2.1.3 聲學模型之建立

在語音辨認系統的訓練或測試流程中，除了上述建立語言模型外，另一項重要工作是聲學模型的建立，在建立前要先決定針對語音訊號抽取的參數類型，由於語音訊號之短時穩定特性及考慮到人耳聽覺效應的補償作用，本研究所使用的參數為 MFCC (Mel-Frequency Ceptral Coefficients，梅爾倒頻譜參數)，它的成分包含 12 維 MFCC 加上 1 維能量共 13 維，並取其 Delta term 和 Delta-Delta term，用以將參數變化訊息提供給辨認器使用，參數一共 39 維，系統相關設定如下表 2.7 所示。此外，基礎語音辨認器之聲

學模型為 411 個音節，每一個音節使用 8 個狀態(state)的隱藏式馬可夫模型(HMM)，並且使用 HTK[23]中之 MMI 鑑別性訓練得到。

表 2.2：MFCC 參數抽取設定檔

音框長度	32ms
音框平移	10ms
Filter bank 個數	24
取樣頻率	16kHz
Pre-emphasis Filter	First order with coefficient 0.97

2.2 基礎實驗結果

接著我們利用 2.1.3 節所建立的聲學模型來對測試音檔做辨認，藉以評估聲學模型的效能，所以本研究先建立一個 free-grammar 音節語言模型，搭配聲學系統做音節辨認率的計算，實驗結果如表 2.8 所示：

表 2.3：音節辨認率(free-grammar)

TCC-300 outside	73.39%
TCC-300 inside	85.74%

實驗數據顯示在 outside test 部分音節辨認率為 73.39%，inside test 可提升至 85.74%，接下來將加入 bigram 語言模型，同時產生詞辨認實驗結果及 word lattice 以觀察詞、字及音節之涵蓋率，然後將產生的 word lattice 利用 trigram 語言模型作展開後再重新評分以產生新的辨認結果，各項實驗結果如下頁各表格所示：

表 2.4：搭配語言模型之詞辨認率

Bigram LM	70.45%
Trigram LM	75.84%

表 2.5：搭配語言模型之字元辨認率

Bigram LM	78.58%
Trigram LM	82.14%

表 2.6：搭配語言模型之音節辨認率

Bigram LM	86.35%
Trigram LM	88.16%

表 2.7：bigram word lattice 之詞、字及音節涵蓋率

詞(word)	90.80%
字(character)	91.12%
音節(syllable)	93.24%

第三章 階層式語音韻律模型

當人類利用語音交談時，能夠影響到對方實質感受的因素很多，除了話語本身的語意之外，說話時音調的抑揚頓挫及音量的高低起伏等皆是，這些語音上的變化稱為韻律變化，其主要表現在語速(speaking rate)、停頓時長(pause duration)、音高軌跡(pitch contour)、音量大小(energy level)等因素上。本章第一節先介紹中文語音韻律階層式架構；第二節說明本研究使用的韻律模型。

3.1 中文語音韻律階層式架構

根據語言學家的研究結果發現[28]，語音的韻律結構會呈現出階層式的架構，而本研究所使用的韻律模型就是建構在這階層式的韻律架構之下，如下圖 3.1 所示，從最底層開始向上發展依序是：音節層次(syllable layer，SYL)、韻律詞層次(prosodic word layer，PW)、韻律短語層次(prosodic phrase layer，PPh)、呼吸組層次(breath group，BG)，以及韻律組句(prosodic phrase group)，這整體架構統稱「階層式多短語韻律句群(Hierarchical Prosodic Phrase Grouping，HPG)」架構[29]。

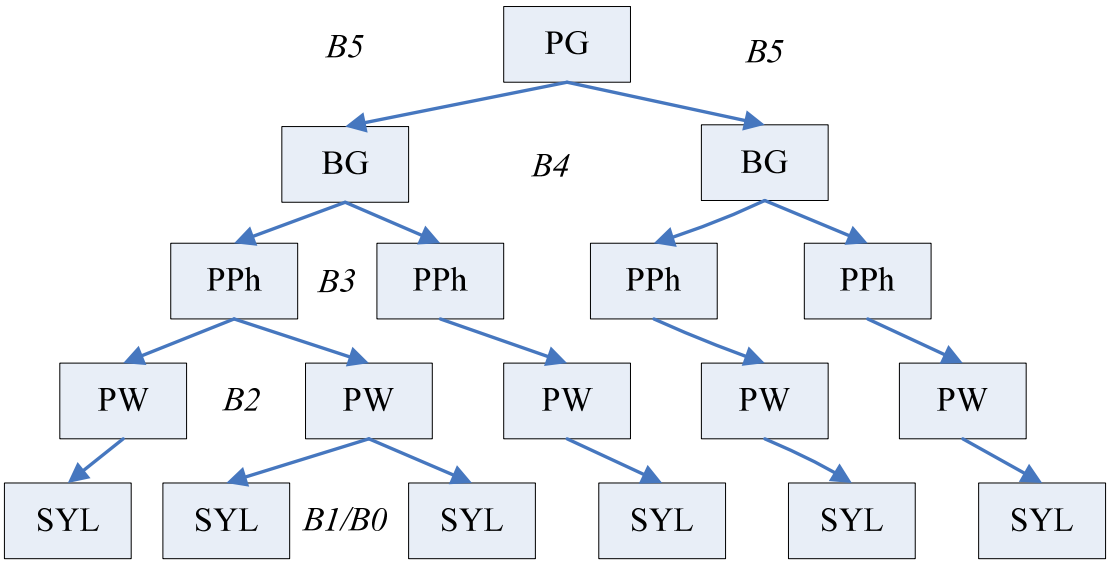


圖 3.1：中文語音韻律之階層式架構概念

這裡我們將使用兩種韻律標記來代表這階層式的韻律架構，第一種是韻律邊界停頓標記，它是用來區分階層式韻律架構中的各層韻律組成份子，如上圖 3.1 所示，其中 B0 和 B1 都是音節層次的邊界，差別在於 B0 表示的是 non-break of reduced syllabic boundary 或 tightly-coupling syllable juncture 而 B1 表示的是 non-break of normal syllabic boundary，通常在 B0 或 B1 的邊界不具有明顯停頓。而 B2 和 B3 分別代表韻律詞和韻律短語的邊界，B4 則代表了呼吸組的邊界，也由於是一個呼吸的停頓，和 B2、B3 比較起來會有個明顯的停頓現象，至於 B5 代表了韻律句組邊界，表示一個文章段落的結束，可以藉此觀察到句尾的音節長度拉長效應及能量減弱等現象。

由於本研究所使用的語料庫是大段落的語音，因此我們以 HPG 架構為基礎並對其作進一步修改，利用修改後的架構來產生本研究所需的韻律模型。首先將 B2 細分為 B2-1、B2-2、B2-3，其中 B2-1、B2-2、B2-3 分別表示含有明顯音高位置(pitch reset)之韻律詞邊界、具有短停頓(short pause)之韻律詞邊界，以及具有音節拉長效應(duration lengthening)後的韻律詞邊界。接下來將 B4、B5 合併為 B4，代表整個韻律階層式架構將從 5 層變回 4 層，如圖 3.2 所示。現在本研究將採用 7 種韻律邊界停頓(prosody break type) **B**={B0, B1, B2-1, B2-2, B2-3, B3, B4}來標記階層式架構中的四種韻律組成單元：音節(SYL)、韻律詞(PW)、韻律短語(PPh)及呼吸組/韻律句組(BG/PG)，用以區分韻律結構中每一層的韻律組成單元，對應如下表 3.1 所示：

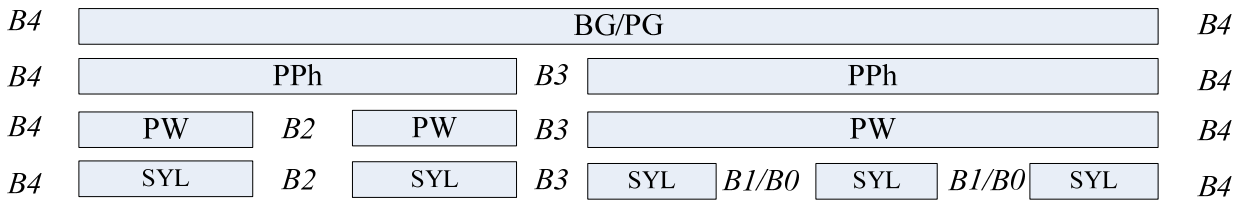


圖 3.2：本研究所採用的階層式韻律架構

表 3.1：韻律結構之停頓標記

韻律結構	停頓標記	意義
韻律群(PG)	B3	長停頓
或呼吸群(BG)	B4	長停頓且含有明顯的基頻跳躍
	B2-1	相鄰兩音節具有明顯的基頻跳躍
韻律詞(PW)	B2-2	短停頓
	B2-3	前一音節發生音節拉長
音節(SYL)	B0	音節邊界相鄰兩音節是緊密連接(tightly coupling)
	B1	音節邊界相鄰兩音節是普通連接(normal coupling)

至於另一種韻律標記是韻律狀態，它代表了每個音節的韻律聲學參數在韻律組成份子中的狀態，在本研究中我們會採用三種不同的韻律狀態，分別是量化正規化後的音節音高、音節音長及音節能量強度。

3.2 階層式韻律模型設計

基於上一章所介紹的的階層式韻律架構，我們便可以設計出多種韻律模型來描述兩種韻律標記、韻律架構中各層級的語言參數及韻律聲學參數之間的關係。在本研究中，要以能幫助於語音辨認的前提下來設計韻律模型，主要任務是在給定聲學參數 $\Lambda_a=\{\mathbf{X}_a,\mathbf{X}_p\}$ 的條件下，找出最佳的語言參數序列 $\Lambda_l=\{\mathbf{W},\mathbf{POS},\mathbf{PM}\}$、韻律標記 $\Lambda_p=\{\mathbf{B},\mathbf{P}\}$ 及 acoustic segmentation Υ_s，以數學的角度來看，就是要滿足如下式的MAP準則:

$$\begin{aligned}\Lambda_l^*,\Lambda_p^*,\Upsilon_s^* &= \arg\max_{\Lambda_l,\Lambda_p,\Upsilon_s} P(\mathbf{W},\mathbf{POS},\mathbf{PM},\mathbf{B},\mathbf{P},\Upsilon_s\mid\mathbf{X}_a,\mathbf{X}_p)\\ &= \arg\max_{\Lambda_l,\Lambda_p,\Upsilon_s} P(\mathbf{W},\mathbf{POS},\mathbf{PM},\mathbf{B},\mathbf{P},\Upsilon_s,\mathbf{X}_a,\mathbf{X}_p)\end{aligned} \quad (3.1)，$$

(3.1)式中 $\mathbf{W}=\{w_1^M\}$ 是代表詞序列；$\mathbf{POS}=\{pos_1^M\}$ 是詞(word)所對應到的詞性(part of speech)序列；至於 $\mathbf{PM}=\{pm_1^M\}$ 是代表標點符號序列；M代表詞的全部數量；$\mathbf{B}=\{B_1^N\}$ 則是韻律邊界停頓標記序列，它包含七種韻律邊界停頓標記：$B_n\in\{B0,\ B1,\ B2\text{-}1,\ B2\text{-}2,\ B2\text{-}3,\ B3,\ B4\}$；$\mathbf{P}=\{\mathbf{p},\mathbf{q},\mathbf{r}\}$ 則代表音節韻律狀態序列，它包含音節音高軌跡 $\mathbf{p}=\{p_1^N\}$、音節長度 $\mathbf{q}=\{q_1^N\}$ 及音節能量強度 $\mathbf{r}=\{r_1^N\}$；N代表音節的全部數量；$\mathbf{X}_a$ 代表一個 frame-based 頻

譜參數序列(i.e., MFCCs 及它們的一階和二階derivatives)； $\mathbf{X}_p = \{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$ 則是一個韻律聲學參數序列，其中**X**代表音節參數、**Y**代表音節邊界參數、**Z**代表音節間的differential參數；然而**X**中又包含了音節音高軌跡 (**sp**)、音節能量強度(**se**)及音節長度 (**sd**)；**Y**中則包含了音節間的停頓長度(**pd**) 及音節間的能量低點 (**ed**)；最後**Z**中包含了正規化的音節內基頻差 (**pj**) 及 兩個經正規化過的音節長度拉長因子 (**dl** and **df**)。在下列表3.2中將對所有在(3.1)式中有包含到的韻律標記、聲學參數及語言參數作一統整。

表 3.2：韻律標記、聲學參數以及語言參數之數學符號

T: prosodic tag	**B**: break type ={*B*0, *B*1, *B*2-1, *B*2-2, *B*2-3, *B*3, *B4* }	
	PS: prosodic state	**p**: pitch prosodic state
		q: duration prosodic state
		r: energy prosodic state
A: prosodic-acoustic 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
		df: normalized duration lengthening factor
L: linguistic feature	**l**: reduced linguistic feature set	
	t: syllable tone sequence	
	s: base-syllable type	
	f: final type	

在本研究中，我們會將(3.1)式作下列五種假設，以方便設計我們的韻律模型：

假設一：就如同傳統的聲學模型，頻譜參數序列 $\mathbf{X}_a$ 只會相依於詞序列 **W**。

假設二：韻律聲學參數序列 $\mathbf{X}_p$ 會相依於韻律標記序列 Λ_p 及語言參數序列 Λ_l。

假設三：音節韻律聲學參數序列 **X** 與音節邊界韻律參數序列 **Y** 及音節間的differential參數序列 **Z** 相互獨立。

假設四：韻律邊界停頓標記序列 **B** 相依於鄰近相關的語言參數序列 Λ_l。

假設五：音節韻律狀態序列 **P** 相依於鄰近的韻律邊界停頓標記 **B**。

經由以上五種假設後，(3.1)式將會簡化成以下形式：

$$\Lambda_l^*, \Lambda_p^*, \Upsilon_s^* \approx \arg\max_{\Lambda_l, \Lambda_p, \Upsilon_s} \{P(\mathbf{X}_a, \Upsilon_s|\mathbf{W})P(\mathbf{W}, \mathbf{POS}, \mathbf{PM}) \cdot P(\mathbf{B}|\Lambda_l)P(\mathbf{P}|\mathbf{B})P(\mathbf{X}|\Upsilon_s, \Lambda_p, \Lambda_l)P(\mathbf{Y}, \mathbf{Z}|\Upsilon_s, \Lambda_p, \Lambda_l)\} \quad (3.2),$$

(3.2)式中 $P(\mathbf{X}_a, \Upsilon_s|\mathbf{W})$ 代表聲學模型**(AM)**；$P(\mathbf{W}, \mathbf{POS}, \mathbf{PM})$ 則是joint syntax model，它描述了W、POS及PM之間的關係；$P(\mathbf{B}|\Lambda_l)$ 是代表break syntax model，它是利用語言參數 $\mathrm{L} = \{\mathbf{W}, \mathbf{POS}, \mathbf{PM}\}$ 來預估隱含著階層結構資訊的韻律邊界停頓 **B** 的模式，連結了階層式韻律與語言資訊之間的關係，$P(\mathbf{P}|\mathbf{B})$ 稱為韻律狀態模型，用來說明韻律狀態P的變化是如何受到韻律邊界停頓 **B** 的影響；$P(\mathbf{X}|\Upsilon_s, \Lambda_p, \Lambda_l)$ 稱為音節韻律模型，用來說明音節韻律參數受到B、P和L的影響而產生的變化；$P(\mathbf{Y}, \mathbf{Z}|\Upsilon_s, \Lambda_p, \Lambda_l)$ 稱為停頓聲學模型，用來說明在各個不同的韻律邊界停頓和語言參數之下，音節內的聲學特性。下圖3.3呈現出以上所提及的四種韻律模型搭配韻律標記、語言參數及韻律聲學參數之間的關係。

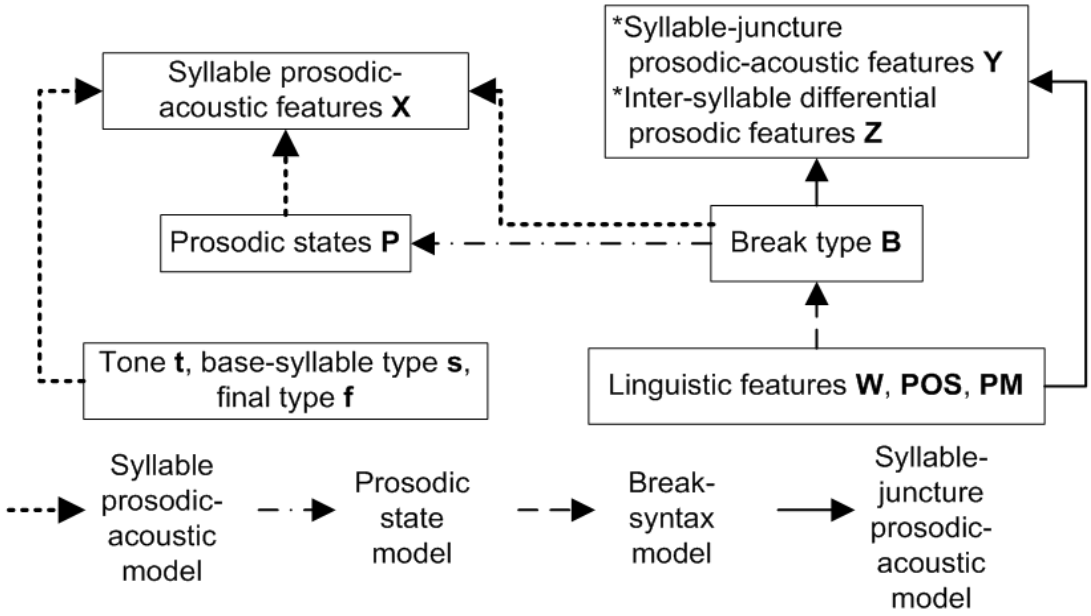


圖 3.3：四種韻律模型搭配韻律標記、語言參數及韻律聲學參數之間的關係

為了實際將韻律資訊運用於語音辨認中，以下我們將針對這四種韻律模型做更深入的探討：

3.2.1 Break Syntax Model

我們將 break syntax model $P(\mathbf{B}\,|\,\Lambda_l)$ 近似成下式：

$$P(\mathbf{B}\,|\,\Lambda_l) \approx \prod_{n=1}^{N-1} P(B_n\,|\,L_n) \qquad (3.3),$$

