國 立 交 通 大 學

電信工程研究所

碩士論文

使用階層式韻律模型於豐富中文語音辨認

Enriching Mandarin Speech Recognition by Incorporating a Hierarchical Prosody Model

研 究 生：張皓翔

指導教授：陳信宏 博士

中華民國九十九年八月

使用階層式韻律模型於豐富中文語音辨認

Enriching Mandarin Speech Recognition by Incorporating a Hierarchical Prosody Model

研究生：張皓翔　　Student：Hao-Hsiang Chang

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

國 立 交 通 大 學

電 信 工 程 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Communication Engineering

College of Electrical and Computer Engineering

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master

in

Communication Engineering

August 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年八月

使用階層式韻律模型於豐富中文語音辨認

研究生：張皓翔　　指導教授：陳信宏　博士

國立交通大學電信工程研究所

中文摘要

人類平常利用語音交換資訊時，語者的聲調高低，抑揚頓挫，這些表現通稱為韻律現象，本研究提出一個新的語音辨認方法，試圖整合這些韻律資訊於語音辨認上，藉由建立階層式韻律模型，利用聲學及語言參數來幫助預估韻律邊界停頓，並且能夠於辨認結果上標記出語言參數及韻律邊界標記，如此能夠幫助我們更容易閱讀辨識結果。實驗的架構採取兩階段式的方法，重新計分時需將每個參與解碼的模型給予適當的權重，挑選出辨識率最高的詞串，本研究利用 DMC 的方法調整，藉以找出最佳的權重分布，實驗語料庫為 TCC300，最後實驗結果顯示加入了韻律模型的詞彙辨認率與基本系統相比提升了 1.67%。

Enriching Mandarin Speech Recognition by Incorporating a Hierarchical Prosody Model

Student：Hao-Hsiang Chang Advisor：Dr. Sin-Horng Chen

Institute of Communication Engineering
National Chiao Tung University

Abstract

This thesis presents a probabilistic model for incorporating hierarchical prosody in the speech recognition task, for improving word recognition directly and for enriching speech recognition output. The model includes higher level linguistic cues (syllable, word, punctuation mark, and part of speech), intermediate prosodic break representation, and prosodic-acoustic feature correlated with break type and linguistic cues. Moreover, our speech recognition system produces not only word sequences but also prosodic label and linguistic information in order to enrich speech recognition output for downstream natural language processing module. We adopted a two-stage rescoring framework to implement our approach, and discriminative model combination method is used for rescoring. We evaluate our approach on TCC300 corpus, and results show that the performance of prosodic model is better than the baseline system. We obtain a 1.67% absolute improvement in word error rate over the baseline system on a read speech task.

keyword: speech recognition, hierarchical prosody, discriminative model combination

誌謝

兩年的研究生活轉眼即逝，能在最後的一年內將論文完成，最主要感謝指導教授陳信宏老師以及王逸如老師，兩位老師教導我對於研究的方法還有態度，使我在研究中不會迷失，能夠一步步確立方向。

本論文的完成還須感謝實驗室的大家，從碩一開始帶領我的合哥，研究上的許多問題還有最後論文的修改，都麻煩合哥許多，提供韻律模型的性獸，在韻律這方面的問題都有賴於獸哥的幫助，最佳助教黃信德，畢業前夕愛唬人的輝哥，開公司的巴金叔，愛護地球的希群。還有已經畢業了的學長們，一起聽 TB 的 Q 哥，一起看棒球的普烏，一起 KGB 的小宋，一起打籃球的小帥哥，感謝鬥陣奮鬥拼畢業的同學們，從碩一就很認真，研究課業都很強的一哥宥余，每天研究做很晚的承燁，對研究的熱忱讓人佩服，不會說客語但是做客語辨識很厲害的小卡，研究只差臨門一腳很愛嗆我的誌宏北鼻，很多外務還能夠準時畢業的依玲，已經有重聽現在又快瞎了的舒姊。還有學弟們，感謝你們為實驗室注入新的熱情與活力，胖胖、啟全、豆腐、彥邦、大胖、智障，以及新加入同一個研究主題的銘傑，你對研究的積極和求學的精神也幫助我非常多。

謝謝我的女友佩茹，總是在我迷惘時給我最大的鼓勵，最後感謝我的父母、哥哥，謝謝你們適時的關心和對我的信任與支持，才能讓我堅持下去完成論文。

目錄

圖目錄

表目錄

第一章 緒論

1.1 研究動機

近年來隨著科技的進步，為了追求更人性化的介面，除了滑鼠點擊或觸控螢幕的技術外，語音辨認科技的發展對人類未來生活也會變的更加重要，目前的語音辨認技術已有相當程度的進展，例如資料驅使(Data-driven)方法、聲學模型和語言模型建立方式，以及基於動態編輯程序(Dynamic Programming-based)之搜尋方法等[1]，經由訓練得當的聲學模型跟語言模型，輔以適當的搜尋演算法，辨識結果都有不錯的效能。

人在講話時會經過腦內組織後，再將一句話完整的說出，其中語調的高低起伏、抑揚頓挫往往帶有語者欲表達的意念，這些現象都包含在我們所要探討的韻律(Prosody)範圍內。在語音合成方面考慮了韻律資訊的關係，其品質有顯著的改善，因此本研究將從語音辨識的角度去探討加入了韻律資訊後，對詞的辨識還有對語意的了解是否有提升的效果，並且將韻律邊界以階層式的方式表現，以 TCC300 多語者的語料為對象，對語音辨識做進一步的探討與研究。

1.2 文獻回顧

近年將韻律資訊利用至語音辨認相關研究主要分為三類，第一類為利用韻律參數對辨認結果所產生之詞格(word lattice)重新計算分數，直接利用韻律參數來驗證(verification)在詞格中不同路徑其對應切割位置之可靠程度[2]；第二類則為以事件為基礎(event-based)的方式增加語音辨認之效能[3]，利用韻律參數建立一個偵測事件之模型，例如：類語句邊界（sentence-like unit）或詞語修補中斷點，並利用事件及詞的序列一起建立語言模型，對辨認結果所產生之詞格重新計算分數；第三類則是利用韻律以及句法的關係建立韻律相關的語言模型（prosody dependent LM），來描述韻律以及詞之結合機率，並利用韻律邊界的資訊建立韻律相關的聲學模型（prosody dependent AM）[4][5]。

1.3 研究方向

由過去研究觀察發現到中文語音辨識的一些問題：

- 中文一字詞數量多，混淆度高，在辨識時容易造成錯誤。
- 詞邊界判斷錯誤造成搶詞問題，使辨認率下降。

本論文主要的研究方向在於探討如何將一個階層式的韻律模型，加入既有的基本辨識系統中，藉以提升辨識效能，並試圖改善上述的錯誤類型，其中基本辨識系統中的聲學模型使用隱藏式馬可夫模型(Hidden Markov Model, HMM)，以音節(Syllable)為模型單位，語言模型則是經由一個大量的文字資料庫統計訓練而出，而韻律模型的部分則是利用[6]所提出之非監督式中文自發性語音韻律標記結果訓練得出，本論文主要探討音節間的韻律停頓(prosodic break)對中文語音辨識的影響。

為了加入韻律模型分數我們採用兩段式(two pass)語音辨認架構，第一階段經基本辨識系統計算其聲學模型與語言模型分數，產生前 N 名的辨認詞串，第二階段加入韻律模型，由聲學及語言學兩方面加強 prosodic break 的預估，計算詞串中每個音節邊界的分數，將詞串重新排名後，挑選出最後的辨認結果，而最後的結果將可達到豐富標記的效果，因為目前的辨認系統大都以辨認出詞為最後結果，要是我們能夠在詞之外另外標記出標點符號、詞性、韻律邊界的結果，將可大幅提升辨認結果的可讀性。