(3.3)式中 $P(B_n\,|\,L_n)$ 是一個描述音節韻律邊界停頓與其相關的語言參數之間關係的模型，同時 $P(B_n\,|\,L_n)$ 可經由分類樹與決策樹(CART)演算法堆導出來，其問題集請參見附錄一。

3.2.2 韻律狀態模型

韻律狀態模型 $P(\mathbf{P}\,|\,\mathbf{B})$ 可以進一步分解成三個子模型，如下所示：

$$P(\mathbf{P}\,|\,\mathbf{B}) = P(\mathbf{p}|\mathbf{B})P(\mathbf{q}|\mathbf{B})P(\mathbf{r}|\mathbf{B}) \approx P(p_1)P(q_1)P(r_1) \cdot \left[\prod_{n=2}^{N} P(p_n\,|\,p_{n-1},B_{n-1})P(q_n\,|\,q_{n-1},B_{n-1})P(r_n\,|\,r_{n-1},B_{n-1})\right] \qquad (3.4),$$

(3.4)式中 $P(p_n\,|\,p_{n-1},B_{n-1})$, $P(q_n\,|\,q_{n-1},B_{n-1})$, and $P(r_n\,|\,r_{n-1},B_{n-1})$ 分別表示各個不同韻律狀態，在給定音節邊界停頓 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。

3.2.3 音節韻律模型

音節韻律模型 $P(\mathbf{X}|\Upsilon_s,\Lambda_p,\Lambda_l)$ 可以進一步分解成三個子模型，如下所示：

$$\begin{aligned} & P(\mathbf{X}|\Upsilon_s,\Lambda_p,\Lambda_l) \\ & \approx P(\mathbf{sp}|\Upsilon_s,\mathbf{B},\mathbf{p},\mathbf{t})P(\mathbf{sd}|\Upsilon_s,\mathbf{B},\mathbf{q},\mathbf{t},\mathbf{s})P(\mathbf{se}|\Upsilon_s,\mathbf{B},\mathbf{r},\mathbf{t},\mathbf{f}) \\ & \approx \prod_{n=1}^{N} P(sp_n|p_n,B_{n-1}^{n},t_{n-1}^{n+1})P(sd_n\,|\,q_n,s_n,t_n)P(se_n\,|\,r_n,f_n,t_n) \end{aligned} \tag{3.5}$$

(3.5)式中 $P(sp_n\,|\,B_{n-1}^{n},p_n,t_{n-1}^{n+1})$、$P(sd_n\,|\,q_n,s_n,t_n)$ 及 $P(se_n\,|\,r_n,f_n,t_n)$ 三個子模型分別模擬音節音高軌跡序列 **sp**、音長序列 **sd** 漢音節能量序列 **se**，並且假設 **sp**、**sd** 和 **se** 的變化在此只受到以下幾個影響因素控制：音節聲調 **t**、基本音節類型 **s**、韻母類型 **f**、韻律狀態 **P={p,q,r}** 和韻律邊界停頓 **B**。在第一個子模型 $P(sp_n|B_{n-1}^{n},p_n,t_{n-1}^{n+1})$ 中，sp_n 會受到下列因素影響：目前音高韻律狀態 p_n、目前聲調 t_n 以及在給定韻律邊界停頓 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 $B_{n-1}^{n}=(B_{n-1},B_n)$，$t_{n-1}^{n+1}=(t_{n-1},t_n,t_{n+1})$。而 sp_n 則為第 n 個音節音高軌跡，是將音節音高軌跡進行正交展開，投影到四個 Legendre 多項式基底所得到的四維正交參數[17]，如下所示：

$$sp_n = sp_n^r + \beta_{t_n} + \beta_{p_n} + \beta_{B_{n-1},t_{n-1}^{n}}^{f} + \beta_{B_n,t_n^{n+1}}^{b} + \mu_{sp} \tag{3.6}$$

在(3.6)式中，β_{t_n} 及 β_{p_n} 則分別是目前音節音調 t_n 及目前音節韻律狀態 p_n 影響效應的 APs；$\beta_{B_{n-1},t_{n-1}^{n}}^{f}$ 及 $\beta_{B_n,t_n^{n+1}}^{b}$ 分別是第 n-1 個和第 n+1 個音節所貢獻的前後音節影響效應的 APs；μ_{sp} 是音高向量的總體平均值(global mean)；sp_n^r 是正規化後的 sp_n，亦可稱為 sp_n 扣除 β_{t_n}、β_{p_n}、$\beta_{B_{n-1},t_{n-1}^{n}}^{f}$、$\beta_{B_n,t_n^{n+1}}^{b}$ 和 μ_{sp} 的殘餘值。藉由假設 sp_n^r 是一 zero-mean 的 normal distribution，即 $N(sp_n^r;0,R_{sp})$，則 $P(sp_n|B_{n-1}^{n},p_n,t_{n-1}^{n+1})$ 可化解成下式：

$$\begin{aligned} & P(sp_n\,|\,p_n,B_{n-1}^{n},t_{n-1}^{n+1}) \\ & = N(sp_n;\beta_{t_n}+\beta_{p_n}+\beta_{B_{n-1},t_{n-1}^{n}}^{f}+\beta_{B_n,t_n^{n+1}}^{b}+\mu_{sp},R_{sp}) \end{aligned} \tag{3.7}$$

在建構第二個子模型 $P(sd_n\,|\,q_n,s_n,t_n)$ 時，我們考慮了三個獨立的影響效應，分別是音節韻律狀態、音節本身及音節聲調，因此我們可以將觀察到的音節長度 sd_n 表示成：

$$sd_n = sd_n^r + \gamma_{t_n} + \gamma_{s_n} + \gamma_{q_n} + \mu_{sd} \tag{3.8}$$

在(3.8)式中 sd_n^r 是正規化後的 sd_n；γ_{t_n}、γ_{s_n} 及 γ_{q_n} 分別是目前音節音調、目前音節本身及目前音節韻律狀態影響效應的 APs；μ_{sd} 是音節音長的總體平均值(global mean)。同樣地，藉由假設 sd_n^r 是一 zero-mean 的 normal distribution，即 $N(sd_n^r;0,R_{sd})$，則 $P(sd_n \mid q_n,s_n,t_n)$ 可化解成下式：

$$P(sd_n \mid q_n,s_n,t_n) = N(sd_n;\gamma_{t_n}+\gamma_{s_n}+\gamma_{q_n}+\mu_{sd},R_{sd}) \quad (3.9)。$$

最後在建構第三個子模型 $P(se_n \mid r_n, f_n,t_n)$ 時，我們同樣考慮了三個獨立的影響效應，分別是音節韻律狀態、音節韻母及音節聲調，因此我們可以將觀察到的音節能量 se_n 表示成：

$$se_n = se_n^r + \omega_{t_n} + \omega_{f_n} + \omega_{r_n} + \mu_{se} \quad (3.10)，$$

在(3.10)式中 se_n^r 是正規化後的 se_n；ω_{t_n}、ω_{s_n} 及 ω_{q_n} 分別是目前音節音調、目前音節韻母及目前音節韻律狀態影響效應的 APs；μ_{se} 是音節能量的(global mean)。同樣地，藉由假設 se_n^r 是一 zero-mean 的 normal distribution，即 $N(se_n^r;0,R_{se})$，則 $P(se_n \mid r_n, f_n,t_n)$ 可化解成下式：

$$P(se_n \mid r_n, f_n,t_n) = N(se_n;\omega_{t_n}+\omega_{f_n}+\omega_{r_n}+\mu_{se},R_{se}) \quad (3.11)。$$

3.2.4 停頓聲學模型

我們將停頓聲學模型進一步化簡，並得到五個子模型，如下所示：

$$\begin{aligned} & P(\mathbf{Y},\mathbf{Z}|\Upsilon_s,\Lambda_p,\Lambda_l) \\ \approx\ & P(\mathbf{pd},\mathbf{ed},\mathbf{pj},\mathbf{dl},\mathbf{df}|\Upsilon_s,\Lambda_p,\Lambda_l) \\ \approx\ & \prod_{n=1}^{N-1}\Big\{g(pd_n;\alpha_{B_n,\Lambda_{l,n}},\beta_{B_n,\Lambda_{l,n}})N(ed_n;\mu_{ed,B_n,\Lambda_{l,n}},\sigma^2_{ed,B_n,\Lambda_{l,n}}) \\ & \cdot N(pj_n;\mu_{pj,B_n,\Lambda_{l,n}},\sigma^2_{pj,B_n,\Lambda_{l,n}})N(dl_n;\mu_{dl,B_n,\Lambda_{l,n}},\sigma^2_{dl,B_n,\Lambda_{l,n}}) \\ & \cdot N(df_n;\mu_{df,B_n,\Lambda_{l,n}},\sigma^2_{df,B_n,\Lambda_{l,n}})\Big\} \end{aligned} \quad (3.12)，$$

在(3.12)式中我們將音節邊界的停頓時長(pause duration) pd_n 以 gamma distribution 來 fit；ed_n 代表了音節間的能量低點(energy dip)，這裡我們則用 normal distribution 來 fit；此外，正規化的音節內基頻差序列(pitch jump)，定義為：

$$pj_n = (sp_{n+1}(1)-\beta_{t_{n+1}}(1))-(sp_n(1)-\beta_{t_n}(1)) \quad (3.13)，$$

在(3.13)式中，$sp_n(1)$ 定義為第一維度的音節音高軌跡；$\beta_{t_n}(1)$ 則定義為第一維度的聲調影響因素。同樣的，我們將正規化的音節內基頻差用 normal distribution 來 fit。最後，還有兩種正規化的音節長度拉長因子 ***dl*** 和 ***df*** 定義為：

$$dl_n = (sd_n - \gamma_{t_n} - \gamma_{s_n}) - (sd_{n-1} - \gamma_{t_{n-1}} - \gamma_{s_{n-1}}) \quad (3.14)，$$

$$df_n = (sd_n - \gamma_{t_n} - \gamma_{s_n}) - (sd_{n+1} - \gamma_{t_{n+1}} - \gamma_{s_{n+1}}) \quad (3.15)，$$

這裡兩種因子我們都用 normal distribution 來 fit。在實作過程中，$P(\mathbf{pd},\mathbf{ed},\mathbf{pj},\mathbf{dl},\mathbf{df}|\Upsilon_s,\Lambda_p,\Lambda_l)$ 是經由分類數與決策樹(CART)推導出來，其節點的分類標準是依據 maximun likelihood gain，CART 演算法可以利用一個已經設計好的問題集(附錄一)，依據不同的韻律邊界停頓同時將所有音節的 pd_n 、 ed_n 、 pj_n 、 dl_n 和 df_n 做好分類。

有了上述四種韻律模型後，最後將使用韻律標記及韻律模式(PLM)演算法[14]從一群未經標記的語料中來訓練出這些韻律模型；PLM 演算法能同時估測韻律模型的參數及對所有語句作韻律標記，過程中先是經過一初始化程序，對所有語句作初始的韻律標記及韻律模型的參數估計，然後在根據ML法則做一連串的最佳化程序、反覆更新所有韻律標記和韻律模型的參數，直到收斂為止。

第四章 使用階層式語音韻律模型於中文大詞彙語音辨認系統

本章第一節會說明如何將這些韻律模型以 two-stage 的方式加入到語音辨認中，這裡我們會先利用傳統 HMM-based 語音辨認器來做 first-stage 語音辨認，待辨認完成後產生辨認結果及 word lattice ，本研究之作法就是在第二個 stage 中利用上述章節所介紹過的四種共 12 個韻律模型及一個 factored 語言模型來對 word lattice 作重新評分的動作，得到新的辨認結果並觀察。第二節說明本論文使用鑑別式模型組合(Discriminative Model Combination)來解決在重新評分過程中針對多個模型之權重問題。

4.1 使用韻律訊息於 two-stage 語音辨認系統

圖 4.1 就是本研究之系統流程圖，以下將針對系統第二個 stage 作詳細介紹。

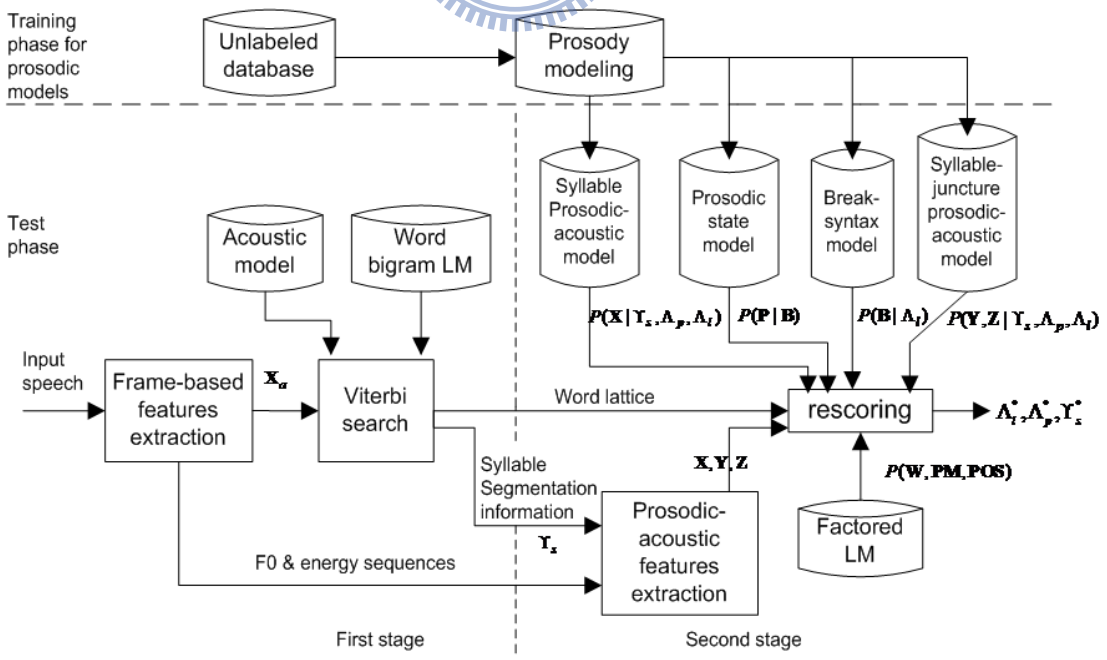


圖 4.1：以 two-stage 方式之韻律輔助中文語音辨認系統流程圖

4.1.1 Joint Syntax Model之架構

由於本研究中會使用到 POS 及 PM 等語言參數資訊來訓練我們的韻律模型，所以針對辨認系統第二個 stage 中語言模型的部份，我們將原本傳統的語言模型(bigram or trigram LM)取代成一個 factored 語言模型(FLM)，用以描述詞序列 **W**、詞性序列 **POS** 及標點符號序列 **PM** 之間的關係。我們最終將這個 factored 語言模型拆解成一個 word trigram 模型、一個 factored POS 模型及一個 factored PM 模型，其數學公式如下所示：

$$
\begin{aligned}
& P(\mathbf{W},\mathbf{PM},\mathbf{POS}) \\
& \approx \prod_{i=1}^{M}\left\{\underbrace{P(W_i \mid W_{i-2}^{i-1})}_{\text{word-trigram LM}}\cdot\underbrace{P(POS_i \mid POS_{i-1},W_i)}_{\text{factored POS model}}\underbrace{P(PM_{i-1} \mid POS_{i-1}^{i},W_{i-1})}_{\text{factored PM model}}\right\} \\
& \approx P(W_1)P(POS_1 \mid W_1)P(W_2 \mid W_1)P(POS_2 \mid POS_1,W_2)P(PM_1 \mid POS_1^2,W_1) \\
& \quad \cdot\prod_{i=3}^{M}\left\{P(W_i \mid W_{i-2}^{i-1})P(POS_i \mid POS_{i-1},W_i)P(PM_{i-1} \mid POS_{i-1}^{i},W_{i-1})\right\}
\end{aligned}
\tag{4.1}
$$

。

這裡我們是使用FLM approach [18]來建構(4.1)式中的兩個factored語言模型(POS及PM)，並使用SRILM toolkit [19]及利用Witten-Bell smoothing的方式來訓練(4.1)式中的三個模型，其中FLM approach的最主要的概念是利用其他相關資訊(factor)的輔助來預估目標，所以這裡將充分利用語言知識來提升預估POS或PM的準確性。當然，若使用多種資訊作預估時恐會面臨到資料量不足的問題，因此FLM會採取backoff的架構來應對，其FLM的Generalized backoff的數學式如 (4.2)式：

$$
P_{GBO}(f \mid f_1,f_2,f_3)=\begin{cases} d_N(f,f_1,f_2,f_3)P_{ML}(f \mid f_1,f_2,f_3) & \text{if } N(f,f_1,f_2,f_3)>\tau \\ \alpha(f_1,f_2,f_3)g(f,f_1,f_2,f_3) & \text{otherwise} \end{cases}
\tag{4.2}
$$