1.4 章節概要說明

本論文的章節內容分配如下：

第一章：緒論。

第二章：TCC300 基本辨認系統建立。

第三章：漢語基本特性及韻律架構。

第四章：整合階層式韻律模型與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 TCC300 基本辨認系統

近年來相關的語音研究，最常採用的聲學模型是利用隱藏式馬可夫模型，藉由這種機率模型，來描述發音過程的狀態(State)轉移現象和輸出結果，此方法的辨識效能佳，故本系統亦採用此模型，並加入語言模型，藉由語言模型來提升辨識率，本研究使用 TCC300 語料庫建立聲學模型，並由一大量文字語料庫訓練語言模型。

2.1 TCC300 語料介紹

本論文中使用 TCC-300 麥克風語音資料庫是由國立交通大學、國立成功大學、國立台灣大學所共同錄製，中華民國計算語言學學會所發行，此語料庫屬於麥克風朗讀語音，主要目的是為提供語音辨認研究，檔案統計資料如表 2.1 所示。台灣大學語料庫主要包含詞以及短句，文字經過設計，考慮音節與其相連出現之機率，共 100 人錄製而成；成功大學及交通大學為長文語料，其語句內容由中研院提供之 500 萬詞詞類標示語料庫中選取，每篇文章包含數百個字，再切割成 3 至 4 段，每段至多 231 字，分別各 100 人朗讀錄製，且每人所朗讀之文章皆不相同。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：TCC300 語料資料統計表

學校名稱	文章屬性	語者總數		總音節數		檔案總數	
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

本研究從中挑選 1/10 的檔案做為測試音檔，其中包含 29 個語者，15 男 14 女，音節總數為 26472 個，其餘 9/10 部分則做為訓練音檔，音節總數為 300836 個。

2.2 辨認系統架構

語音辨識為輸入欲辨識音檔聲學特徵向量序列 $\mathbf{X}_a$，經辨識器後，輸出相對應最有可能的詞串(Word Sequence)，其基本數學式及系統方塊圖如下：

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

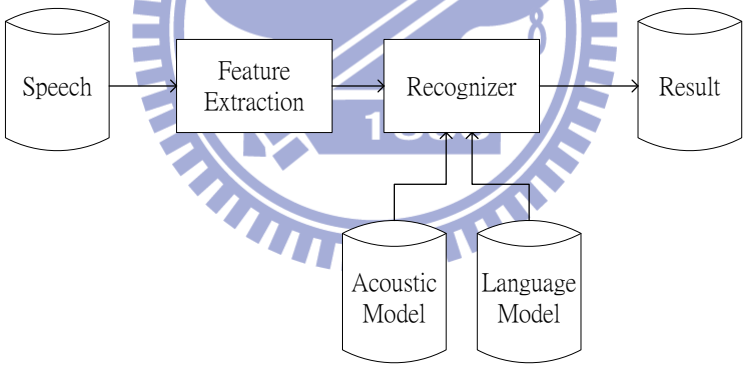


圖 2.1：傳統語音辨識流程圖

其中 $\mathbf{W}$ 代表詞，$\mathbf{X}_a$ 為聲學特徵參數，由最大事後機率法則(Maximum a posterior, MAP)再拆解成兩部分，其中 $P(\mathbf{W})$ 由語言模型(language model，LM)計算得到，$P(\mathbf{X}_a \mid \mathbf{W})$ 則由聲學模型(acoustic model，AM)計算得到。

2.2.1 聲學模型的建立

進行語音辨識系統之訓練、測試，首先的前處理工作就是將語音參數從輸入語音中抽出來。因為語音訊號之短時間穩定特性(Short Term Stationary)，加上考慮到人耳聽覺效應的補償作用，使用的參數為 MFCC(Mel-Frequency Ceptral Coefficients，梅爾倒頻譜參數)。他的成分包括 12 維 MFCC 加上能量共 13 維，取其 Delta 和 Delta-Delta，將參數變化的訊息也提供給辨識器使用，總共 38 維(最後拿掉能量這一維)，系統參數設定如下表：

表 2.2：參數抽取設定檔

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

聲學模型為 411 個音節，每一個音節使用 8 個 state 的 HMM，並且使用 HTK 之 MMI 鑑別性訓練[7]。

2.2.2 語言模型的建立

語言模型可區分為兩種，一種是根據語言的文法、詞性，訂定文章出現一定符合規則之語言模型(Rule-based LM)；另一種則是藉由處理大量的文字資訊，利用統計的方式，計算詞與詞之間的連結規則而建立的語言模型(Statistic-Based LM)。

2.2.2.1 語言模型簡介

事實上所有的語言都有其文法規則，而利用這類文法規則所建立出的機率模型，則稱之為語言模型。若在進行語音辨認時，能將語言模型配合聲學模型共同使用，通常能夠大幅提升辨識系統的效能。在漢語中文的情況下，建立語言模型時，一般是以詞做為基本單位。因為在中文語言中，以「詞」為基本單位建構而成的句子比較符合語言規則，以「環保」這個二字詞來解釋，若將它拆成「環」跟「保」這兩個字(Character)，則不如原本的詞那樣有意義。

2.2.2.2 n-gram 語言模型

假設有一個句子(Sentence)，其內容以詞為單位所組成，總共有 m 個詞，也就是「$w_1,w_2,\ldots,w_m$」，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(Condition Probability)之連乘：

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

但是因為記憶體的大小有限，而且要求得所有詞的條件機率是不可能的，所以若是給予適當的假設，則可以使用 n-gram 的機率去趨近(2.2)式。

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

其中每個 n-gram 的機率，可藉由在大量文章中詞串所累積的出現次數而得，如下式所示：

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

上式中，$Count(\cdot)$ 表示詞串的出現次數。而語言模型也就是由大量的 n-gram 之機率所組合而成。

2.2.2.3 機率的 smoothing

由(2.4)式可以知道，如果在分子的 $Count(\cdot)$ 的值為 0 時，則此 n-gram 的機率會等於 0，但是一個詞串在部分文章中沒有出現過，不代表辨識結果中絕不會有這種組合出現，因此這種情況下給定的機率不合理，而且在消息理論(Information Theorem)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，當 $Count(\cdot)$ 的值很小的時候，所計算出的 n-gram 機率也不準確、信心度不足，所以還必須對利用(2.4)式所計算出的機率做 smoothing，使所有的 n-gram 機率均能被良好的估計，一般常見的 smoothing 方式如下：

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

式中 $a(w_{i-n+1},...,w_{i-1})$ 表示為 back-off 係數，也就是當計算 n-gram 機率所用的詞串出現次數為 0 時，則利用其(n-1)-gram 的機率，再乘上 back-off 係數，這樣便可避免機率 0 的出現，並分配給它一個適當的機率值。而 $a(w_{i-n+1},...,w_{i-1})$ 的選定，還會經過 normalization，令其滿足：

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

而關於 $Count(\cdot)$ 的值很小時所造成的 n-gram 機率不準確的問題，解決方法是當詞串的出現次數小於 k 次時，則將這個 n-gram 機率乘上一個依據 Good-Turning discounting 所計算出隻小於 1 的值 d_a (Discount Coefficient Factor)，以減低其機率(相對於對它的值較沒信心)，並將扣除的這些機率分給詞串沒有出現的 n-gram 機率使用。

2.2.2.4 語言模型訓練流程

利用大量的文字資料訓練出一個涵蓋範圍廣泛，是用於各個領域的語言模型，基於此種模型的普遍性，$a(w_{i-n+1},...,w_{i-1})$稱為「General LM」。要訓練一個好的語言模型，必須要又大量的文字資料庫，本研究使用的文字資料庫有三個：

第一個是光華雜誌(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；其次為 NTCIR，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；最後是中研院平衡語料庫(Sinica)，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，本研究將這三個語料庫簡稱 NSS。

要建立一個完善的語言模型，詞典的選擇是其中一項重要因素。辨識結果只會輸出有收錄在詞典的詞，所以最終目的是希望能將一種語言中所有存在的詞都納入其中，但受限於記憶體大小，我們僅能將較常出現、較重要的詞整理出來包含在詞典之中提供建立 LM 使用。

NSS 先經過 CRF[8]斷詞器斷詞後，再經過文字正規化[9]的處理，得到詞的總數量為 122,541,303 個，字數為 231225705，本研究詞典的選擇方式是用最直接的選詞方法，即是由斷詞結果中統計出各詞的詞頻，並依據詞頻大小來決定詞的重要性，納入詞典中，表 2.3 及圖 2.2 為不同的詞典數量對應的涵蓋率(coverage rate)，表 2.4 則為六萬詞詞典的詞長分佈情況，詞典平均詞長為 1.73 個字，圖 2.3 是六萬詞詞典中詞長分布情況及每種詞長的總共 character 數量，決定詞典後便可依此來訓練語言模型，訓練流程如圖 2.4 所示。

表 2.3：不同詞典數目大小對應的詞與字元涵蓋率

詞典 size	60K	70K	80K	90K	200K
word coverage rate (WCR)	96.7%	97.1%	97.5%	97.8%	99%
character coverage rate (CCR)	94.8%	95.5%	96.0%	96.4%	98.3%

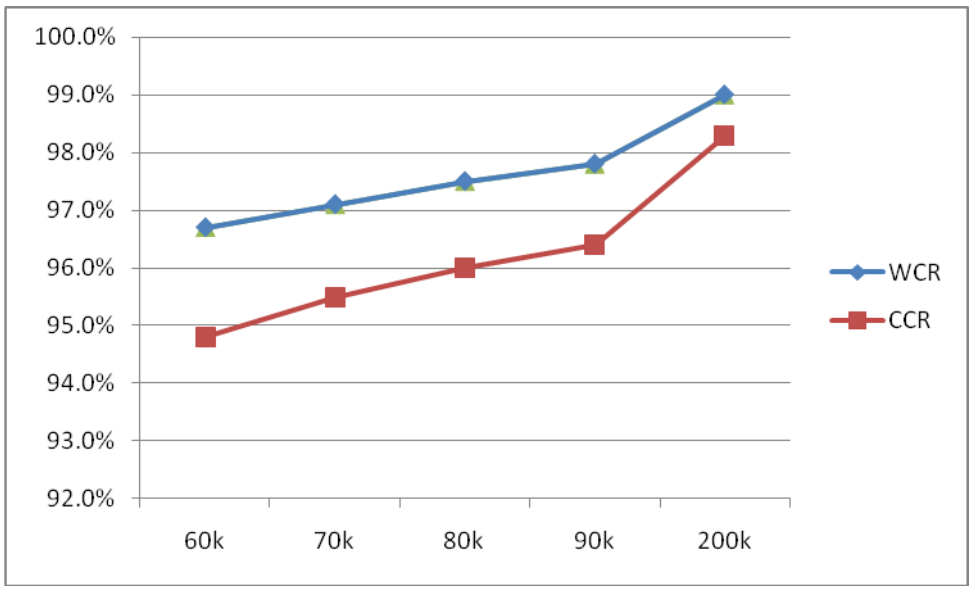


圖 2.2：不同詞典數目大小對應的詞與字元涵蓋率

表 2.4：六萬詞詞典的詞長分佈

詞長	1	2	3	4	5	6	7
word 個數	2,954	37,034	15,810	3,893	270	32	7
character 數量	42374164	133203210	24690870	4431408	493225	39258	7245

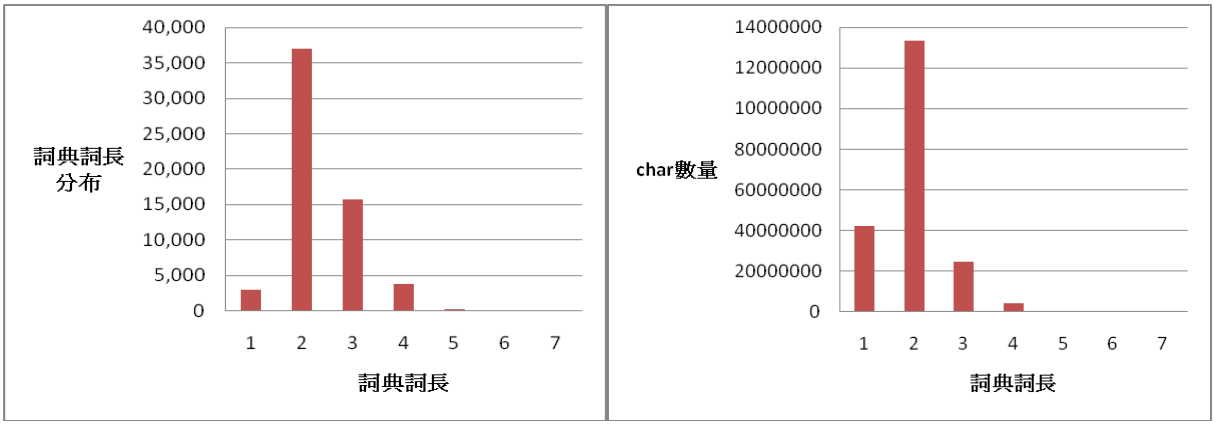


圖 2.3：六萬詞詞典詞長分佈及 character 數量

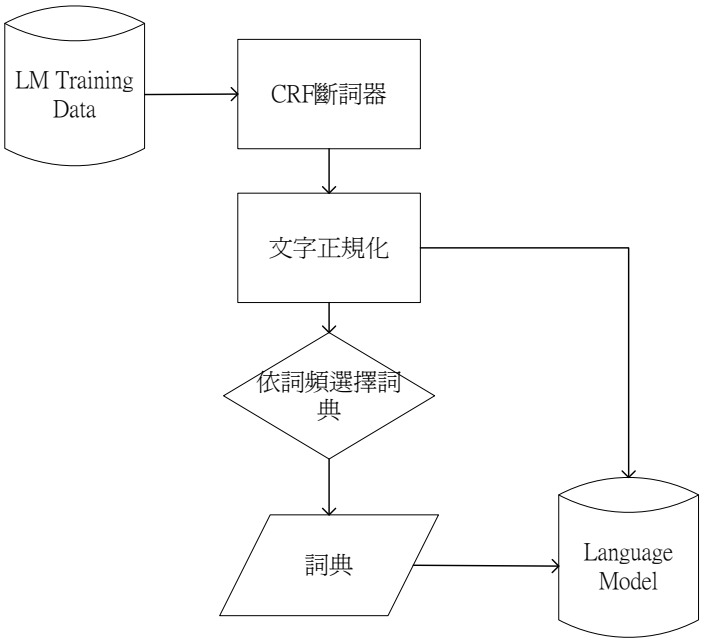


圖 2.4：語言模型訓練流程圖

在詞典中，我們對每個中文詞的表式方式均採用「Big5 碼_漢語拼音」，例如「電纜」這一個二字詞，在詞典中表示格式為「B971C66C_dian4lan3」。

2.2.2.5 Perplexity 計算

評估語言模型通常是以計算其混淆度(perplexity, ppl)來判斷。混淆度是根據消息理論(information theory)而得，如下式：

$$H = -\frac{1}{m}\log P\left(W = w_1, w_2, \ldots, w_m\right) \tag{2.7}$$

上式為一個詞串 $W = w_1, w_2, \ldots, w_m$，對於每個新詞提供的平均資訊量(entropy)，經過 ergodic 的假設與適當化簡而得。而混淆度可直接使用(2.7)式進一步定義為：

$$PP = \exp\left(H\right) \tag{2.8}$$

若 $P\left(W = w_1, w_2, \ldots, w_m\right) = \prod_{i=1}^{m} P\left(w_i \mid w_1, w_2, \ldots, w_{i-1}\right)$ 則可發現，混淆度就是 $P\left(w_i \mid w_1, w_2, \ldots, w_{i-1}\right)$ 的幾何平均數的倒數。因此混淆度可以解讀為語言模型估測一個歷史詞串後面，平均可能的