，

(4.2)式中各項份子代表意義如下：

1. $P_{ML}(f \mid f_1,f_2,f_3)=\dfrac{N(f,f_1,f_2,f_3)}{N(f_1,f_2,f_3)}$ ，它是代表 maximum likelihood distribution。

2. $g(f,f_1,f_2,f_3)$ 代表 backoff distribution。

3. $N(f,f_1,f_2,f_3)$ 代表 f,f_1,f_2,f_3 這樣的組合出現在訓練語料的次數，當它的數值大於門

檻值 τ 時，$P_{GBO}(f \mid f_1, f_2, f_3) = d_N(f, f_1, f_2, f_3) P_{ML}(f \mid f_1, f_2, f_3)$。

4. $d_N(f, f_1, f_2, f_3)$ 代表 discount，它是一個介於 0 到 1 的值，其作用會轉移部分 $P_{ML}(f \mid f_1, f_2, f_3)$ 的機率值給 backoff distribution $g(f, f_1, f_2, f_3)$ 做平滑化(smoothing)。

5. $\alpha(f_1, f_2, f_3)$ 代表 backoff weight，它是為了確保 $\sum_f P_{GBO}(f \mid f_1, f_2, f_3) = 1$，經由推導可得下式：

$$\alpha(f_1, f_2, f_3) = \frac{1 - \sum_{f:N(f, f_1, f_2, f_3) > \tau} d_{N(f, f_1, f_2, f_3)} P_{ML}(f \mid f_1, f_2, f_3)}{\sum_{f:N(f, f_1, f_2, f_3) <= \tau} g(f, f_1, f_2, f_3)} \quad (4.3)。$$

在本研究裡，factored POS model的backoff路徑結構如下圖4.2所示，在最上層的情況，期望以目前的詞 W_i 及前一個POS等語言factors來預估 POS_i，若此機率的組合沒有出現，則丟棄一個factor POS_{i-1}，若仍是沒有出現的話，就退化到最下層的狀態，此時就一定有此機率；

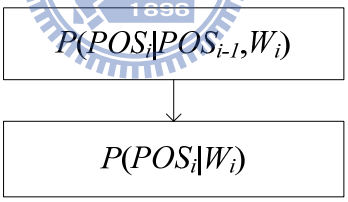


圖 4.2：factored POS model 的 backoff 路徑

factored PM model亦是如此，其backoff路徑的架構如圖4.3。我們使用了前一個word、前一個POS及目前POS的資訊，來預估前一個PM的機率為何，依照圖4.3設定，我們首先丟掉 POS_i，接著是 POS_{i-1}，然後 W_{i-1}，最終退化到 $P(PM_{i-1})$。

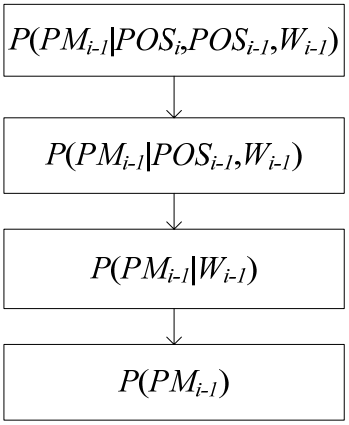


圖 4.3：factored PM model 的 backoff 路徑

4.1.2 第二級辨認器之實作

如上圖 4.1 所示，經由辨認器第一個 stage 流程，會產生 bigram word lattice，接下來我們利用 HTK[23]中的指令及搭配 trigram 語言模型，將 bigram word lattice 展開成 trigram word lattice 輸出，由於本研究未來要加入的韻律資訊是 syllable level，所以在提供 word lattice 給第二個 stage 使用前，要先將 word lattice 中標示出音節切割位置。

至於從第二個 stage 開始，要加入的韻律模型及 joint syntax model 之種類高達 16 種，為了觀察每個模型分別所貢獻的影響力，本研究將分成三階段逐次加入模型資訊並觀察實驗結果，詳細流程圖如下所示：

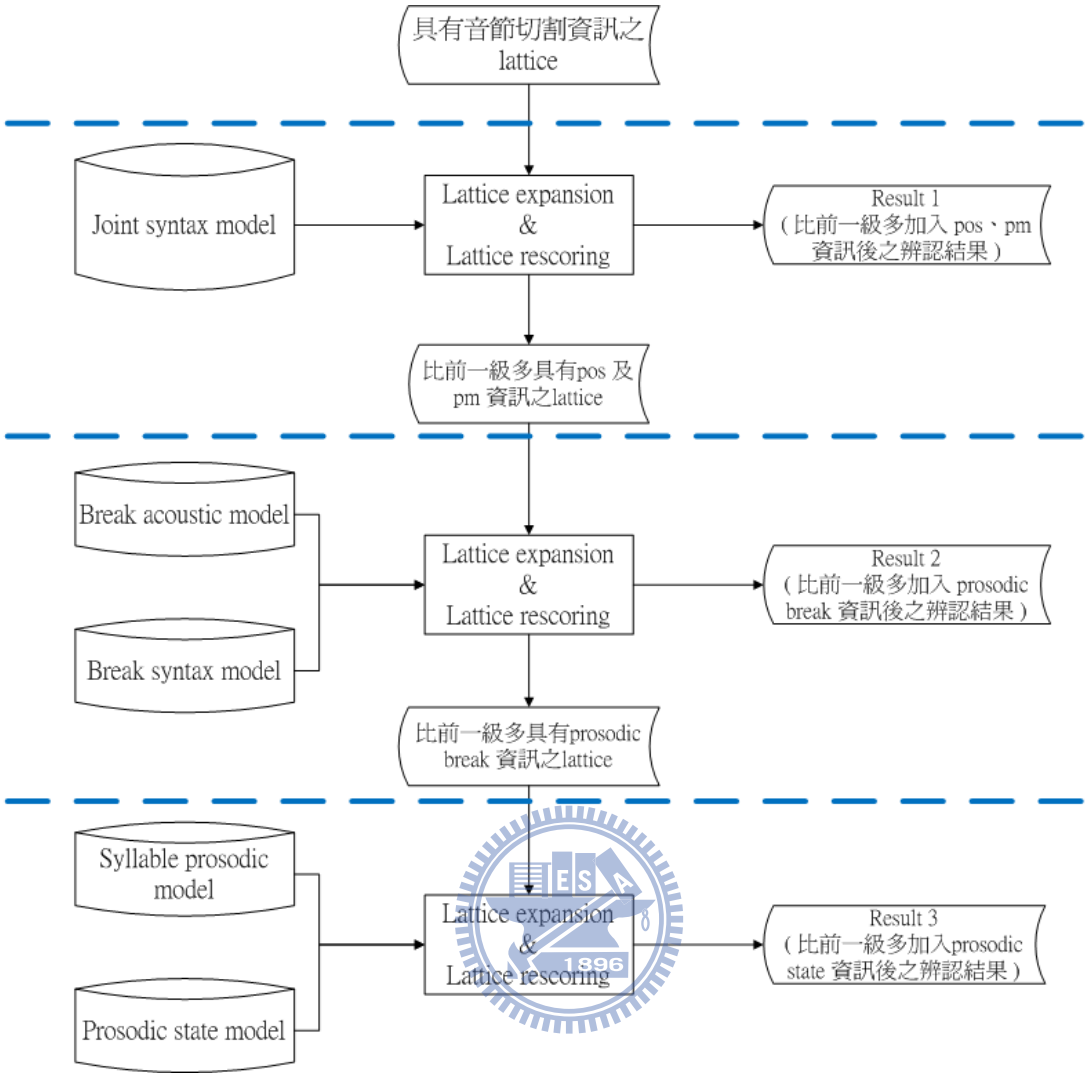


圖 4.4：辨認器第二級三階段實作流程圖

如上圖 4.4 所示，辨認器第二級第一個階段是引入更多語言參數資訊(POS 及 PM)，需要加入的模型為 joint syntax model，未來在加入韻律資訊時除了需要更多語言資訊外，整套系統最終是希望能解碼出多種語言參數，如詞(word)、詞性(POS)及標點符號(PM)；第二個階段我們主要是引入韻律邊界停頓的資訊，要加入的韻律模型分別是 break syntax model 及停頓聲學模型，在這一階段最後除了會解碼出多種語言參數外，同時會將每個詞中每個音節後所接的韻律邊界停頓一併解碼出；最後一個階段是引入音節韻律狀態資訊，要加入的韻律模型分別是音節韻律模型及韻律狀態模型，在這一階段最後除了會解碼出如同上一階段的語言參數及韻律邊界停頓外，每個詞中每個音節所屬的三種韻律狀態也將一併解碼出。

4.1.2.1 參數正規化

針對圖 4.4 中第二級辨認器第二個階段開始，工作過程中如有牽涉到音節能量、音節長度或音節音高等參數資訊時，必須要先經過正規化的步驟，這是為了要克服不同語者先天發音上的差異。首先我們利用第一級系統所產生的最佳(top 1)辨認結果來做正規化，利用此辨認結果及 HTK 中之指令對抽取好的測試音檔特徵參數作 force alignment，接下來就是計算各測試音檔的音節能量、音節長度或音節基頻等統計平均值，完成後方可執行正規化，如下所示：

經過高斯正規化後第 s 個音檔的第 i 個音框之對數基頻數值 f_i 如下：

$$\hat{f}_i^s = \left(\frac{f_i^s - \mu_s}{\sigma_s} \right) \cdot \sigma_{average} + \mu_{global} \quad (4.4)，$$

其中 μ_s 及 σ_s 分別為第 s 個音檔基頻數值之平均值及標準差，其數學式分別如下：

$$\mu_s = \frac{\sum_{i=1}^{I(s)} f_i^s}{I(s)} \quad (4.5)，$$

$$\sigma_s = \sqrt{\frac{\sum_{i=1}^{I(s)} (f_i^s - \mu_s)^2}{I(s)}} \quad (4.6)，$$

在(4.5)式及(4.6)式中 $I(s)$ 為第 s 個音檔基頻之總音框數，而 μ_{global} 及 $\sigma_{average}$ 分別為訓練韻律模型時所統計出的結果，代表所有語者基頻之總體平均及平均標準差。

最後要對音節能量、音節長度作正規化，方法與上述基頻正規化相同，對音節能量而言，(4.4)式中的 μ_s 、 σ_s 、 μ_{global} 及 $\sigma_{average}$ 分別改為代表第 s 個音檔音節能量之平均值、標準差、訓練韻律模型時統計之所有語者能量之總體平均及平均標準差；對音節長度而言，(4.4)式中的 μ_s 、 σ_s 、 μ_{global} 及 $\sigma_{average}$ 則分別改為代表第 s 個音檔音節長度之平均值、標準差、訓練韻律模型時統計之所有語者音節長度之總體平均及平均標準差。

4.1.2.2 各階段 lattice 之展開

在圖 4.4 當中，第二級辨認器中的三個階段會各自針對上一階段或第一級所產生的 lattice 依據所需之資訊作展開以方便未來作重新評分，這裡本研究是使用維特比搜尋演算法(viterbi search algorithm)來作 lattice 重新評分，以下將對其 lattice 展開過程作簡單說明：

- **第一階段：加入多種語言資訊**

針對 4.1.2 節一開始所述，產生帶有音節切割資訊的 trigram word lattice，然後再根據其他語言資訊作展開，展開方法如下：

(1).node expansion :

針對原始 lattice 中各個 node 所帶有的 word 資訊，找出相對應的 POS 數目(P)及 PM 數目(M)，再將各個 node 展開至 P*M 倍。

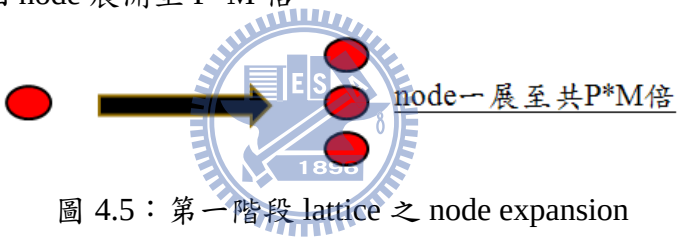


圖 4.5：第一階段 lattice 之 node expansion

(2).arc expansion :

針對原始 lattice 中各個 arc 所帶有的 word 資訊，找出相對應的 POS 數目(P)及 PM 數目(M)，並對上一個 arc 中所帶有的 word 資訊，找出相對應的 POS 數目(P2)及 PM 數目(M2)，再將各個 arc 展開至 P*M*P2*M2 倍。

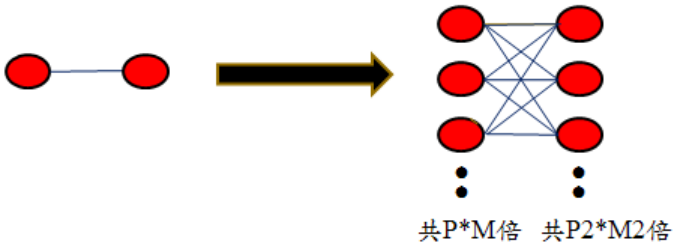


圖 4.6：第一階段 lattice 之 arc expansion

經過以上的展開流程後，產生的 lattice 裡每一個 arc 和 node 中皆包含了 word、POS、PM 等三種語言資訊。

- **第二階段：加入韻律邊界停頓資訊**

針對第一階段所產生的 lattice 作展開，由於這一步是引入停頓聲學模型及 break syntax model，所以如何作 lattice 展開必須先觀察數學式(3.3)及(3.12)，由於韻律資訊是 syllable level，所以現在 lattice 中每一個 arc 上會累積除 LM score、POS score、PM score、AM score 外，還有各個音節的 prosodic score，但由式(3.3)及(3.12)可發現，針對 intraword syllable 的部分，可由程式內部處理，不需要將 lattice 作展開的動作，但是針對 interword syllable 的部分就要多加考量，因此我們必須先對各 differential 參數所包含到的數學式(3.13)、(3.14)及(3.15)作觀察，從這三式中發現為了計算某個音節的 prosodic score，我們必須取得前後各一個音節的資訊，也就是說整個 lattice 將會根據每個 word 中第一個音節及最後一個音節的資訊作展開，也就是 interword syllable 的部分，但為了考量實作中記憶體用量有限，所以這裡將只針對前一個音節資訊，也就是針對 word 中最後一個音節的長度(duration)作展開，當然如果要針對音節長度作展開，那音節聲調(syllable tone)的部分為何不用展開，因為每一個 arc 如果最終是進入同一個 node，其語言參數是一樣的。接下來，要解釋為何可以不需要針對後一個音節資訊作處理，也就是不用將 word 中第一個音節的資訊作展開，因為我們利用到 viterbi search 的特性，以下將詳細說明。

經由上述停頓聲學模型及 break syntax model 的數學式(3.3)及(3.12)可發現，word 中每個音節的分數可以獨立計算，所以我們可以先算出每一個 word 中 intraword syllable 的 prosodic score，並結合 LM score、POS score、PM score 及 AM score 於 lattice 中各個 arc 上。接下來就是跑 backward viterbi search，這一步將會針對 interword syllable 的部分作處理，以下我們將以圖形解說的方式解釋整套演算法：

首先，由於是使用 backward viterbi 的關係，所以是先從 leaf node 開始往前推算出每一個 node 的 backward score，如下圖 4.7 所示：

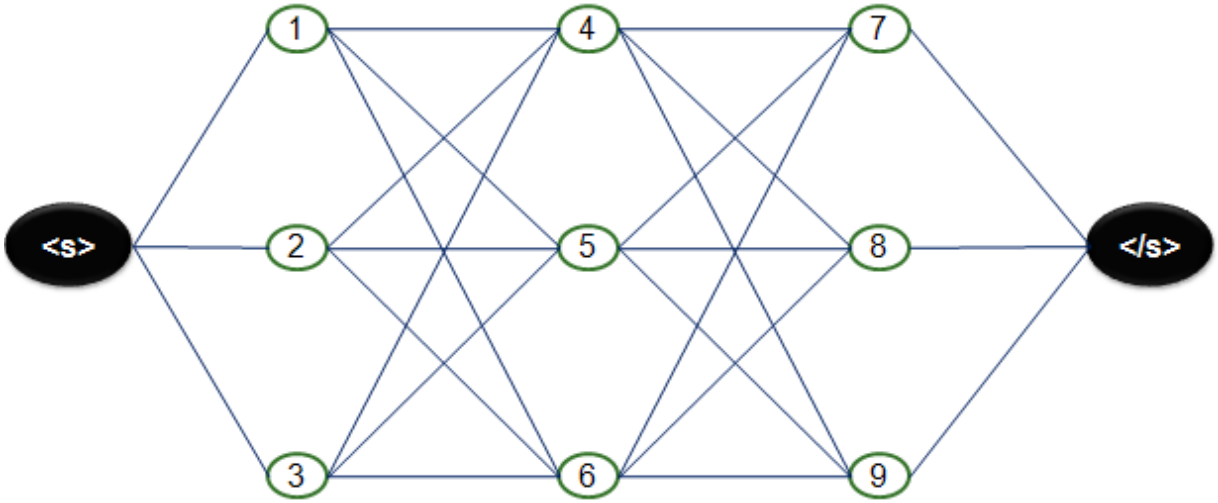


圖 4.7：第二階段 lattice 之架構範例

觀察上圖 4.7 中，由於 leaf node 所連接的 arc 中並未包含語言參數，所以 lattice 中最後一個 word 的資訊便是在 node7、node8 或 node9 的來源 arc 中，針對 node4 當例子，經由它分出的 arc 共有三條，如下圖 4.8 所示，這時我們將各路徑所累積分數的作比較後選出一條最佳路徑(如下圖 4.8 中深色部分所示)，並將這一條路徑上所帶的 word 資訊，也就是第一個音節的資訊以及所找出的最佳 backward score，記錄在 node4 中(如下圖 4.8 中紅線部分所示)，因為從倒數第二個 word 開始就要計算 interword syllable 的 prosodic score。而 node5 及 node6 也沿用相同作法算出其 backward score。

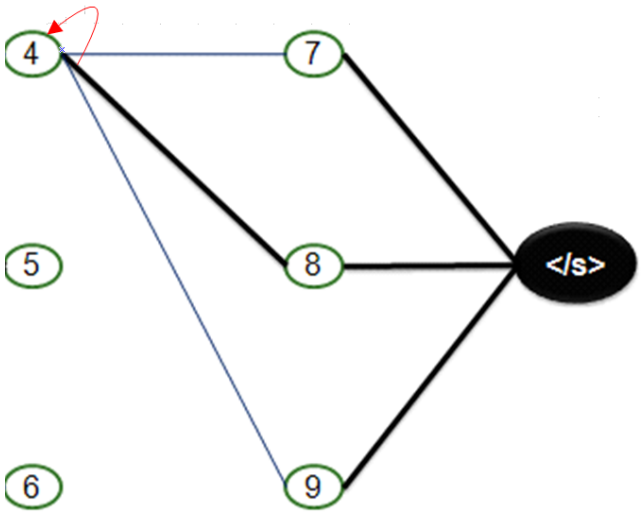


圖 4.8：第二階段 lattice 之內部工作流程範例 1

接下來我們針對圖 4.8 中的 node1 作分析，經由它分出的 arc 共有三條，如下圖 4.9 所示：

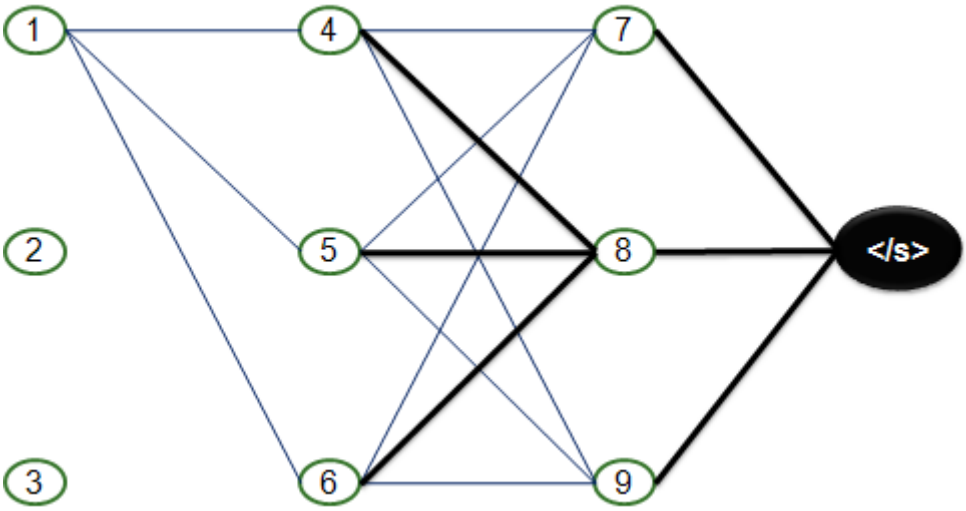


圖 4.9：第二階段 lattice 之內部工作流程範例 2

這時原本應將各路徑所累積分數的作比較後選出一條最佳路徑，但從倒數第二個 word 開始，每一個 arc 上的最後一個音節的 prosodic score 還沒計算，因為它會相依於下一個音節資訊，所以理論上要將 lattice 作展開，但在上一段的敘述中，已分別在 node4、node5 及 node6 中存入了下一個 word 的第一個音節資訊，所以根本不需要對 lattice 作展開，直接沿用這些 node 所傳遞過來的資訊來計算出目前最後一個音節的 prosodic score，這正是 viterbi 的精神所在，針對每一個 node 都找出到達終點的最佳路徑，所以對上一個 word 的而言，下一個 word 的第一個音節早已決定好了，不需要再對 lattice 作額外的展開，如此一來，我們便可針對 node1 找出一條最佳路徑，如下圖 4.10 所示，圖中紅線的部分就是代表訊息傳遞：

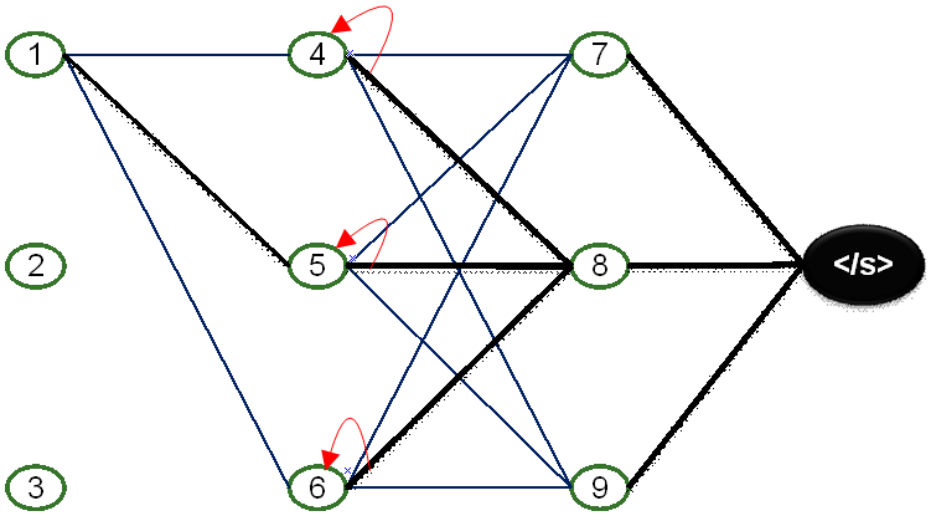


圖 4.10：第二階段 lattice 之內部工作流程範例 3

往後針對 lattice 中每一個 node 的 backward score 的計算都如同上述所言，如下圖 4.11 所示，圖中紅線的部分就是代表訊息傳遞：

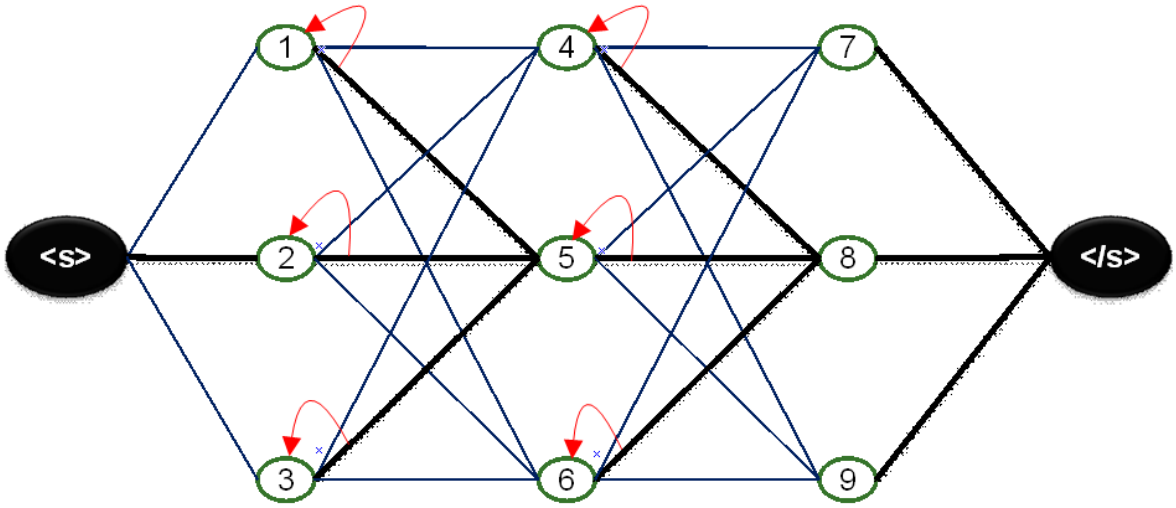


圖 4.11：第二階段 lattice 之內部工作流程範例 4

當計算到起始節點(start node)時，也就等於找到了最佳路徑，經由最終的路徑回溯，即可解碼出最佳的詞序列(word sequence)：

圖 4.12：第二階段 lattice 之內部工作流程範例 5

經由以上解說，現在 lattice 只須對 word 中最後一個音節的長度作展開，以下為展開流程：

(1).node expansion :

針對原始 lattice 中各個 node，觀察其來源 arc 中所帶有的 word 資訊(最後一個音節的長度)，現假設來源 arc 中最後一個音節的長度共有 M 種，則原始 node 將展開至 M 倍。以下為範例圖 4.13，圖中左側右邊的 node 根據來源 arc 中最後一個音節的長度作展開，假設一共有兩種不同的長度，則右側就是針對 node 進行展開後的結果：

圖 4.13：第二階段 lattice 之 node expansion

(2).arc expansion :

針對原始 lattice 中各個 arc，觀察其 start node 的特性，假設從第(1)步中，已知 start node 將被展至 M 倍，則原始 arc 也將展開至 M 倍，如下範例圖 4.14 所示：

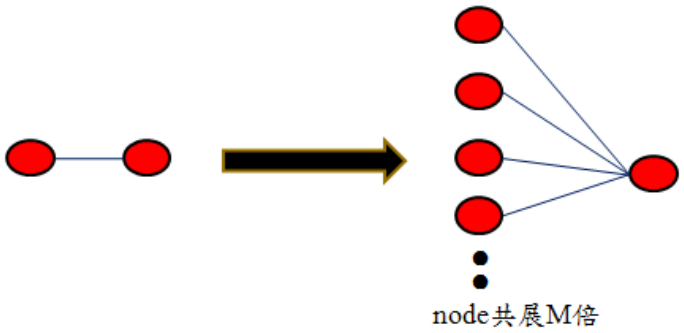


圖 4.14：第二階段 lattice 之 arc expansion

- **第三階段：加入音節韻律狀態資訊**

針對第二階段所產生的 lattice 作展開，由於這一步是引入音節韻律模型及韻律狀態模型，所以如何作 lattice 展開一樣要先觀察其數學式(3.4)及(3.5)，從這兩式中發現為了計算某一音節的 prosodic score，必須得到前後各一個音節資訊，而針對 intraword syllable 的部分，處理方式可能不像第二階段中只加入韻律邊界停頓資訊那樣簡單(註：計算 intraword syllable 的 prosodic break score 不用考慮到上一個音節的韻律邊界停頓類型，所以可以在跑 backward viterbi search 以前就先計算完) ，因為現在每一個音節的 prosodic state score 會相依於上一個音節的韻律狀態，所以 intraword syllable 的處理方式也較複雜，基本上為了實作上對記憶體的有效控制，在 intraword syllable 的部分我們不將 lattice 作展開(如何處理將會在稍後作補充說明)，但對 interword syllable 而言就要多加考量，所以說整個 lattice 將會根據每個 word 中第一個音節及最後一個音節的資訊作展開，但同樣為了考量實作中記憶體用量有限，在這只根據後一個音節資訊作展開，也就是 lattice 將根據每個 word 中第一個音節資訊作展開，而參照以上數學式(3.4)及(3.5)，word 中第一個音節資訊正好就是音節所對應的聲調資訊。

接下來要解釋為何不需要針對前一個音節資訊作展開，其原理如第二階段中敘述相同，利用 viterbi search 可以將資訊傳遞的特性，只是這一次我們是採用 forward viterbi search，跟第二階段的方式比較起來只是演算法運算的方向相反，其於觀念是一樣的，以下同樣以圖解的方式作說明：

首先，由於是使用 forward viterbi 的關係，所以是先從 start node 開始往前推算出每一個 node 的 forward score，如下圖 4.15 所示：

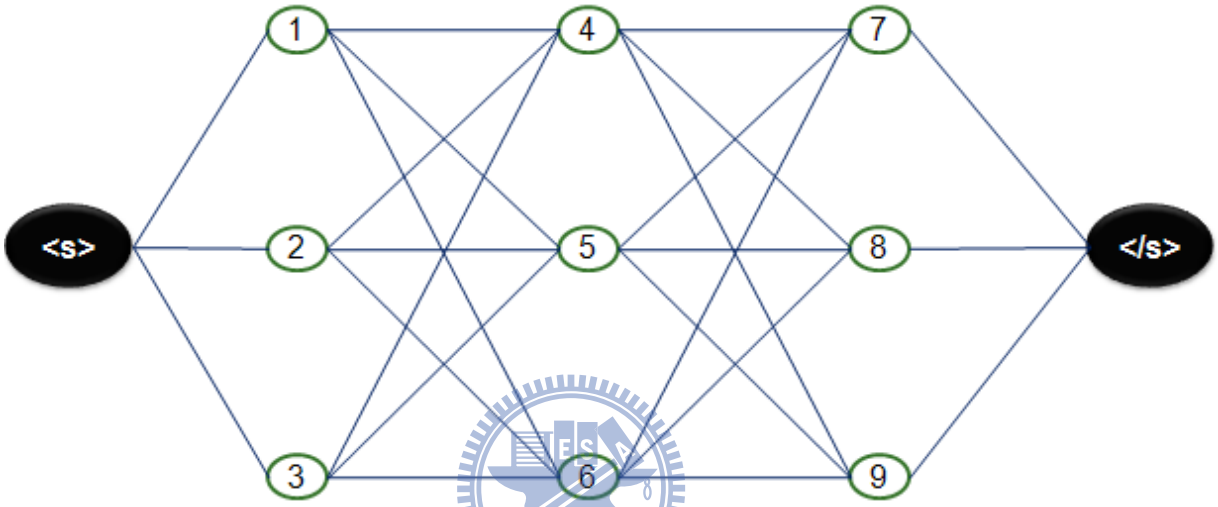


圖 4.15：第三階段 lattice 之架構範例

觀察上圖 4.15 中，針對 node1 當例子，它的來源 arc 只有一條，因為是第一個 word，如下圖 4.16 所示，這時我們將這一條路徑上所帶的 word 資訊，也就是最後一個音節的資訊(包含三種 prosodic state)以及所找出的最佳 forward score，記錄在 node1 中(如下圖 4.16 中紅線部分所示)，因為從第二個 word 開始就要計算 interword syllable 的 prosodic score。而 node2 及 node3 也沿用相同作法算出其 forward score。

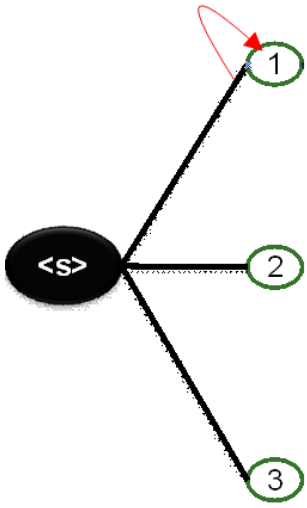


圖 4.16：第三階段 lattice 之內部工作流程範例 1

接下來我們針對圖 4.16 中 node4 作分析，它的來源 arc 共有三條，如下圖 4.17 所示：

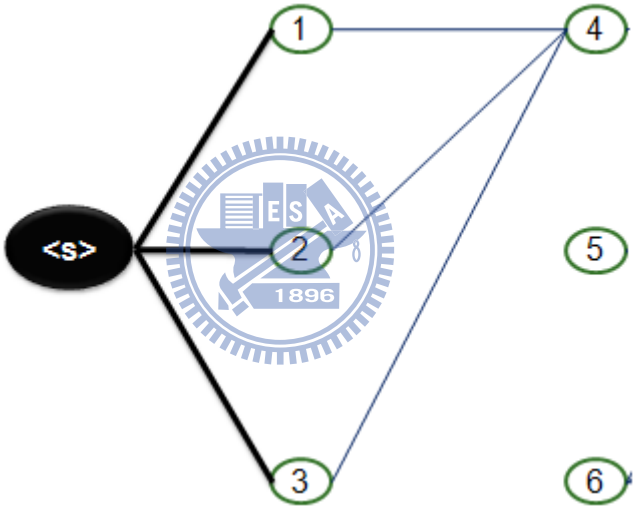


圖 4.17：第三階段 lattice 之內部工作流程範例 2

這時原本應將各路徑所累積之分數作比較後選出一條最佳路徑，但從第二個 word 開始，為了計算每一個 arc 上的第一個音節的 prosodic score，必須知道上一個音節資訊，所以理論上要將 lattice 作展開，但在上一段的敘述中，已分別在 node1、node2 及 node3 中存入了上一個 word 的最後一個音節資訊，所以根本不需要對 lattice 作展開，直接沿用這些 node 所傳遞過來的資訊來計算出目前第一個音節的 prosodic score；針對每一個 node 都找出從多條來源 arc 中的最佳路徑，所以對下一個 word 的而言，上一個 word 的最後一個音節早已決定好了，不需要再對 lattice 作額外的展開，如此一來，我們便可針對 node4 找出一條最佳路徑，如下圖 4.18 所示，圖中紅線的部分就是代表訊息傳遞：