可接詞數；混淆度越高，表示一個歷史詞串後接詞有較多的選擇，辨認時就越難找到確切的答案；反之，則較易找到正確答案。

以下我們分別利用不同的測試文章，經過混淆度的計算來評估先前所訓練出來的語言模型效能：

Inside test of LM

測試文章：部分 NSS 檔案

1377855 sentences, 44070460 words

Database	lexicon size	n-gram order	smoothing	logprob	ppl	ppl1
NSS	60K	2	GT	-1.17887e+08	392.53	473.12
NSS	60K	3	GT	-1.01961e+08	175.16	205.87

Outside test of LM

測試文章：TCC300 測試語料

845 sentences, 17921 words

Database	lexicon size	n-gram order	smoothing	logprob	ppl	ppl1
NSS	60K	2	GT	-52179.7	626.62	848.97
NSS	60K	3	GT	-49336.5	434.68	579.45

測試文章: TCC300 訓練語料

8037 sentences, 172562 words

Database	lexicon size	n-gram order	smoothing	logprob	ppl	ppl1
NSS	60K	2	GT	-52179.7	637.77	861.57
NSS	60K	3	GT	-475107	427.29	566.56

註：

GT:：Good Turing algorithm

ppl = 10^(-logprob / (words - OOVs + sentences))
ppl1 = 10^(-logprob / (words - OOVs))

其中 words 表示總共詞的數量，OOVs 表示 out-of-vocabulary 數量，sentences 表示為測試的句子數量。因為對數機率值 logprob 包含</s>句子結束符號，所以平均每個 word 的 perplexity 是以 ppl 為計算方式，若排除</s>句子結束符號的話，則 perplexity 以 ppl1 為計算方式。

2.3 實驗結果

接著我們利用 2.2.1 節所建立的聲學模型來對測試音檔做辨識，藉以評估聲學模型的效能，所以本研究先建立一個 free-gram 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 2.5 所示：

表 2.5：音節辨識率(free-gram)

TCC300 outside	73.39%
TCC300 inside	85.74%

實驗數據顯示在 outside test 部分音節辨認率可以達到 73.39%，inside test 更是提升至 85.74%，所以聲學模型的效能已經到達一個不錯的水準。

表 2.6 為加入 2.2.2 節所建立的語言模型詞辨認實驗結果，由於語料庫中音檔長短的屬性不同，所以分析結果時我們特別將依據錄製音檔學校及男女分開統計，並輸出前 100 名辨認結果，觀察其涵蓋率的分布情況：

表 2.6：TCC 300 辨認結果分析 (依據學校、男女)

	N=1	N=100
台大 男性 338 句音檔 共 1609 個詞 音檔平均詞彙數量 4.7	38.78%	62.09%
台大 女性 281 句音檔 共 1321 個詞 音檔平均詞彙數量 4.7	48.90%	70.55%
成大 男性 63 句音檔 共 3774 個詞 音檔平均詞彙數量 59.9	62.75%	67.46%
成大 女性 53 句音檔 共 3648 個詞 音檔平均詞彙數量 68.8	75.14%	79.30%
交大 男性 49 句音檔 共 3264 個詞 音檔平均詞彙數量 66.6	64.55%	68.41%
交大 女性 61 句音檔 共 4304 個詞 音檔平均詞彙數量 70.5	68.42%	72.47%

上表中 N=1 的欄位代表辨認結果，N=100 的欄位代表輸出的 100 條詞串中，辨認率最高的結果，也就是涵蓋率，經觀察可以發現成大及交大所錄製長句，涵蓋率與辨認率相去不遠，而台大所錄製音檔多是短句，可以看到涵蓋率比起辨認率有相當程度的提升，但是可能錄音品質不佳，造成辨認率相對較低，所以在本研究之後的測試音檔中，會將台大的錄製的音檔拿掉。

由於長文在辨識產生 top N 詞串時，一個小的變化就可能產生一條新的詞串，兩條詞串之間可能僅僅只差了一個音節，這樣的情況造成在產生 top N 詞串時必須將 N 設大，才能達到較好的涵蓋率，所以我們經過一個切短句的機制，將長文辨認音檔斷開成較小段落，使每個短句段落都有其 top N 的辨認結果。斷成短句後，每段短句段落會有 N 種不同的辨認結果，相對於長文，短句的辨識結果包含較多變化，如此一來便可提升涵蓋率，切短句的流程如圖 2.5：

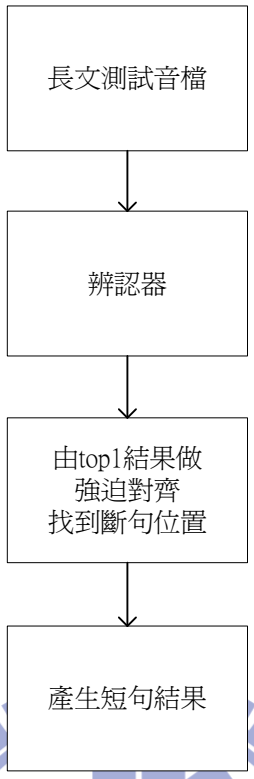


圖 2.5：切短句流程圖

切短句是依據 top 1 辨認結果做強迫對齊後得到的 short pause 長短(在此設定為 20 個 frame，大於 20 個 frame 即切斷)，並保留前後 short pause，依此原則求得每段短句段落的起始與結束位置。抽取特徵參數時，若只先求 13 維的 MFCC，短句邊界的 frame 所抽取 MFCC 的 delta 與 acceleration 係數會與原來長文用的不一致，所以我們先對原先長文音檔求出 39 維的 MFCC，再對 39 維的 MFCC 依短句段落的切割位置做特徵參數抽取。

- **製作短句答案**

利用正確文本對長句音檔做強迫對齊，得到每個詞的起始與結束 frame 位置，判斷這個詞的位置落在哪段短句段落較多，藉此決定詞屬於哪個短句段落，。

表 2.7 為斷成短句後的辨認結果，我們可以看出切完短句後辨認率比起原長句結果降低 0.58%~1.86%，但是相對涵蓋率卻有 7.54%~12.84%大幅的提升，證明了在相同 N 的設定下，短句的確包含了較多變化的輸出結果，讓我們在往後能夠提升空間變大。

表 2.7：TCC 300 短句辨認結果分析 (依據學校、男女)

	N=1	N=100
成大 男性 635 句音檔 共 3774 個詞 音檔詞彙數量 5.94	60.89%	79.04%
成大 女性 442 句音檔 共 3648 個詞 音檔詞彙數量 8.25	73.96%	86.84%
交大 男性 466 句音檔 共 3264 個詞 音檔詞彙數量 7.00	63.97%	81.25%
交大 女性 641 句音檔 共 4304 個詞 音檔詞彙數量 6.71	66.895	82.92%

第三章 漢語基本特性及韻律架構

當我們利用聲音與人溝通時，除了話中詞的意思外，我們說話音調的抑揚頓挫或是音量的高低起伏都會讓對方有不同的感受，這些語音上的變化我們稱為韻律變化，其主要表現在(1)音量大小(energy level)、(2)音高的高低(pitch contour)、(3)說話速度快慢(speaking rate)、(4)停頓時長(pause duration)等因素上，本章會先介紹漢語的語音特性，之後進一步討論漢語階層韻律架構，我們將可知道這些特性如何影響韻律變化，而從階層韻律架構中，我們可以了解韻律變化確實不是隨意所組成。