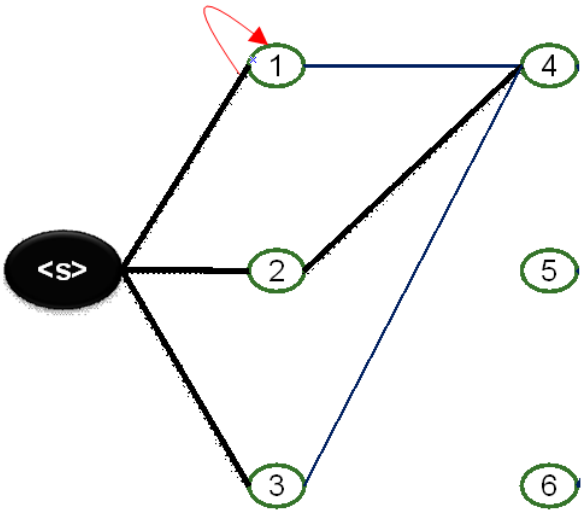


圖 4.18：第三階段 lattice 之內部工作流程範例 3

往後針對 lattice 中每一個 node 的 forward score 的計算都如同上述所言，如下圖 4.19 所示，圖中紅線的部分就是代表訊息傳遞：

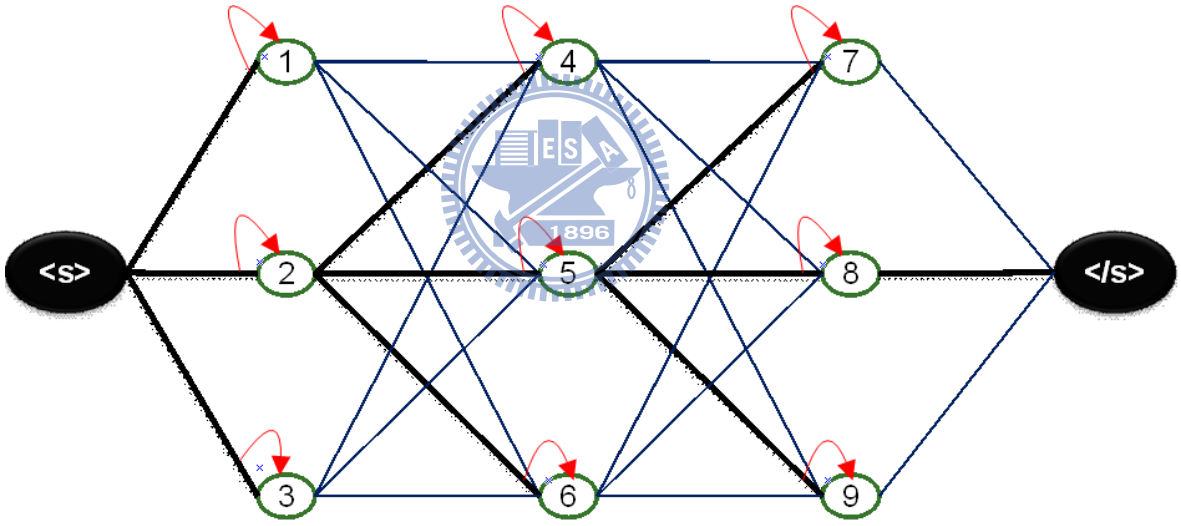


圖 4.19：第三階段 lattice 之內部工作流程範例 4

當計算到終止節點(leaf node)時，也就等於找到了最佳路徑，經由最終的路徑回溯，即可解碼出最佳的詞序列(word sequence)：

圖 4.20：第三階段 lattice 之內部工作流程範例 5

經由以上解說，現在 lattice 只須對 word 中第一個音節的聲調作展開，以下為展開流程：

(1).node expansion :

針對原始 lattice 中各個 node，觀察其分出的 arc 中所帶有的 word 資訊(第一個音節的聲調)，現假設分出的 arc 中第一個音節的聲調共有 M 種，則原始 node 將展開至 M 倍；以下為範例圖 4.21，圖中左側左邊的 node 根據分出的 arc 中第一個音節的聲調作展開，假設一共有兩種不同的聲調，則右側就是針對 node 進行展開後的結果：

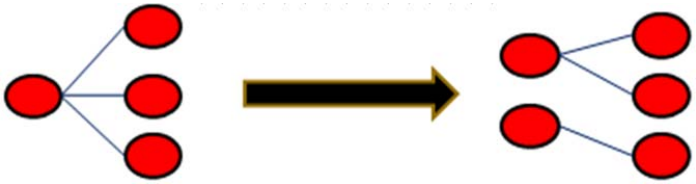


圖 4.21：第三階段 lattice 之 node expansion

(2).arc expansion :

針對原始 lattice 中各個 arc，觀察其 end node 的特性，假設從第(1)步中，已知 end node 將被展至 M 倍，則原始 arc 也將展開至 M 倍，如下範例圖 4.22 所示：

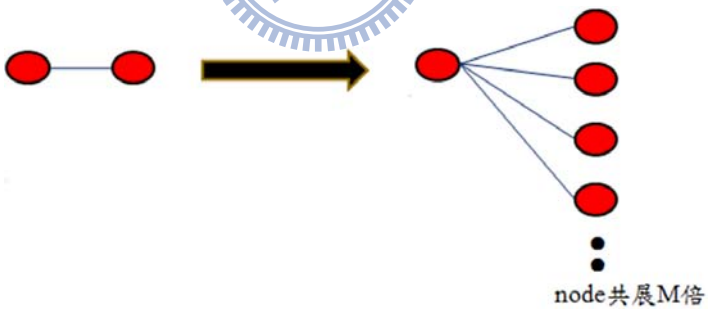


圖 4.22：第三階段 lattice 之 arc expansion

以上是介紹 lattice 中面對到 interword syllable 的必要展開動作；但若是針對 intraword syllable 的部分，我們之所以不將 lattice 作展開，是因為我們將各個 word 中 intraword syllable 的部分獨立處理，可以說是在 lattice 中的一個 arc 內跑一個 viterbi search，將最後一個音節所累積的最佳資訊傳遞到下一個 node 中，以下將以圖解的方式作說明：

下圖 4.23 為一範例，是針對一個三字詞作範例，圖中最左邊的 node1 會傳遞給下一個 word 的第一個音節一些訊息(如圖中左側紅線所示)，這些訊息代表上一個 word 最後

一個音節的韻律狀態,所以當計算第一個音節的 prosodic state score 時就需要上一個 node 所傳遞的訊息才能計算分數，然後就如同跑 viterbi 的方式，針對一個音節中的每一種 prosodic state 都找到最佳的分數(如圖中深綠線所示)，當跑完最後一個音節時，再從這個音節中 16 個 prosodic state 裡選出一個累積分數最多的 state，並將這個 state 資訊存到下一個 node(node2)中供下一個 arc 來擷取(如圖 4.23 中右側紅線所示),而最終當下的 arc 所累積的分數就包含上一個 node 的 forward score、目前 word 的 FLM score (包含 LM score、POS score 及 PM score)、目前 word 的 prodosic break score 及之前所累積最高的 prosodic state score。

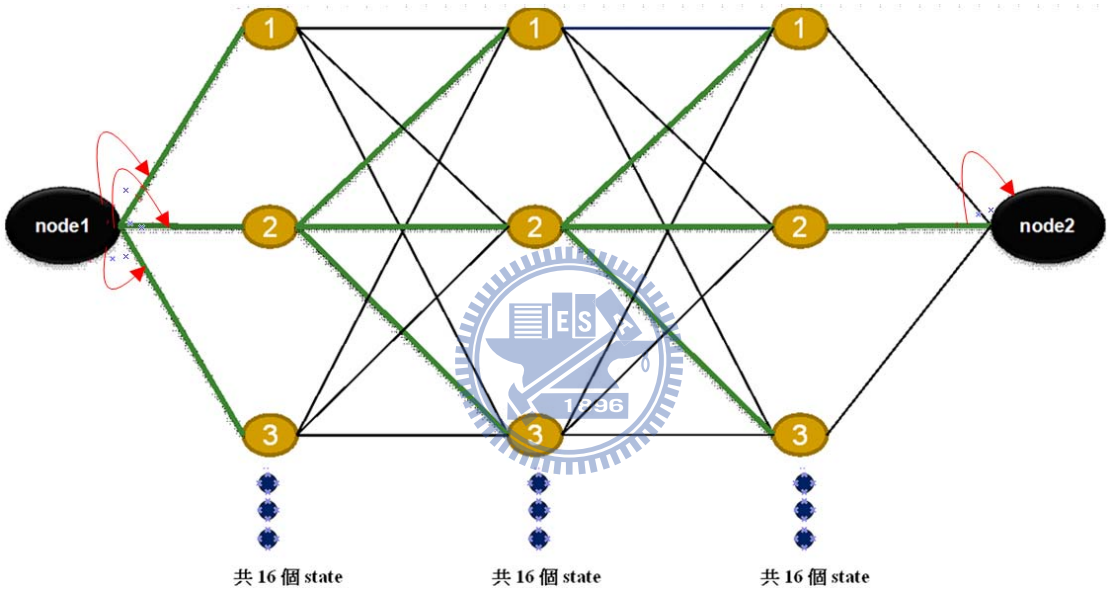


圖 4.23：第三階段 lattice 之內部工作流程範例 6

4.2 鑑別式模型組合

在本研究裡，第二個 stage 中拿來作重新評分的模型共有 16 個，因此如何找出一組權重使這 16 個模型作結合後能夠得到最小的詞錯誤率便是非常重要之課題，如果使用 trail-and-error 的方式來決定 16 個模型的權重將非常耗費時間且缺乏效率，所以本研究使用鑑別式模型組合(Discriminative Model Combination) [20]的方法(簡稱 DMC)來決定權重，以下將作簡單的說明：

DMC 的方法是先定義一個 decision error rate 的鑑別式函數(discriminant function)如

(4.7)式，目標是找到一組權重使此函數的 decision error rate 最佳化。

$$\begin{aligned} & g(x_1^T, w_1^S, w'^{S'}_1) \\ & = \log P(w_1^S \mid x_1^T) - \log P(w'^{S'}_1 \mid x_1^T) \\ & = \log[P(w_1^S)P(x_1^T \mid w_1^S)] - \log[P(w'^{S'}_1)P(x_1^T \mid w'^{S'}_1)] \end{aligned} \quad (4.7)，$$

(4.7)式中 $w_1^S = (w_1, ...w_S)$ 代表詞串，$x_1^S = (x_1, ...x_T)$ 代表特徵參數向量，$P(w_1^S \mid x_1^T)$ 代表在給定特徵參數條件下得到**正確詞串**的分數；而 $P(w'^{S'}_1 \mid x_1^T)$ 則代表在給定同樣特徵參數條件下得到**辨認結果詞串**的分數，雖然當這兩種分數愈趨近代表愈好，但分數最接近者不代表詞錯誤率(WER)會是最小。現在假如 $P(w_1^S \mid x_1^T)$ 將拆可解成 M 個不同模型，其線性對數(log-linearly)組合如下：

$$P_{\{\Lambda\}}^{\prod}(x_1^T \mid w_1^S) = \exp\{\log C(\Lambda) + \sum_{j=1}^{M} \lambda_j \log P_j(x_1^T \mid w_1^S)\} \quad (4.8)，$$

(4.8)式中 $\Lambda = (\lambda_1, ..., \lambda_M)^{tr}$ 代表模型 P_j 分數組合時的權重；$C(\Lambda)$ 代表正規化因子。由於現在 $P(w_1^S \mid x_1^T)$ 可拆成 M 個不同模型，其鑑別式函數將改寫成下式：

$$g(x_1^T, w_1^S, w'^{S'}_1) = \sum_{j=1}^{M} \lambda_j (\log P_j(w_1^S \mid x_1^T) - \log P_j(w'^{S'}_1 \mid x_1^T)) \quad (4.9)。$$

最後將定義一個 smooth misclassification function $\ell(x_n, k_{n0}, \Lambda)$，並搭配 Generalized Probabilistic Descent (GDP) algorithm[20]來得到多個模型的權重值 $\boldsymbol{\Lambda}$，以下在簡述其作法前，先對未來會使用到的符號作些定義：

定義一：詞串 w_1^S 表示為 class k；而每個句子 x_1^T 表示為特徵參數向量 x。

定義二：訓練資料表示為 $(x_n, k_{nr}), n = 1, ..., N, r = 0, ..., K$, 其中 N 代表句子數目；k_{n0} 代表特徵參數向量 x_n 的標準答案；$k_{nr}, r = 1, ..., K$ 代表 k_{n0} 的競爭者，意即 K-best 序列。

定義三：$LD(k_{nr}, k_{n0})$ 代表 Levenshtein-distance，意即 hypothesis k_{nr} 的錯誤數量，錯誤包含插入性、刪除性、取代性等。

定義四：訓練語料(或 held-out data)的 smoothed empirical error rate $L(\Lambda)$ 為：

$$L(\Lambda) = \frac{1}{N} \sum_{n=1}^{N} \ell(x_n, k_{n0}, \Lambda) \quad (4.10)，$$

$$\ell(x_n,k_{n0},\Lambda)^{-1}=1+A\cdot\left(\frac{1}{K}\sum_{r=1}^{K}e^{\left\{-\eta LD(k_{nr},k_{n0})\log\frac{p^{\Pi}_{\{\Lambda\}}(k_{n0}|x_n)}{p^{\Pi}_{\{\Lambda\}}(k_{nr}|x_n)}\right\}}\right)^{-\frac{B}{\eta}},\ \mathrm{A}>0,\mathrm{B}>0,\eta>0 \quad (4.11)。$$

有了上列各定義後並利用式(4.13)的遞迴架構可以求出權重值λ_j：

For j=1,...,M

$$\lambda_j^{(0)}=1$$

$$\lambda_j^{(I+1)}=\lambda_j^{(I)}+\varepsilon\sum_{n=1}^{N}\ell(x_n,k_{n0},\Lambda^{(I)})\left(1-\ell(x_n,k_{n0},\Lambda^{(I)})\right)\cdot \frac{\sum_{r=1}^{K}LD(k_{nr},k_{n0})\log\left(\frac{p_j(k_{n0}\mid x_n)}{p_j(k_{nr}\mid x_n)}\right)\left[p^{\Pi}_{\{\Lambda^{(I)}\}}(k_{nr}\mid x_n)\right]^{\eta LD(k_{nr},k_{n0})}}{\sum_{r=1}^{K}\left[p^{\Pi}_{\{\Lambda^{(I)}\}}(k_{nr}\mid x_n)\right]^{\eta LD(k_{nr},k_{n0})}}$$

$$\Lambda^{(I+1)}=(\lambda_1^{(I+1)},...,\lambda_M^{(I+1)})^{tr} \quad (4.12)，$$

在(4.12)式中ε代表 stepsize，式中呈現λ_j在多次遞迴中決定於鑑別式函數$\log\left(\frac{p_j(k_{n0}\mid x_n)}{p_j(k_{nr}\mid x_n)}\right)$的權重和。

第五章 實驗結果與分析

本章將介紹本研究所作的實驗並進一步分析結果，整套分析流程分為三大部分，第一部份是 joint syntax model 的建立；第二部份是階層式韻律模型的訓練；最後則是將韻律訊息加入語音辨認的實驗，這部分包含圖 4.4 中第二級辨認器裡三階段的實驗結果及分析。

5.1 Joint Syntax Model 之建立

在本研究所使用的 joint syntax model 裡包含一個 trigram LM、一個 factored POS model 與一個 factored PM model，其中 trigram LM 的建立已於第二章簡單介紹過，而其餘兩種 factored model 則是使用 SRILM toolkit[19]所建構而成，然而在建立模型之前要先對提供訓練的文字資料作些前處理，包含將資料中所有的詞(word)都進行詞性(POS)及標點符號(PM)的標記，而本研究的 PM 標記一共分成 4 類，分別是逗號(COM)、頓號(DOT)及其他類標點符號(OTH)；POS 則一共分成 46 類，其中針對本研究所使用的六萬詞字典而言，每一個詞彙平均可對應到 1.5 種詞性。在有了完整的訓練文字資料庫後，factored model 的訓練流程主要分成兩個階段，完整訓練流程如圖 5.1 所示：

- 第 1 階段：從訓練文字資料中產生 fngram-count 檔案，代表統計 factored model 中所有層級組合出現在訓練文字資料中的次數。
- 第 2 階段：從第 1 階段產生的 fngram-count 檔案中訓練出 factored model。

此外，本研究在訓練 factored model 時皆給定一條固定的 backoff 路徑，如圖 4.2、圖 4.3 所示，另外還需注意的是上述的兩個訓練步驟中都需匯入一個 flm 檔案，它主要是設定 factored model 內每一層 backoff 結構中所考慮的 factor。

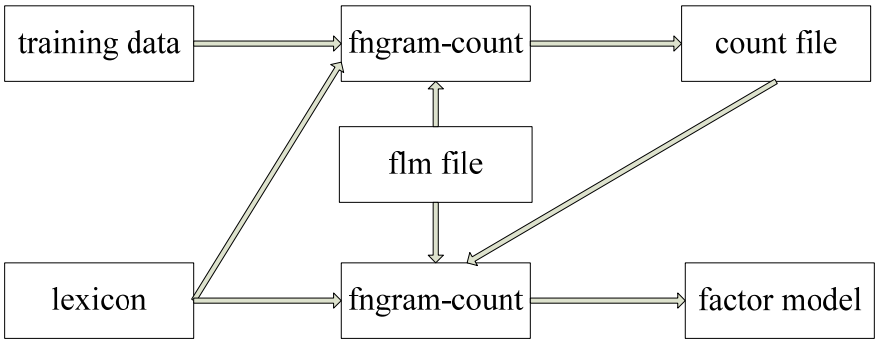


圖 5.1：factored model 訓練架構流程圖

兩種 factored model 訓練完成後，其 perplexity 效能評估於表 5.1、5.2 所示，藉由觀察 perplexity 的下降，表示每加入一個 factor 都有助於預估 POS 或 PM。

表 5.1：factored PM model 的 perplexity

factored PM model with different factors	perplexity	
	ppl	ppl1
$P(PM_{i-1} \mid W_{i-1})$	1.6572	1.6699
$P(PM_{i-1} \mid W_{i-1}, POS_{i-1})$	1.6299	1.6420
$P(PM_{i-1} \mid W_{i-1}, POS_{i-1}, POS_i)$	1.4458	1.4539

表 5.2：factored POS model 的 perplexity

factored POS model with different factors	perplexity	
	ppl	ppl1
$P(POS_i \mid W_i)$	1.3120	1.3286
$P(POS_i \mid W_i, POS_{i-1})$	1.2577	1.2712

5.2 階層式韻律模型之訓練

5.2.1 Break Syntax Model

Break syntax model 是根據多種語言參數將不同類的韻律邊界停頓作分類而得到一顆決策樹，其方法是利用 CART 演算法並搭配一個設計好的問題集(如附錄一)推導而出，決策樹中的每一個終止節點(leaf node)將存入每一類韻律邊界停頓的機率值，同時針對樹中的各非終止節點(nonterminal node)所問到的問題來分析其重要程度。

本研究所使用的韻律模型是利用 TCC300 中約十萬個音節來訓練而成，下圖 5.2 及圖 5.3 就是訓練出來的決策樹架構，同時在利用 CART 演算法訓練韻律模型時還需要兩個基本設定如下所示：

設定一：決策樹中各終止節點(leaf node)內最小樣本數量(音節數量)必須大於 700。

設定二：訓練韻律模型過程中其相對相似度增益(relative likelihood gain)要大於 0.001。

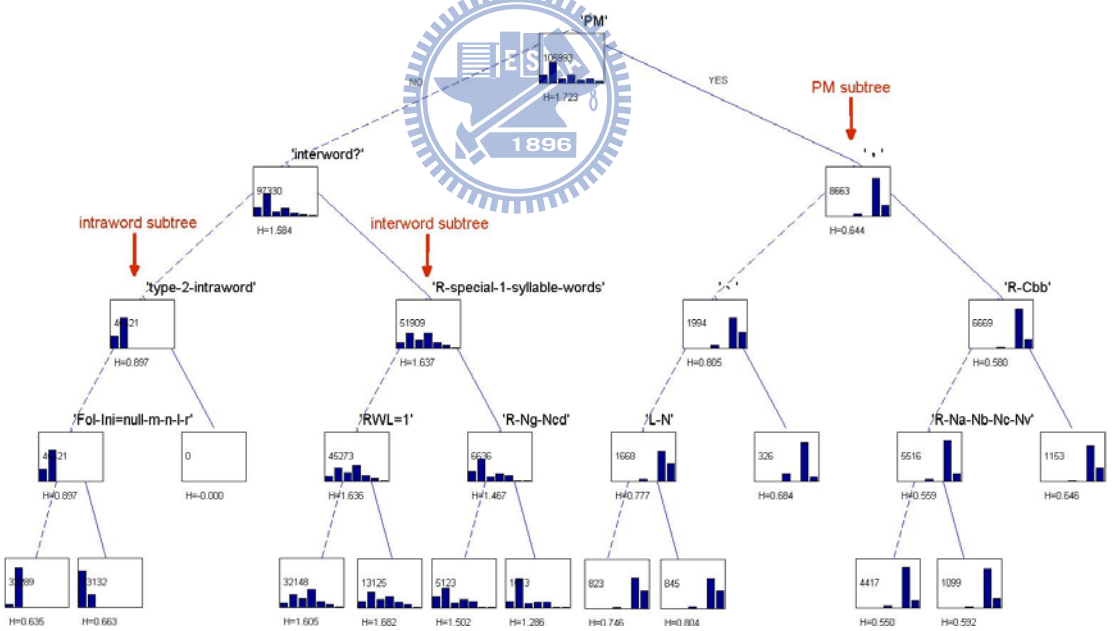


圖 5.2：break syntax model 的決策樹架構

上圖 5.2 中每一個節點(node)裡代表 7 種韻律邊界停頓類型之分布圖(由左至右分別為 *B*0, *B*1, *B*2-1, *B*2-2, *B*2-3, *B*3, *B*4)，*H* 則代表 Shannon entropy，用以評估韻律邊界停頓類型分布之不確定性。

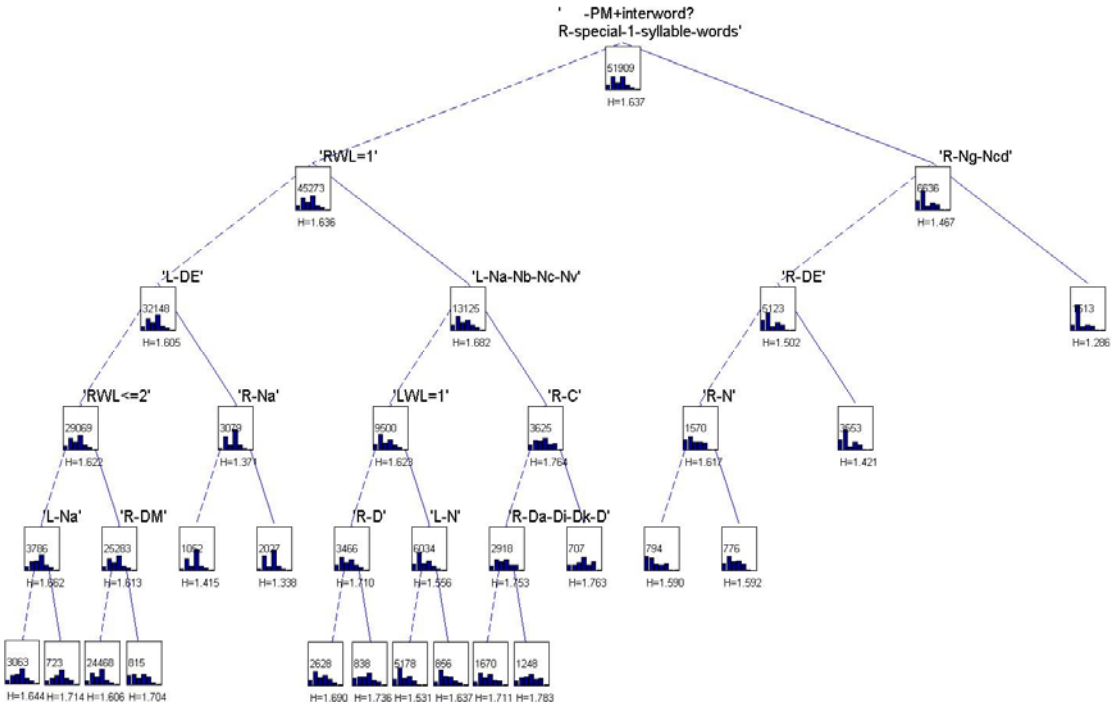


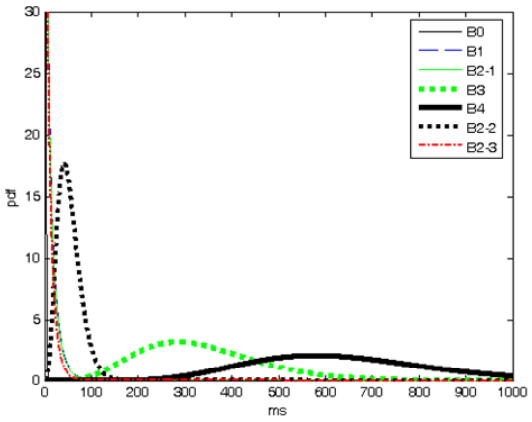
圖 5.3：圖 5.2 中 break syntax model 的決策樹架構更深層部分

觀察圖 5.2 的決策樹，從根節點(root node)開始往下生長，長的愈深就代表為了要預估出韻律邊界停頓所考慮的語言資訊就愈多，而韻律邊界停頓之機率分佈的 entropy 也愈來愈低；我們觀察這棵樹中最重要的兩種語言資訊分別是 PM 及 interword/intraword，而樹中又分成三顆子樹：PM 子樹、interword 子樹及 intraword 子樹，其中 PM 子樹及 intraword 子樹的韻律邊界停頓之機率分佈的 entropy 隨著節點愈往下長其下降幅度愈大，也愈早到達 leaf node，從樹中我們不難發現在 PM 子樹中，根節點的韻律邊界停頓之機率分布大多集中在 B3、B4 等長停頓；反之在 intraword 子樹中則大多集中在 B0、B1。而針對 interword 子樹而言，韻律邊界停頓之機率分佈的 entropy 就下降的緩慢許多，而樹本身結構相較於其他兩顆子樹要複雜得多，也因此為了要使預估更精確，必須向下問更多重要問題，像是「右邊或左邊之特殊一字詞」，這是一個很重要的問題，與本研究想解決中文語音辨認上的一字詞易混淆及搶詞錯誤有很大的關連性，圖 5.3 就是針對 interword 子樹的更深層結構。

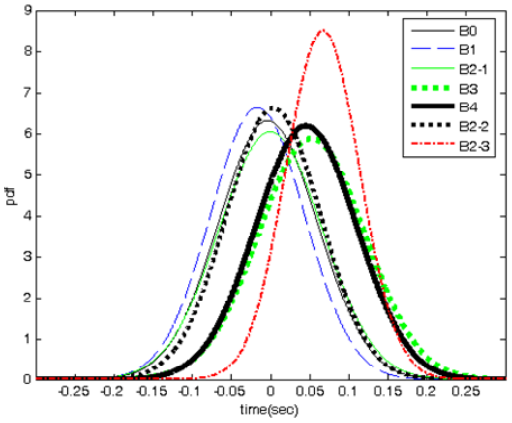
5.2.2 停頓聲學模型

在這部分的實驗，首先我們將各種停頓標記之下，決策樹根節點中各個參數的機率密度函數畫出，即不考慮語言參數之下各停頓標記之參數分佈，如圖 5.4。從圖 5.4 (a)可以看出 B0 的停頓時長最短，接著 B1、B2-1、B2-3 的停頓時長次之且機率密度函數幾乎重疊在一起，B2-2、B3、B4 的停頓時長依序明顯增加。觀察圖 5.4 (b)、5.4 (c)，雖然正規化音節長度拉長現象不明顯，但還是可以看出主要分成兩個部分，B2-3、B3、B4 的音節長度延長現象普遍會大於 B0、B1、B2-1、B2-2。圖 5.4 (d)是正規化音節間基頻差的分布，除了 B3、B4，B2-1 也表現出有明顯基頻差，B0、B1、B2-3 在正規化音節間基頻差則沒有大的鑑別度。從圖 5.4 (e)可看出韻律邊界停頓在音節間能量低點(energy dip)上分布的情況，對照停頓時長分布來看，確實在有比較長停頓時長的音節邊界如 B2-2、B3、B4 來看，energy dip 會比停頓時長較短的邊界還來的低。

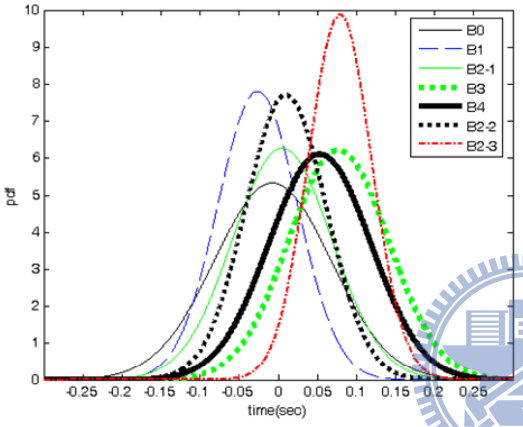
接下來將停頓聲學模型中七種韻律邊界停頓的決策樹結構之主要部分給畫出來，如圖 5.5 所示，一般來說，在韻律階層結構中，用來區分愈高階層韻律組成份子的韻律邊界停頓通常會具有較長的停頓時長(pause duration)、較低的音節間能量低點(energy-dip)、較大的正規化基頻跳躍值(normalized pitch-level jump)、及較大的音節長度影響因子(duration lengthening factors)。對於每一種韻律邊界停頓來說，隨著決策樹長的愈深，代表需要更多語言資訊來預估停頓聲學參數，各節點中的相似度(likelihood) $P(\mathbf{pd},\mathbf{ed},\mathbf{pj},\mathbf{dl},\mathbf{df}|\Upsilon_s,\Lambda_p,\Lambda_l)$ 便會增加，這意謂著語言參數確實能對於建立停頓聲學模型有一定的幫助、對預估停頓聲學參數更加準確。這裡值得注意的是，B4 決策樹的根節點並沒有向下分裂，因為在節點中各個停頓聲學參數的分佈相對均勻；另外針對 B3 及 B2-2 等 pause-related 韻律邊界停頓，用來分裂決策樹的問題會與較高階層的語法參數相關，像是 PM 和 POS；相對地，B0、 B1、B2-1 及 B2-3 等 non-pause 韻律邊界停頓，用來分裂決策樹的問題會與較低階層的語言參數相關，像 interword/intraword 和 phonetic features。



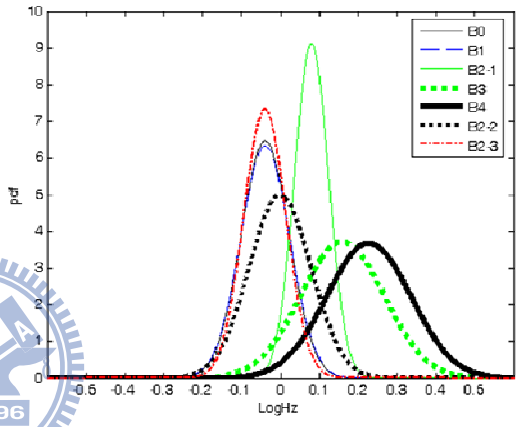
(a)



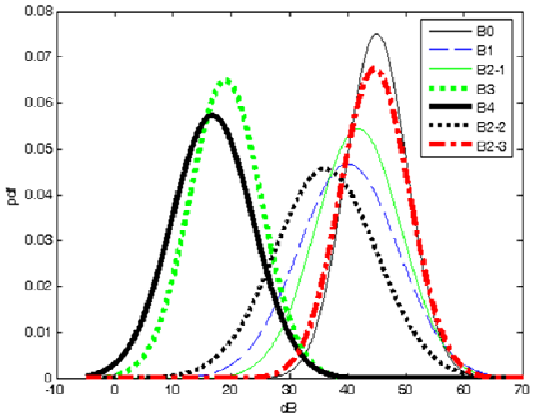
(b)



(c)



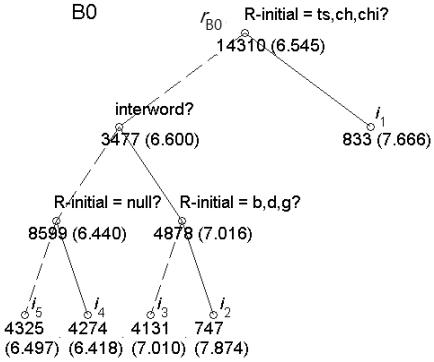
(d)



(e)

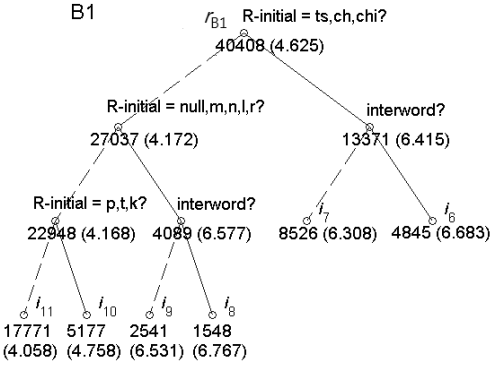
圖 5.4：(a)音節停頓長度 (b)正規化音節延長因子 1 (c)正規化音節延長因子 2 (d) 正規化基頻跳躍值之分布圖 (e)音節間能量低點

B0



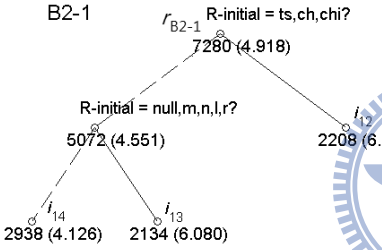
r_{B0} (1, 45, -0.039, -3, -7)
i_1 (1, 46, -0.035, 59, 97)
i_2 (1, 44, 0.024, -6, -24)
i_3 (1, 45, -0.048, -18, -32)
i_4 (1, 45, -0.042, -9, -15)
i_5 (1, 45, -0.038, 5, 7)