3.1 漢語語音特性

漢語語音基本單元，是由一個音節搭配一個聲調所組成；漢語語言的基本單位為字，一個字的發音就對應一個音節。漢語的音節結構圖如圖 3.1 所示，音節的結構是由聲母(initial)及韻母(final)所組成，韻母又可以再細分為介音(medial)與韻腳(rime)，而韻腳還可以再分為主要元音(nucleus)和韻尾(coda)，音節可以只由一個主要元音構成，也可以多到由聲母、介音、主要元音跟韻尾組成一個音節，且漢語音節也允許空聲母或是空韻母的情況發生，並非每個音節都有聲母或韻母，漢語的聲母和韻母各有 22 和 40 類，如表 3.1、表 3.2 所示，依其發音特性又可分別分為 7 類和 17 類。一般來說只有單一主要元音的音節會比較短，四個音節成份都備齊的音節長度會比較長；聲母為摩擦音(fricative，如 f, s, sh, h)或為塞擦音(affricate，如 zh, z, j)的音節較長，而聲母為爆破音(如 b, d, g)的音節其長度會較短。

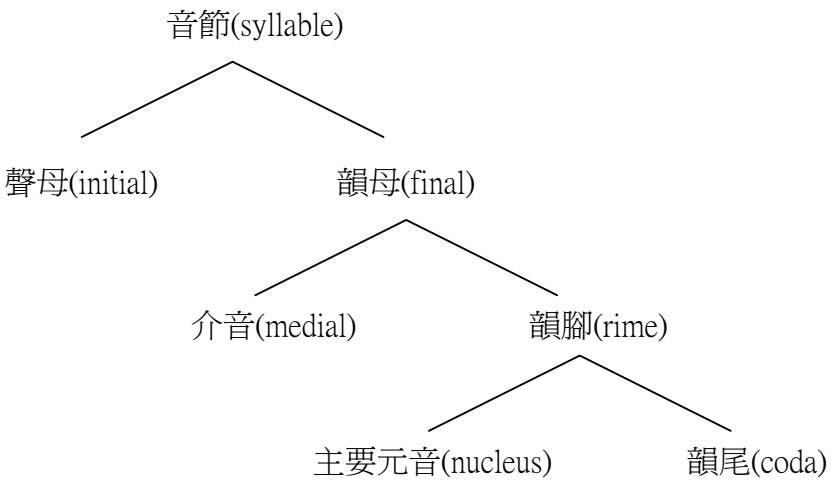


圖 3.1：漢語音節結構圖

表 3.1：聲母分類表，依照發音特性分成 7 類

類別		聲母(Initial)
鼻音_濁音	1	ㄇ、ㄋ、ㄌ、ㄖ
摩擦音_清音	2	ㄈ、ㄒ、ㄕ、ㄏ、ㄙ
爆破音_不送氣	3	ㄅ、ㄉ、ㄍ
塞擦音_不送氣	4	ㄐ、ㄓ、ㄗ
爆破音_送氣	5	ㄆ、ㄊ、ㄎ
塞擦音_送氣	6	ㄑ、ㄔ、ㄘ
	7	空聲母

表 3.2：韻母分類表，依照發音特性分成 17 類

類別	韻母(Final)	類別	韻母(Final)
1	空韻母	10	ㄢ、ㄧㄢ、ㄨㄢ、ㄩㄢ
2	ㄚ、ㄧㄚ、ㄨㄚ	11	ㄣ、ㄧㄣ、ㄨㄣ、ㄩㄣ
3	ㄛ、ㄧㄛ、ㄨㄛ	12	ㄤ、ㄧㄤ、ㄨㄤ
4	ㄜ	13	ㄥ、ㄧㄥ、ㄨㄥ、ㄩㄥ
5	ㄝ、ㄧㄝ、ㄩㄝ	14	ㄧ
6	ㄞ、ㄧㄞ、ㄨㄞ	15	ㄨ
7	ㄟ、ㄨㄟ	16	ㄩ
8	ㄠ、ㄧㄠ	17	ㄦ
9	ㄡ、ㄧㄡ		

相對於語調語言(intonation language)如英語，漢語屬於聲調語言(tonal language)，以聲調可用來區分語意，如「買」跟「賣」二字，只是聲調的不同就足以造成截然不同的意思。漢語的聲調共有五種，分別是一聲、二聲、三聲、四聲和輕聲。聲調是最明顯影響漢語韻律的特徵，其所附帶的資訊最主要呈現在音高軌跡上，圖 3.2 為一聲到四聲的基頻軌跡形狀示意圖，一聲的基頻軌跡為一水平線；二聲的基頻軌跡呈現低到高的趨勢；三聲的基頻軌跡為一勺狀曲線；四聲的基頻軌跡則呈現高到低的趨勢，至於輕聲的基頻軌跡一般會受到前後音節的聲調所影響，沒有固定的形狀，且其音長通常比其他四個聲調短，能量也比較小，綜合這些特性我們可以知道音高軌跡受到聲調顯著的影響。常見的漢語文字大約有 12000 多個，帶聲調的音節組合約有 1300 個，如果扣掉同音字的情形和聲調的部分，可以分類成 411 種基本音節。

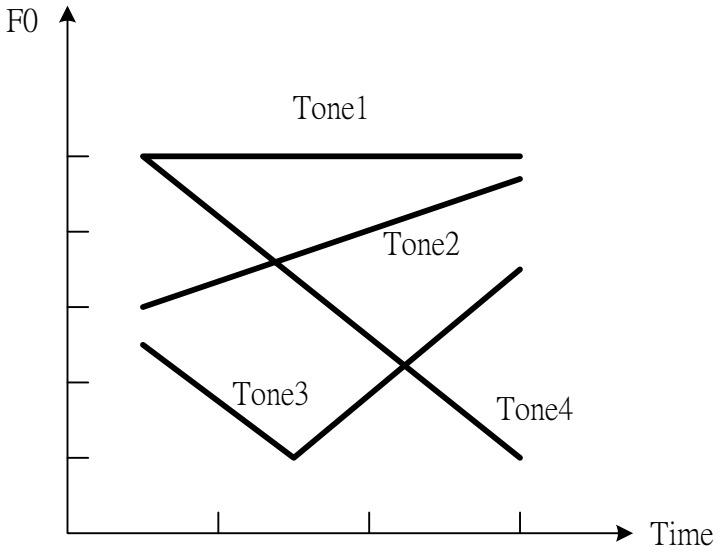


圖 3.2：漢語四個聲調的基頻軌跡圖

3.2 漢語語音階層韻律架構

相同一段文字隨著說話方式的不同，例如斷句的地方或停頓時間長短，會產生不同的語意，令人誤解，因此光是從逐個音節來探討語音中所代表的意義是不夠的，必須要再往上延伸討論。

根據語言學家的研究發現[10]。語音的韻律結構呈現階層式的架構，從最底層的音節層次(syllable layer，SYL)，往上逐步發展成韻律詞層次(prosodic word layer，PW)、韻律短語層次(prosodic phrase layer，PPh)和呼吸組層次(Breath Group，BG)，及韻律組句(Prosodic Phrase Group)，稱為「階層式多短語韻律句群(Hierarchical Prosodic Phrase Grouping，HPG)」架構[11]。

在漢語系統中，一個字不僅代表一個音節，也代表一個最基本的字義，因此音節層次是最底層的韻律單元，而聲調影響音高軌跡甚深，為此層最重要的韻律影響因素。第二層的韻律詞層次在音節層次之上，表是一個雙音節或是多音節的詞組，這些詞組通常在句法和語意上關係緊密或常常合在一起成為詞組。第三層的韻律短語層次則式由一個或是多個韻律詞所組成，其結尾通常帶有可察覺但不明顯的停頓。呼吸組層次則是由一個或數個韻律短語構成，至於第五層韻律組句是由一個或是數個呼吸組構成，但是在較短的口語段落，通常是人

們可以一口氣就能完整表達的口語段落中，呼吸組層次就已經夠用了，此時可以把呼吸組跟韻律句組合併，但是當一個呼吸組已不足以表達完整的語意段落時，就需要最上層的韻律句組。

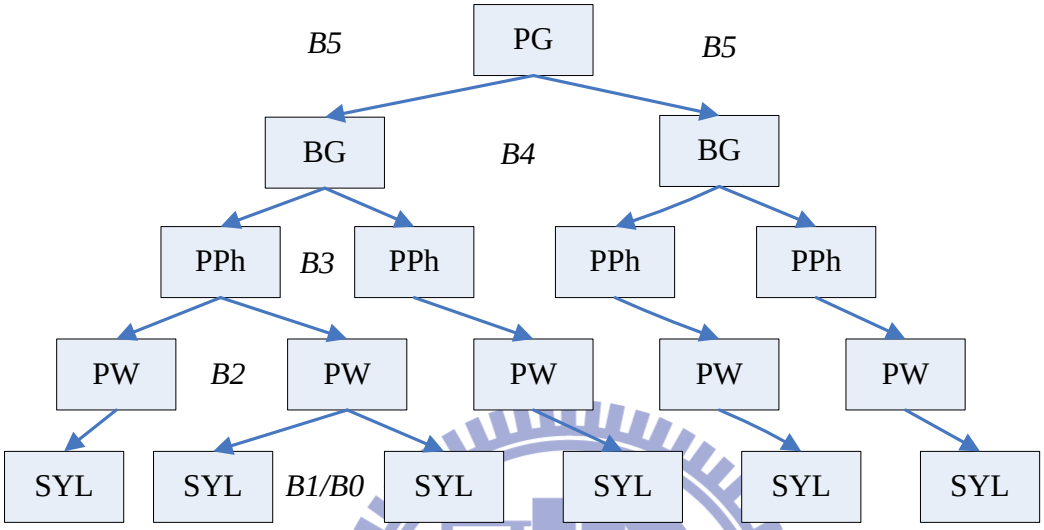


圖 3.3：中文語音韻律之階層式架構概念

如圖 3.3 所示，這 5 種韻律單元被 6 種標記所分類，B0 和 B1 都是 SYL 的邊界，差別在於 B0 表示的是 reduced syllabic boundary 而 B1 表示的是 normal syllabic boundary，通常在 B0 或 B1 的邊界聽不出停頓。B2 和 B3 分別是韻律詞和韻律短語的邊界，B4 則定義了呼吸組的邊界，一個呼吸的停頓，和 B2、B3 比起來會有個明顯的停頓，至於 B5 定義了韻律句組邊界，代表一個完整的段落結束，可以觀察到句尾的音節長度被拉長及能量減弱現象。

由於我們所採用的語料庫也是大段落的語音，因此我們以 HPG 架構為基礎，進一步對其做一些修改，利用修改後的架構來產生我們所採用的韻律模型。首先我們將 B2 再細分為 B2-1、B2-2、B2-3，其中 B2-1、B2-2、B2-3 分別代表明顯音高位置(pitch reset)之韻律詞邊界、短停頓(short pause)之韻律詞邊界以及含有音節拉長效應(duration lengthening)之後的韻律詞邊界。再來，我們將 B4、B5 合併為 B4，整個架構從 5 層變回 4 層，如圖 3.2 所示。現在我們採用這 7 種韻律邊界停頓(break type) **B**={B0, B1, B2-1, B2-2, B2-3, B3, B4}來標記這四種韻律單元：音節(SYL)、韻律詞(PW)、韻律短語(PPh)、呼吸組/韻律句組(BG/PG)，來區分韻

律結構中每一層的韻律單元，對應如表 3.3。

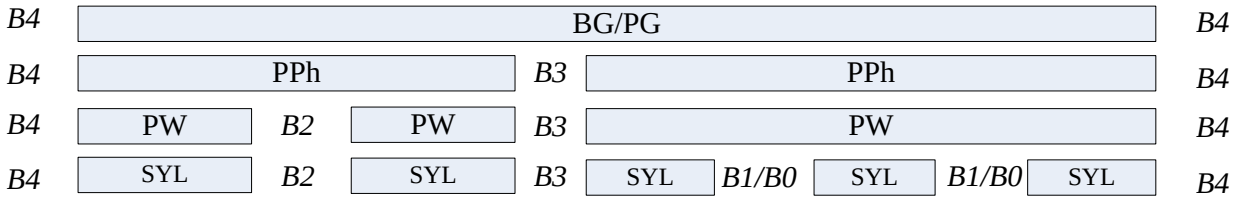


圖 3.4：修改後的階層式韻律架構

表 3.3：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)

在此將 B2 分成 3 類是因為雖然同屬於韻律詞層次，它們的聲學特性仍然有不同，原先的單一類別不足以將這相異的特性描述出；而將 B4、B5 合併主要是因為他們的聲學性質相近，故我們不多用額外的韻律邊界停頓而將其從兩類併為一類。

第四章 整合階層式韻律模型與中文大詞彙語音辨認系統

第一節先介紹我們提出新的語音辨認方法，階層式韻律之語音辨認系統，此方法結合語言與韻律資訊來改善中文大詞彙語音辨認中常見的一字詞問題及搶詞問題，除此之外，同時作豐富標註(rich transcription)產生標點符號、詞性、韻律停頓等重要資訊，接著一一介紹所使用到的模型，第二節說明本論文使用 discrimination model combination (DMC)來解決多個模型的權重問題。

4.1 階層式韻律之語音辨認系統

本研究提出一個 rich transcription 的方式，它結合階層式韻律資訊及語言資訊，使得語音可以轉寫出 word(**W**)，prosodic break (**B**)，punctuation mark (**PM**)，part-of-speech (**POS**)，和音節切割位置(Υ_s)等重要資訊，此方法之數學推導式如(4.1)式：

$$
\begin{aligned}
&\mathbf{W}^*,\mathbf{B}^*,\mathbf{PM}^*,\mathbf{POS}^*,\Upsilon_s^* \\
&=\arg\max_{\mathbf{W},\mathbf{B},\mathbf{PM},\mathbf{POS},\Upsilon_s} P(\mathbf{W},\mathbf{B},\mathbf{PM},\mathbf{POS},\Upsilon_s|\mathbf{X}_a,\mathbf{X}_p) \\
&=\arg\max_{\mathbf{W},\mathbf{B},\mathbf{PM},\mathbf{POS},\Upsilon_s} P(\mathbf{W},\mathbf{B},\mathbf{PM},\mathbf{POS},\Upsilon_s,\mathbf{X}_a,\mathbf{X}_p) \\
&=\arg\max_{\mathbf{W},\mathbf{B},\mathbf{PM},\mathbf{POS},\Upsilon_s} P(\mathbf{X}_a,\mathbf{X}_p,\Upsilon_s|\mathbf{W},\mathbf{B},\mathbf{PM},\mathbf{POS})P(\mathbf{W},\mathbf{B},\mathbf{PM},\mathbf{POS}) \\
&\approx\arg\max_{\mathbf{W},\mathbf{B},\mathbf{PM},\mathbf{POS},\Upsilon_s} P(\mathbf{X}_a,\Upsilon_s|\mathbf{W})P(\mathbf{X}_p|\Upsilon_s,\mathbf{W},\mathbf{B},\mathbf{PM},\mathbf{POS})P(\mathbf{W},\mathbf{B},\mathbf{PM},\mathbf{POS})
\end{aligned}
\tag{4.1}
$$