B1



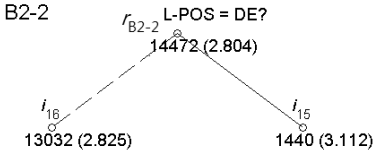
r_{B1} (9, 40, -0.039, -17, -26)
i_6 (1, 41, -0.048, -12, -34)
i_7 (1, 41, -0.035, -14, -19)
i_8 (1, 29, -0.061, -37, -56)
i_9 (1, 29, -0.030, -54, -57)
i_{10} (8, 41, -0.042, -15, -19)
i_{11} (18, 41, -0.037, -12, -23)

B2-1



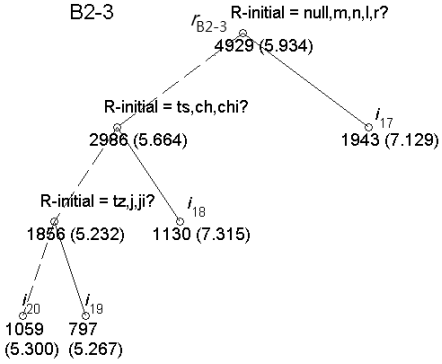
r_{B2-1} (9, 42, 0.080, -2, 5)
i_{12} (1, 42, 0.081, 2, 6)
i_{13} (2, 42, 0.076, 3, 9)
i_{14} (20, 42, 0.083, -3, 2)

B2-2



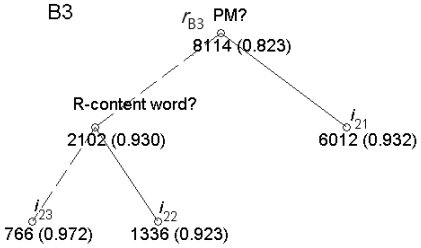
r_{B2-2} (55, 36, 0.000, 4, 10)
i_{15} (55, 29, -0.024, 12, 3)
i_{16} (55, 37, 0.003, 3, 11)

B2-3

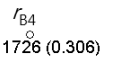


r_{B2-3} (7, 45, -0.039, 67, 79)
i_{17} (1, 45, -0.045, 70, 92)
i_{18} (1, 45, -0.040, 64, 78)
i_{19} (19, 44, -0.040, 62, 62)
i_{20} (14, 44, -0.024, 70, 69)

B3



B4



r_{B3} (339, 19, 0.160, 52, 77)
i_{21} (360, 19, 0.178, 51, 78)
i_{22} (279, 20, 0.099, 57, 73)
i_{23} (279, 20, 0.123, 58, 76)
r_{B4} (642, 17, 0.227, 46, 53)

圖 5.5：停頓聲學模型針對 7 種韻律邊界停頓之決策樹架構

上圖 5.5 中實線(虛線)代表針對問提回答正確(錯誤)之走向；另外在每一個節點(node)中的數字代表樣本數量及對樣本之平均相似度(括號中之數值)；對於每一個節點之統計值以列表方式呈現在各決策樹之下，注意 *r*'s 代表每一種停頓類型的根節點(root node)，表格中括號內的數值由左至右分別代表平均音節間停頓長度(pause duration)(ms)、音節間能量低點(energy-dip)大小(dB)、正規化之基頻跳躍(normalized pitch jump)(log-Hz)及兩種音節長度影響因子(duration lengthening factors)(ms)。

5.2.3 韻律狀態模型

圖 5.6 是針對在給定韻律邊界停頓的情況下之音節音高韻律狀態轉移機率 $P(p_n \mid p_{n-1}, B_{n-1})$，對於B0或B1，可以觀察到狀態轉移趨勢(high-to-low)，以及一次轉移的幅度幾乎都是nearby-state transitions，證明了在韻律詞(PW)內音節音高韻律狀態轉移是由高階緩慢下降的；對於B2-2而言，狀態轉移有兩種趨勢(high-to-low及low-to-high)；對於B2-1、B3、及B4而言，可以從它們的low-to-high狀態轉移趨勢裡發現到明顯的pitch reset現象，這些現象通常會在跨越韻律詞(PW)、韻律短語(PPh)及呼吸組/韻律句組(BG/PG)時發生，與這些明顯的reset現象相比，在B2-2條件下reset現象就不明顯；最後，B2-3的狀態轉移趨勢跟B0及B1非常類似，這代表了在音節拉長效應(duration lengthening)之後的韻律詞邊界裡沒有明顯的pitch reset。

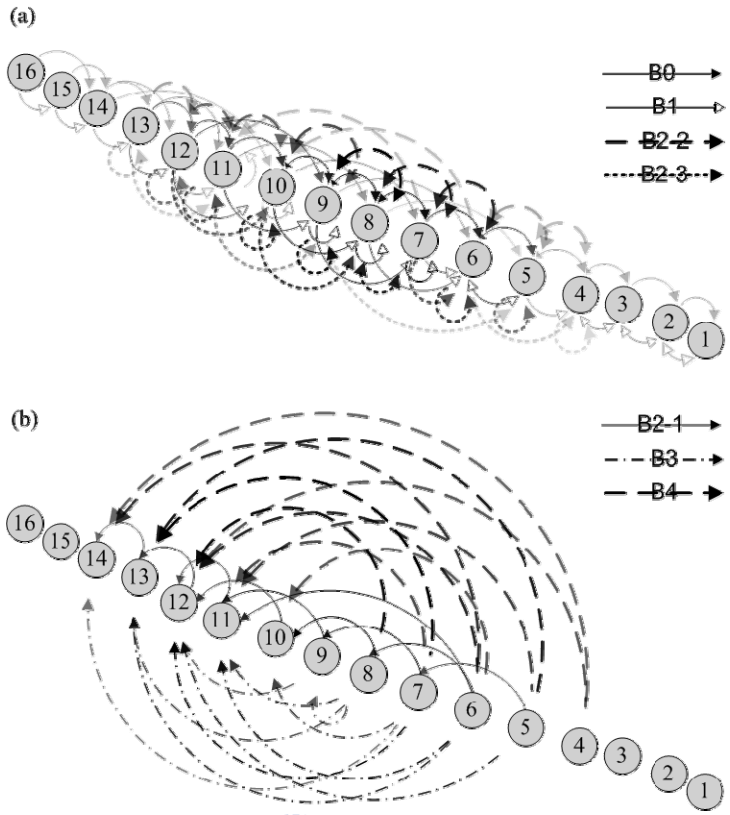


圖 5.6：基於不同韻律邊界停頓類型之的音節音高韻律狀態轉移

上圖 5.6 中(a)是基於 B0、B1、B2-2 及 B2-3 之下；(b)則基於 B2-1、B3 及 B4 之下。圖中每一個節點(node)代表韻律狀態的編號，編號愈大代表音節擁有較高的 log-F0 值，同時圖中較深的線則代表較重要的狀態轉移。

5.3 使用韻律訊息於語音辨認

針對實驗中第一個 stage 的部份，聲學模型的部份是採用 411 個音節 8 個 state 的 HMM 模型，訓練方式是使用 HTK 3.4[23]以 MMIE 法則[24]訓練得到；至於對音檔抽取的聲學參數是 MFCC，它的成分包括 12 維 MFCC 加上能量共 13 維，取 Delta 和 Delta-Delta，將參數變化的訊息也提供給辨認器使用，總共 39 維；語言模型的部份包含 bigram 及 trigram，以下將會呈現出如圖 4.4 所示之辨認器第二個 stage 中各階段的多項語言參數辨認率，包含詞(word)、字(character)、基本音節(base syllable)、帶聲調音節(tonal syllable)、音節聲調(syllable tone)、詞性(POS)及標點符號(PM)。

這邊做個說明，以下數據表格中第一列的 baseline 數據是指圖 4.4 中第一個階段之

實驗結果；第二列數據代表圖 4.4 中第二個階段之實驗結果；第三列數據則代表圖 4.4 中第三個階段之實驗結果。

表 5.3：詞(word)辨認率

	Accuracy(%)
Baseline	75.78
+Prosodic break	79.00
+Prosodic break +Prosodic state	79.88

表 5.4：字(character)辨認率

	Accuracy(%)
Baseline	82.11
+Prosodic break	85.29
+Prosodic break +Prosodic state	86.15

表 5.5：音節(syllable)辨認率

	Accuracy(%)
Baseline	88.18
+Prosodic break	89.88
+Prosodic break +Prosodic state	90.62

表 5.6：帶聲調音節(tonal syllable)辨認率

	Accuracy(%)
Baseline	83.99
+Prosodic break	86.99
+Prosodic break +Prosodic state	88.06

表 5.7：音節聲調(syllable tone)辨認率

	Precision	**Recall**	**F-Measure**
Baseline	88.00%	87.66%	87.83%
+Prosodic break	90.98%	90.58%	90.78%
+Prosodic break +Prosodic state	92.11%	92.07%	92.09%

表 5.8：詞性(POS)辨認率

	Precision	**Recall**	**F-Measure**
Baseline	93.43%	76.71%	84.25%
+Prosodic break	93.46%	79.92%	86.16%
+Prosodic break +Prosodic state	93.49%	80.79%	86.68%

表 5.9：標點符號(PM)辨認率

	Precision	Recall	F-Measure
Baseline	55.79%	37.43%	44.80%
+Prosodic break	69.54%	56.69%	62.46%
+Prosodic break +Prosodic state	66.28%	49.85%	56.90%

針對以上表格中 POS、PM 及音節聲調辨認率的計算，我們是使用 F-measure 的方式，以下將詳細說明。

5.3.1 Syllable Tone 辨認率算法

計算音節聲調(syllable tone)辨認率的部分，是先針對 base syllable 的標準答案及 base syllable 的辨認結果作 syllable align，使 syllable 之間有正確的對應，然後再根據前述所作好的 syllable align 來進行聲調標準答案和聲調辨認結果的比對，在比對過程中如果 syllable 的部分為 NULL(代表是 syllable insertion 或 syllable deletion)，則將聲調標示為 NONE，表示這個位置不具有聲調標記，最後根據比對的結果統計出 confusion matrix，以及利用統計出的 confusion matrix 來計算 F-measure score。

5.3.2 POS 辨認率算法

計算 POS 辨認率並不是直接針對辨認結果以及正確答案來計算，因為通常辨認出來的 word 也要正確，這樣辨認出來的 POS 才有意義，所以我們使用另一種計算方式(F-measure)，就是先統計出在 word 辨認正確的條件之下 POS 辨認正確的數量 ***H***，以及在 word 辨認正確的條件之下 POS 總數 ***N***，最後則是 POS 答案中的總數量 ***R***。

有了以上統計結果，接下來則分別計算 POS 的 Recall (***H*** / ***R***) 及在 word 辨認正確的條件之下，POS 的 Precision (***H*** / ***N***)，最後有了 Precision 及 Recall 就能算出 F-measure score。

5.3.3 PM 辨認率算法

這裡計算 PM 辨認率是先將 syllable 的標準答案及 syllable 的辨認結果與 PM 標準答案及 PM 辨認結果做結合，完成後接著做 syllable + PM align，使 syllable 及 PM 之間有正確的對應，然後再將前述做好 align 的檔案中非 PM 的部份(包含 syllable 及 NULL 的部份)標示成 NONE，表示這個位置不具有 PM 標記，最後根據這份檔案統計出 PM 的 confusion matrix 並計算 F-measure score。

5.3.4 辨認結果分析

5.3.4.1 OOV 的分析

首先我們必須統計測試音檔中 OOV (out of vocabulary)的數量，統計後得到 OOV rate 為 **4.28%**；然而，除了統計 OOV rate 外，我們也會針對 OOV 所造成的影響作分析，對每一個屬於 OOV 的 word 而言，它有可能對辨認結果造成取代性錯誤或是額外的插入性錯誤(如下辨認結果範例所示)，這些都是值得進一步統計，我們將針對最終加入 prosodic break 及 prosodic state 資訊後的辨認結果來進行分析，結果發現，4.28%的 OOV rate 會造成約 **8.07%**的錯誤率。

辨認結果範例:

(a)…牙醫師 公會 理事長 郭振興…

(b)…牙醫師 公會 理事 張國政 新…

(c) ...牙醫師 B2-2 公會 B0 理事長 B3 或(or) B2-2 真心 B3…

上述是一個因 OOV 影響辨認率的範例：(a)文本標準答案，(b)baseline system 的辨認結果，(c)加入韻律資訊(prosodic break + prosodic state)後的辨認結果。

接下來，根據前述章節 2.2 所統計過的數據已知，baseline lattice 中的 word coverage rate 為 90.80%，扣掉 4.28%的 OOV rate，剩下來的就是辨認結果的 upper bound，約為 86.52%，而目前的辨認率為 79.88%，也就表示還有約 6.64%的提升空間；因此，為了大幅度提升辨認率，如何有效解決 OOV 的問題便是未來值得進一步研究的課題。

5.3.4.2 針對韻律邊界停頓標記及音節韻律狀態的分析

我們將帶有韻律邊界停頓標記(尤其是針對 B2-1、B2-2、B2-3、B3、B4 等長停頓)及音節韻律狀態的 word 辨認結果輸出，觀察其合理性，如下圖 5.7 範例所示：

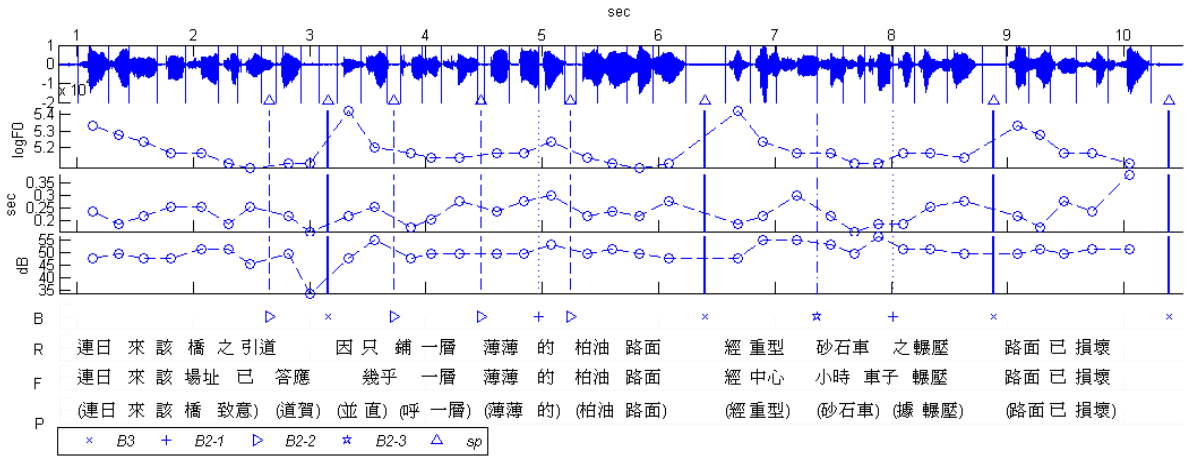


圖 5.7：一片斷語句之辨認結果範例

上圖 5.7 中八列欄位分別代表語音信號波形、音節韻律狀態影響因子加上音節 log-F0 值之總體平均值(global mean)、音節長度、音節能量大小、韻律邊界停頓 (B)、標準答案文本 (R)、baseline 系統之辨認結果 (F) 及加入所有韻律資訊後之辨認結果 (P)。

5.3.4.3 各級辨認結果之比較

以下我們將辨認結果中所改善的搶詞情況、一次詞辨認及聲調修正的部分分別列出。下頁各表格中，第一欄是正確文本，第二欄為 baseline 系統(加入 joint syntax model)的辨認結果，第三欄則是加入韻律模型(加入 prosodic break 及 prosodic state 資訊)後的辨認結果，並且將解碼出針對詞內最後一個音節所連接的韻律邊界停頓標示在右。

表 5.10：搶詞狀況的改善

Ref.	Baseline.rec	Rescored.rec
撞車	撞車	撞車(B0 or B1)
事件	事件	事件(B3)
所幸	所幸	所幸(B2-1)
未	尾盤	未(B0 or B1)
翻覆	福和	翻覆(B3)
或	我國	或(B0 or B1)
惡意	已	惡意(B3)
隱瞞	溢滿	隱瞞(B0 or B1)
造成	造成	造成(B2-1)
而	NULL	而(B2-1)
四歲	二十歲	四歲(B0 or B1)
以上	以上	以上(B0 or B1)
則	則	則(B2-1)
有	有	有(B2-2)
牙醫師	牙醫師	牙醫師(B2-2)
公會	公會	公會(B0 or B1)
理事長	理事	理事長(B3)
郭振興(OOV)	張國政	或(B2-2)
	新	真心(B3)

執勤 的 特殊 族群 郭振興(OOV)	執勤 的 特殊 處 尋獲 真心	執勤(B0 or B1) 的(B2-1) 特殊(B0 or B1) 族群(B3) 或(B0 or B1) 真心(B0 or B1)

表 5.11：一字詞辨認的改善

Ref.	**Baseline.rec**	**Rescored.rec**
任何 買 受 的 一 方 去 承擔	任何 NULL 理首 的 方 去 承擔	任何(B2-1) 買(B0 or B1) 受(B0 or B1) 的(B0 or B1) 一(B2-2) 方(B0 or B1) 去(B2-1) 承擔(B3)
例如 我 管 你 是 為了 這個	例如 NULL NULL 觀念 是 為了 這個	例如(B0 or B1) 我(B0 or B1) 管(B2-1) 你(B0 or B1) 是(B2-1) 為了(B2-1) 這個(B0 or B1)