其中 $\mathbf{X}_a$ 代表聲學特徵參數向量，$\mathbf{X}_p$ 代表韻律的聲學特徵參數，Υ_s 代表音節的切割位置，$\mathbf{W}=\{w_1,w_2,...,w_M\}$ 代表詞序列，$\mathbf{POS}=\{POS_1,POS_2,....,POS_M\}$ 代表詞性序列，**PM** 代表標點符號序列，**B** 代表韻律停頓序列，$P(\mathbf{X}_a,\Upsilon_s|\mathbf{W})$ 為傳統聲學模型，$P(\mathbf{X}_p|\Upsilon_s,\mathbf{W},\mathbf{B},\mathbf{PM},\mathbf{POS})$ 為韻律聲學模型(break-acoustic model)，及 general prosody-syntax model $P(\mathbf{W},\mathbf{B},\mathbf{PM},\mathbf{POS})$，它結合了語言與韻律資訊，由於我們的階層式韻律是藉由語言上面的訊息來預估的，因此，$P(\mathbf{W},\mathbf{B},\mathbf{PM},\mathbf{POS})$ 可以拆解成 break syntax model $P(\mathbf{B}|\mathbf{L})$，它是利用語言參數

$\mathbf{L}=\{\mathbf{POS},\mathbf{PM},\mathbf{W}\}$ 來預估隱含著階層結構資訊的韻律停頓 $\mathbf{B}$ 的模式，連結了階層式韻律與語言資訊之間的關係，以及用來描述語言參數 $\mathbf{L}=\{\mathbf{POS},\mathbf{PM},\mathbf{W}\}$ 的 joint syntax model $P(\mathbf{PM},\mathbf{POS},\mathbf{W})$，所以可得到式(4.2)：

$$P(\mathbf{B},\mathbf{P},\mathbf{PM},\mathbf{POS},\mathbf{W})=P(\mathbf{B}|\mathbf{L})P(\mathbf{PM},\mathbf{POS},\mathbf{W}) \tag{4.2}$$

其中 $P(\mathbf{B}|\mathbf{L})$ 代表 break syntax model， $P(\mathbf{PM},\mathbf{POS},\mathbf{W})$ 代表 joint syntax model。

本研究中，為了將這些模型加入到語音辨認系統，我們採取兩階段式(two-stage)的語音辨認系統來檢驗這些模型的效果，如圖 4.1

- 第一階段**:**

  此階段為標準辨認程序 Viterbi decode，使用 acoustic model 及 language model(word-bigram)產生 N-best list。

- 第二階段**:**

  在 N-best list 的基礎上加入階層式韻律資訊(break syntax model 和 break acoustic model)及 joint syntax model 來更正搶詞錯誤、降低一字詞的混淆程度等等，因此，最後 rescore 是將 prosodic model (prosodic score)、joint syntax model (LM score)、acoustic model (AM score)等模型的 scores 作 fusion 得到新的 score，如下(4.3)式如示

$$\begin{aligned}&\text{new score}=\alpha\cdot(\text{AM score})+\beta\cdot(\text{LM score})+\gamma\cdot(\text{prosodic score})\\&\textit{where } \alpha+\beta+\gamma=1\end{aligned} \tag{4.3}$$

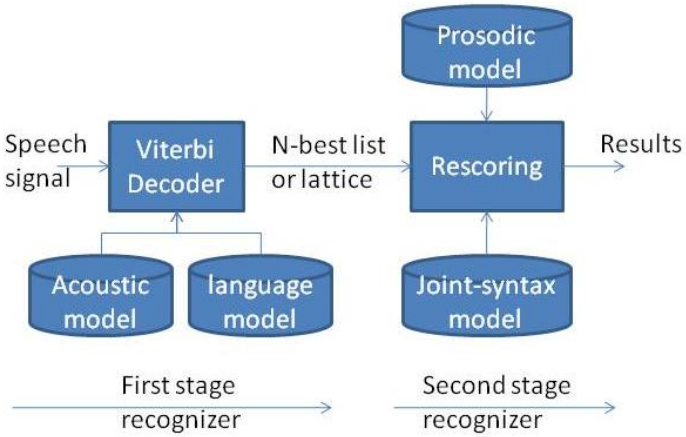


圖 4.1：兩階段式的語音辨認系統方塊圖

4.1.1 Joint Syntax Model

本論文中，factored language model (FLM)[12]用以實現此 joint syntax model $P(\mathbf{PM},\mathbf{POS},\mathbf{W})$如下式(4.4)所示：

$$\begin{aligned} &P(\mathbf{PM},\mathbf{POS},\mathbf{W}) \\ &= P(W_1)P(POS_1 \mid W_1)P(PM_1 \mid POS_1, W_1) \\ &P(W_2 \mid W_1)P(POS_2 \mid W_2, POS_1, PM_1)P(PM_2 \mid POS_2, W_2, PM_1) \\ &\prod_{i=3}^{M} \underbrace{P(W_i \mid W_{i-1}, W_{i-2})}_{\text{trigram language model}} \underbrace{P(POS_i \mid W_i, POS_{i-1}, PM_{i-1})}_{\text{factor POS model}} \underbrace{P(PM_i \mid POS_i, W_i, PM_{i-1})}_{\text{factor PM model}} \end{aligned} \tag{4.4}$$

其中 $P(W_1)P(W_2 \mid W_1)\prod_{i=3}^{M} P(W_i \mid W_{i-1}, W_{i-2})$ 代表word language model，於第一級辨認時簡化為word bi-gram，第二級時則考慮成更複雜的word tri-gram，期望再提升詞的辨認率，$P(POS_1 \mid W_1)P(POS_2 \mid W_2, POS_1, PM_1)\prod_{i=3}^{M} P(POS_i \mid W_i, POS_{i-1}, PM_{i-1})$ 代表factor POS model，$P(PM_1 \mid POS_1, W_1)P(PM_2 \mid POS_2, W_2, PM_1)\prod_{i=3}^{M} P(PM_i \mid POS_i, W_i, PM_{i-1})$ 代表factor PM model

FLM方法的最主要的概念是利用其他相關的資訊(factor)來幫忙預估目標，於是本研究期望在預估POS或PM時，能充分利用語言知識來提升預估的準確性。當然，想使用許多資訊作預估時將會遭遇到資料量不足的問題，因此，FLM就會採取 back off的架構以避免此情況的發生，其back off的數學式如 (4.5)式

- **FLM 的 Generalized back off 數學式**

$$P_{GBO}(f \mid f_1, f_2, f_3) = \begin{cases} d_N(f, f_1, f_2, f_3)P_{ML}(f \mid f_1, f_2, f_3) & \text{if } N(f, f_1, f_2, f_3) > \tau \\ \alpha(f_1, f_2, f_3)g(f, f_1, f_2, f_3) & \text{otherwise} \end{cases} \tag{4.5}$$

其中 $P_{ML}(f \mid f_1, f_2, f_3) = \dfrac{N(f, f_1, f_2, f_3)}{N(f_1, f_2, f_3)}$ 為 maximum likelihood distribution，函式 $g(f, f_1, f_2, f_3)$ 為 back off distribution，$N(f, f_1, f_2, f_3)$ 表示 f, f_1, f_2, f_3 這樣的組合出現在訓練語料的次數，當 $N(f, f_1, f_2, f_3)$ 大於 threshold τ 時，$P_{GBO}(f \mid f_1, f_2, f_3) = d_N(f_1, f_2, f_3)P_{ML}(f_1 \mid f_2$。

discount $d_N(f, f_1, f_2, f_3)$ 則為一個介於 0 到 1 的值，它會轉移一部分 $P_{ML}(f \mid f_1, f_2, f_3)$ 的機率給 back off distribution $g(f, f_1, f_2, f_3)$ 做 smoothing。

其中， Back off weight $\alpha(f_1, f_2, f_3)$ 是為了確保 $\sum_f P_{GBO}(f \mid f_1, f_2, f_3) = 1$，推導得到下式：

$$\alpha(f_1, f_2, f_3) = \frac{1 - \sum_{f:N(f,f_1,f_2,f_3)>\tau} d_{N(f,f_1,f_2,f_3)} P_{ML}(f \mid f_1, f_2, f_3)}{\sum_{f:N(f,f_1,f_2,f_3)<=\tau} g(f, f_1, f_2, f_3)} \tag{4.6}$$