家	家	家(B0 or B1)
好	好	好(B3)
賭博	NULL	賭博(B2-1)
新	不過	新(B0 or B1)
方法	警方	方法(B3)
且	發現	且(B0 or B1)
已	已	已(B0 or B1)
漸漸	漸漸	漸漸(B0 or B1)
蔓延到	蔓延到	蔓延到(B2-1)
中市	中市	中市(B0 or B1)
各	各	各(B0 or B1)
階層	階層	階層(B3)
增加	增加	增加(B0 or B1)
了	了	了(B0 or B1)
聯賽	連載	連載(B0 or B1)
的	的	的(B0 or B1)
成績	NULL	成績(B3)
但	NULL	但(B0 or B1)
因	績單	因(B0 or B1)
未	因為	未(B0 or B1)
特別	特別	特別(B0 or B1)
告知	告知	告知(B0 or B1)

所以	所以	所以(B0 or B1)
願	原	願(B0 or B1)
將	想	將(B0 or B1)
跟隨	跟隨	跟隨(B0 or B1)
他	他	他(B0 or B1)
多年	多年	多年(B0 or B1)

表 5.12：聲調修正

Ref.	Baseline.rec	Rescored.rec
影響 因素	影響 應 屬	影響(B0 or B1) 因素(B3)
台北縣 的 執政 成了	台北縣 的 紀政 成了	台北縣(B0 or B1) 的(B0 or B1) 執政(B3) 成了(B2-2)
因此 彩券 發行	一致 裁決 發行	因此(B3) 彩券(B0 or B1) 發行(B2-1)
儘管 忽略 絕大部分	警官 忽略 絕大部分	儘管(B3) 忽略(B2-2) 絕大部分(B0 or B1)

人 的	人 的	人(B2-2) 的(B2-2)
優異 之 士 空降	NULL 友誼 只是 空	優異(B0 or B1) 之(B0 or B1) 士(B3) 空降(B0 or B1)

5.3.4.4 各級 lattice 複雜度總整理

本研究之測試語料一共分成 226 段長句音檔，其中共包含 19 位語者，總長度約為 2 小時，總詞彙數量為 14993，每句包含的音節數平均為 117.2 個音節。以下表格 5.13 將整理出本研究第一級辨認器中的 bigram word lattice、trigram word lattice 以及圖 4.4 中第二級辨認器裡三個階段作展開後(如章節 4.1.2.2 所述)的 lattice 之複雜度(包含每句音檔的平均 node 及 arc 數)和對測試語料重新評分所需的所有時間。表格中 FLM lattice 代表第二級辨認器裡第一個階段所使用之 lattice；Prosodic lattice 1 代表第二級辨認器裡第二個階段所使用之 lattice；Prosodic lattice 2 代表第二級辨認器裡第三個階段所使用之 lattice。

表 5.13：實驗中各層級之 lattice 複雜度

	平均 node 數	平均 arc 數	Rescoring 所需時間
Bigram lattice	8017	76085	0.39min
Trigram lattice	33056	306799	2.02min
FLM lattice	132113	4892790	5.34hr
Prosodic lattice 1	139405	5130643	5.79hr
Prosodic lattice 2	269557	9907172	13.20hr

第六章 結論與未來展望

6.1 結論

本研究提出一套利用系統化韻律建模方式訓練出的 12 種韻律模型來改善語音辨認效能，其中包含從測試音檔中解碼出除詞以外的更多資訊，實驗辨認結果證實了本研究之有效性，下將歸納出其優點。第一，本研究所使用的 12 種韻律模型皆是利用大量未經人工標記的語料訓練出，不但節省了大量人力成本，同時也避免了因人工標記造成的不一致性；第二，本研究所使用的 12 種韻律模型清楚描述了韻律標記與 4 階層式韻律架構、文本中之語言參數及語音信號中韻律聲學參數之間的關係，實驗結果[14]顯示這 12 種韻律模型中的各種參數皆有其意義存在；第三，利用本研究中 two-stage 辨認器可以明顯改善由傳統 HMM 辨認器所產生之聲調辨認率及詞辨認率；第四，本研究整套辨認流程除了最終會解碼出詞、詞性和標點符號等多種語言參數外，同時還解碼出韻律邊界停頓標記和音節韻律狀態等資訊，用以建構出測試音檔之階層式韻律架構。

6.2 未來展望

從本研究可以延伸出四項議題值得未來進一步探討。第一，經由本研究之辨認結果作分析後發現，詞錯誤率深受 OOVs 的影響，而絕大部分的 OOVs 屬於人名，所以未來可以多加入一個專屬人名的語言模型來解決此一問題；第二，本研究中尚未利用到高層次的語言參數，例如詞綴、片語或語法等，未來可以針對這些高層次語言參數重新設計韻律模型，待加入到辨認系統後除了有望再次提升效能外，同時還能解碼出各測試音檔之語法架構；第三，將本研究運用至其他語言(例如英文)；第四，目前本研究由於只對朗讀式語音作辨認，未來若能延展到更貼近生活化的自發性語音，相信對語音辨認領域相信又是一重要研究貢獻。

參考文獻

【1】 S. Ananthakrishnan and S. Narayanan, “Unsupervised adaptation of categorical prosody models for prosody labeling and speech recognition,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 138-149, Jan. 2009.

【2】 S. Ananthakrishnan and S. Narayanan, “Improved speech recognition using acoustic and lexical correlates of pitch accent in a n-best rescoring framework,” in *Proc.of ICASSP 2007*, pp. IV-873-IV876.

【3】 S. Ananthakrishnan and S. Narayanan, “Prosody-enriched lattices for improved syllable recognition,” in *Proc. INTERSPEECH* 2007, pp. 1813-1816.

【4】 K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, S.-S. Kim, J. Cole, and J.-Y. Choi, “Prosody dependent speech recognition on radio news corpus of American English,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14 no.1, pp.232-245, January 2006.

【5】 D. H. Milone and A. J. Rubio, “Prosodic and accentual information for automatic speech recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 11, no. 4, pp. 321-333, July 2003.

【6】 D. Vergyri, A. Stolcke, V. R. R. Gadde, L. Ferrer, and E. Shriberg, “Prosodic knowledge sources for automatic speech recognition,” in *Proc. ICASSP* 2003, pp. I-208-I-211.

【7】 M. Ostendorf, I. Shafran, and R. Bates, “Prosody models for conversational speech recognition,” in Proc. 2nd Plenary Meeting Symp. Prosody and Speech Process 2003, pp. 147-154.

【8】 X. Lei and M. Ostendorf, “Word-level tone modeling for Mandarin speech recognition,” in *Proc. ICASSP* 2007, pp. IV-665-IV-668.

【9】 C. Ni, W. Liu, and B. Xu, “Improved large vocabulary Mandarin speech recognition using prosodic and lexical information in maximum entropy framework,” in *Proc. of CCPR* 2009.

【10】 W.-J. Wang, Y.-F. Liao, and S.-H. Chen, “Prosodic modeling of Mandarin speech and its application to lexical decoding,” in *Proc. EUROSPEECH*, 1999, vol. 2, pp. 743-746.

【11】 Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, “Enriching speech recognition with automatic detection of sentence boundaries and disfluencies,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1526-1540, September 2006.

【12】 E. Shriberg and A. Stolcke, “Prosody modeling for automatic speech recognition and understanding,” in *Proc. workshop on mathematical foundations of natural language modeling* 2002.

【13】 K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, “ToBI: A standard for labeling English prosody,” in *Proc. ICSLP*, 1992, vol. 2, pp. 867-870.

【14】 C.-Y. Chiang, S.-H. Chen, H.-M. Yu, and Y.-R. Wang, “Unsupervised joint prosody labeling and modeling for Mandarin speech,” *Journal of the Acoustic Society of America*, vol. 125, no. 2, pp.1164-1183, Feb. 2009.

【15】 C.-Y. Tseng, S.-H. Pin, Y.-L. Lee, H.-M. Wang, and Y.-C. Chen, “Fluent speech prosody: Framework and modeling,” Speech Communication, 46, pp. 284-309, 2005.

【16】 L. Breiman, J. Friedman, R. Olshen, and C. Stone, “Classification and Regression Tree,” Wadsworth, Belmont, 1984.

【17】 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, September 1990.

【18】 J. A. Bilmes and K. Kirchhoff, “Factor language models and generalized parallel backoff,” in *Proc. of HLT/NACCL*, 2003, pp. 4-6.

【19】 A. Stolcke, “SRILM – An extensible language modeling toolkit,” in *Proc.* ICSLP, 2002.

【20】 P. Beyerlein, “Discriminative model combination,” in *Proc*. *ICASSP* 1998, pp. 481-484.

【21】 B.-H. Juang, W. Chou, C.-H. Lee, “Statistical and discriminative methods for speech recognition”, in *Speech Recognition and Coding - New Advances and Trends*, ed. A.J. Rubio Ayuso, J.M. Lopez Soler, Springer-Verlag, Berlin-Hheidelberg, 1995.

【22】 Mandarin microphone speech corpus – TCC300, http://www.aclclp.org.tw/use_mat.php#tcc300edu.

【23】 “HTK Web-Site”, http://htk.eng.cam.ac.uk. Accessed 2009

【24】 L.R. Bahl, R. F. Brown, P. V. de Souza, and R.L. Mercer, “Maximum mutual information estimation of hidden markov model parameters for speech recognition,” in *Proc. ICASSP* 1986, pp. 49-52.

【25】 C.-R. Huang, K.-J. Chen, F.-Y. Chen, Z.-M. Gao and K.-Y. Chen. 2000, “Sinica treebank: design criteria, annotation guidelines, and on-line interface”, in *Proceedings of 2nd Chinese Language Processing Workshop* 2000, Hong Kong, pp. 29-37.

【26】 S.-H. Chen, W.-H. Lai, and Y.-R. Wang, “A statistics-based pitch contour model for Mandarin speech,” *Journal of the Acoustical Society of America*, vol. 117, no. 2, pp. 908–925, February 2005

【27】 S.-H. Chen, W.-H. Lai, and Y.-R. Wang, “A new duration modeling approach for Mandarin speech,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 11, no. 4, pp. 308–320, July 2003.

【28】 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, 2008

【29】 C.-Y. Tseng, S.-H. Pin, Y.-L. Lee. H.-M. Wang, and Y.-C Chen, “Fluent speech prosody:Framwork and modeling,” Speech Commun. Special issue on quantitive prosody modeling for natural speech description and generation, 46, 284-309 2005

【30】 F. Sha and F. Pereira. Shallow parsing with conditional random fields.

【31】 周建邦, “中文大詞彙語音辨認知語言模型改進”, 國立交通大學碩士論文, 民國九十八年十二月。

【32】 張皓翔, “使用階層式韻律模型於豐富中文語音辨認”, 國立交通大學碩士論文, 民國九十九年八月。

附錄一：決策樹之問題集

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

' Is the inter-syllable location an utterance boundary?'

' Is the inter-syllable location an interword?'

' Does a PM exist at the inter-syllable location'

' Does a Major PM exist at the inter-syllable location '

' Does a。exist at the inter-syllable location '

' Does a，exist at the inter-syllable location '

' Does a、exist at the inter-syllable location '

' Does a．exist at the inter-syllable location '

' Does a；exist at the inter-syllable location '

' Does a：exist at the inter-syllable location '

' Does a？exist at the inter-syllable location '

' Does a！exist at the inter-syllable location '

' Does a（exist at the inter-syllable location '

' Does a）exist at the inter-syllable location '

' Is the the preceding special prefix words + special 1-syllable words: Ng, Ncd, Di, DE, I, T'

' Is the POS of the preceding word A'

' Is the POS of the preceding word C'

' Is the POS of the preceding word D'

' Is the POS of the preceding word N'

' Is the POS of the preceding word I or T'

' Is the POS of the preceding word P'

' Is the POS of the preceding word V'

' Is the POS of the preceding word DE'

' Is the POS of the preceding word SHI'

' Is the POS of the preceding word FW'

' Is the POS of the preceding word DM'

' Is the POS of the preceding word Da Di Dk D'

' Is the POS of the preceding word Dfa'

' Is the POS of the preceding word Dfb'

' Is the POS of the preceding word Na Nb Nc Nv'

' Is the POS of the preceding word Nd'

' Is the POS of the preceding word Neu Nes Nep Neqa Neqb Nf'

' Is the POS of the preceding word Ng Ncd'

' Is the POS of the preceding word Nh'

' Is the POS of the preceding word VA VAC VG'

' Is the POS of the preceding word VB VC VCL VD VE VF VJ VK VL'

' Is the POS of the preceding word VH VHC VI'

' Is the POS of the preceding word V_2'

' Is the POS of the preceding word Caa'

' Is the POS of the preceding word Cab'

' Is the POS of the preceding word Cba'

' Is the POS of the preceding word Cbb'

' Is the POS of the preceding word Da'

' Is the POS of the preceding word Di'

' Is the POS of the preceding word Dk'

' Is the POS of the preceding word D'

' Is the POS of the preceding word Na'

' Is the POS of the preceding word Nb'

' Is the POS of the preceding word Nc'

' Is the POS of the preceding word Ncd'

' Is the POS of the preceding word Neu'

' Is the POS of the preceding word Nes'

' Is the POS of the preceding word Nep'

' Is the POS of the preceding word Neqa'

' Is the POS of the preceding word Neqb'

' Is the POS of the preceding word Nf'

' Is the POS of the preceding word Ng'

' Is the POS of the preceding word Nv'

' Is the POS of the preceding word I'

' Is the POS of the preceding word T'

' Is the POS of the preceding word VA'

' Is the POS of the preceding word VAC'

' Is the POS of the preceding word VB'

' Is the POS of the preceding word VC'

' Is the POS of the preceding word VCL'

' Is the POS of the preceding word VD'

' Is the POS of the preceding word VE'

' Is the POS of the preceding word VF'

' Is the POS of the preceding word VG'

' Is the POS of the preceding word VH'

' Is the POS of the preceding word VHC'

' Is the POS of the preceding word VI'

' Is the POS of the preceding word VJ'

' Is the POS of the preceding word VK'

' Is the POS of the preceding word VL'

' Is the length of the preceding word 1'

' Is the length of the preceding word 2'

' Is the length of the preceding word 3'

' Is the length of the preceding word 4'

' Is the length of the preceding word 5'

' Is the length of the preceding word 6'

' Is the length of the preceding word less than 2'

' Is the length of the preceding word less than 3'

' Is the length of the preceding word less than 4'

' Is the length of the preceding word less than 5'

' Is the length of the preceding word less than 6'

' Is the following special 1-syllable words: Ng, Ncd, Di, DE, I, T + special suffix words'

' Is the POS of the following word A'

' Is the POS of the following word C'

' Is the POS of the following word D'

' Is the POS of the following word N'

' Is the POS of the following word I or T'

' Is the POS of the following word P'

' Is the POS of the following word V'

' Is the POS of the following word DE'

' Is the POS of the following word SHI'

' Is the POS of the following word FW'

' Is the POS of the following word DM'

' Is the POS of the following word Da Di Dk D'

' Is the POS of the following word Dfa'

' Is the POS of the following word Dfb'

' Is the POS of the following word Na Nb Nc Nv'

' Is the POS of the following word Nd'

' Is the POS of the following word Neu Nes Nep Neqa Neqb Nf'

' Is the POS of the following word Ng Ncd'

' Is the POS of the following word Nh'

' Is the POS of the following word VA VAC VG'

' Is the POS of the following word VB VC VCL VD VE VF VJ VK VL'

' Is the POS of the following word VH VHC VI'

' Is the POS of the following word V_2'

' Is the POS of the following word Caa'

' Is the POS of the following word Cab'

' Is the POS of the following word Cba'

' Is the POS of the following word Cbb'

' Is the POS of the following word Da'

' Is the POS of the following word Di'

' Is the POS of the following word Dk'

' Is the POS of the following word D'

' Is the POS of the following word Na'

' Is the POS of the following word Nb'

' Is the POS of the following word Nc'

' Is the POS of the following word Ncd'

' Is the POS of the following word Neu'

' Is the POS of the following word Nes'

' Is the POS of the following word Nep'

' Is the POS of the following word Neqa'

' Is the POS of the following word Neqb'

' Is the POS of the following word Nf'

' Is the POS of the following word Ng'

' Is the POS of the following word Nv'

' Is the POS of the following word I'

' Is the POS of the following word T'

' Is the POS of the following word VA'

' Is the POS of the following word VAC'

' Is the POS of the following word VB'

' Is the POS of the following word VC'

' Is the POS of the following word VCL'

' Is the POS of the following word VD'

' Is the POS of the following word VE'

' Is the POS of the following word VF'

' Is the POS of the following word VG'

' Is the POS of the following word VH'

' Is the POS of the following word VHC'

' Is the POS of the following word VI'

' Is the POS of the following word VJ'

' Is the POS of the following word VK'

' Is the POS of the following word VL'

' Is the length of the following word 1'

' Is the length of the following word 2'

' Is the length of the following word 3'

' Is the length of the following word 4'

' Is the length of the following word 5'

' Is the length of the following word 6'

' Is the length of the following word less than 2'

' Is the length of the following word less than 3'

' Is the length of the following word less than 4'

' Is the length of the following word less than 5'

' Is the length of the following word less than 6'

Is the initial of the following syllable a null one or in { m, n, l, r}?

Is the initial of the following syllable a null one or in { b, d, g}?

Is the initial of the following syllable a null one or in { f, s, sh, h}?

Is the initial of the following syllable a null one or in { c, ch, q}?

Is the initial of the following syllable a null one or in { p, t, k}?

Is the initial of the following syllable a null one or in { z, zh, j}?