在本研究裡，factor POS model的back off path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 、前一個POS及前一個PM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 PM_{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；

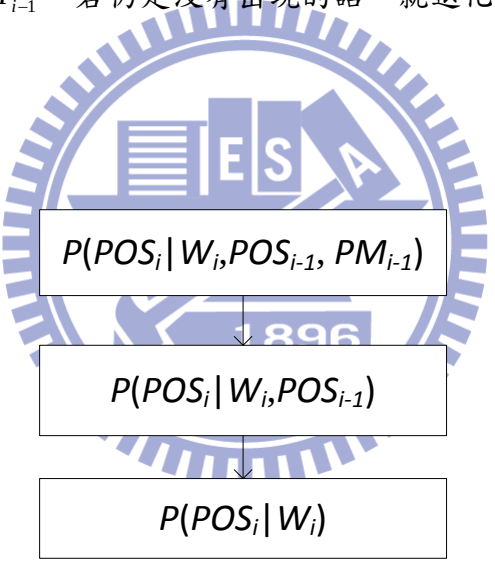


圖 4.2：factor POS model back off path

factor PM model亦是如此，其back off path的架構如圖4.3。我們使用了目前的Word、POS及前一個PM的資訊，來預估現在PM的機率為何，依照圖4.3設定，我們首先丟掉 PM_{i-1} ，接著是 POS_i ，然後 W_i ，最終退化到 $P(PM_i)$ 。

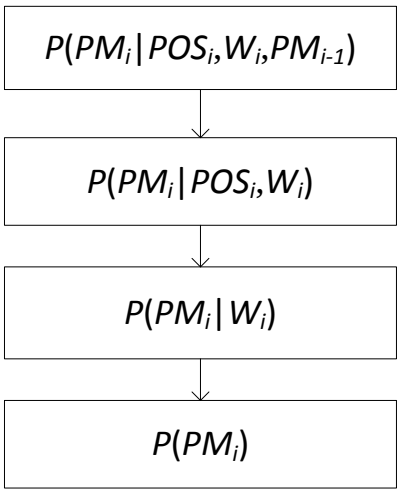


圖 4.3：factor PM model back off path

接下來 4.1.2 要介紹的 break syntax model 就是以這些語言參數為基礎來作 break type 的預估，這個物理意義即是利用上層的語言結構來猜測韻律停頓各種 type。

4.1.2 Break Syntax Model

在實驗中，Classification and Regression Tress (CART)方法被使用來描述 break type 與語言參數之間的關係，藉由一些語言上的特徵資訊(又稱問題集)幫忙對 break type 作正確的分類，完整的問題集請參考附錄 1 所示，最後 break syntax model 的數學表示式如(4.7)

數學式為：

$$P(\mathbf{B}|\mathbf{L}) = \prod_{n=1}^{N-1} P(B_n|L_n) \tag{4.7}$$

其中 N 表示句子的音節數目，$P(B_n|L_n)$ 是使用 CART 預估而得。

在分類的過程中，問題的設定是依據本研究想解決的問題，中文一字詞的混淆及搶詞的錯誤，因此，我們問了一些特殊一字詞，它們的特性是非常容易與前後的詞相連而形成一個韻律詞，舉例來說，「軍事+化」、「價值+觀」等。此時，我們已經可以預估韻律停頓，但是強度上仍覺不足，所以在 4.1.3 節中將介紹 break acoustic model，藉由 break type 上的韻律聲

學特徵來幫忙韻律停頓的正確性。

4.1.3 Break Acoustic Model

在式子(4.1)提到的 break acoustic model $P(\mathbf{X}_p|\Upsilon_s,\mathbf{W},\mathbf{B},\mathbf{PM},\mathbf{POS})$，在本論文中，$\mathbf{X}_p$ 用 5 種韻律聲學特徵參數，所以數學式將推演成式(4.8)

$$\begin{aligned} P(\mathbf{X}_p|\Upsilon_s,\mathbf{W},\mathbf{B},\mathbf{PM},\mathbf{POS}) &\approx P(\mathbf{pd},\mathbf{ed},\mathbf{pj},\mathbf{dl},\mathbf{df}|\mathbf{B},\mathbf{L}) \\ &\approx \prod_{n=1}^{N-1} P(pd_n,ed_n,pj_n,dl_n,df_n|B_n,\mathrm{L}_n) \end{aligned} \tag{4.8}$$

其中 pd_n 和 ed_n 代表第 n 個音節與第 n+1 個音節之間(第 n 個 juncture)的 pause 長度和 energy-dip level，它們表達了句子中 juncture 部份的聲學特徵，另外，pj_n、dl_n 和 df_n 用來表達出 differential 韻律特徵參數，分別是 pitch-jump level 和兩個正規化音節長度拉長因子(normalized syllable duration lengthening factor)於第 n 個 juncture。分別定義為：

$$pj_n=(\mathbf{sp}_{n+1}(1)-\boldsymbol{\beta}_{t_{n+1}}(1))-(\mathbf{sp}_n(1)-\boldsymbol{\beta}_{t_n}(1)) \tag{4.9}$$

在此**x(1)**定義為向量**x**的第一維度，下標 n 表示為第 n 個音節，$\boldsymbol{\beta}_{t_n}$ 為聲調影響因素 t_n 的 affecting patterns(APs)，正規化音節長度拉長因子**dl**和**df**定義為

$$dl_n=(sd_n-\gamma_{t_n}-\gamma_{s_n})-(sd_{n-1}-\gamma_{t_{n-1}}-\gamma_{s_{n-1}}) \tag{4.10}$$

$$df_n=(sd_n-\gamma_{t_n}-\gamma_{s_n})-(sd_{n+1}-\gamma_{t_{n+1}}-\gamma_{s_{n+1}}) \tag{4.11}$$

sd_n 為第 n 個音節長度，γ_{t_n} 和 γ_{s_n} 分別表示聲調與基本音節類型影響因素在音長的 APs。

4.2 Discriminative Model Combination

經過前面 4.1 節的介紹，在本研究中，最後拿來作 rescore 的 models 共有十個，分別為 acoustic model 一個、joint syntax model 有三個、break syntax model 一個、和 break acoustic model 有五個，因此產生一個問題，如何找出一組權重使這十個 models 作 combination 後能夠得到最小的詞錯誤率?若使用 trail-and-error 的方式決定權重的話，則既沒有效率也無法得

知是否找出最佳權重，因此本研究使用 Discriminative Model Combination(DMC)[13]的方法來決定權重，接著對 DMC 方法作一些簡單的說明，DMC 的作法是先定義一個 decision error rate 的鑑別式函數(discriminant function)，期望找到一組權重使此函數的 decision error rate 最佳化。

一開始定義一個 discriminant function 寫成：

$$\begin{aligned} & g(x_1^T, w_1^S, w'^{S'}_1) \\ &= \log P(w_1^S \mid x_1^T) - \log P(w'^{S'}_1 \mid x_1^T) \\ &= \log[P(w_1^S)P(x_1^T \mid w_1^S)] - \log[P(w'^{S'}_1)P(x_1^T \mid w'^{S'}_1)] \end{aligned} \tag{4.12}$$

其中 $w_1^S = (w_1, ...w_S)$ 為詞串，$x_1^S = (x_1, ...x_T)$ 為特徵參數向量，$P(w_1^S \mid x_1^T)$ 是給定特徵參數下看到**正確詞串**的分數，而 $P(w'^{S'}_1 \mid x_1^T)$ 則是給定同樣特徵參數下**辨識出的詞串**分數，當 $P(w'^{S'}_1 \mid x_1^T)$ 分數越接近 $P(w_1^S \mid x_1^T)$ 越好，但分數最接近者詞錯誤率(WER)確不一定最小。

在一般情況下語音辨認只會拆解成 LM 及 AM 兩部分。我們可以去調整 discriminative 參數，LM factor λ 的值，此時考慮了 λ 的 discriminant function 寫成：

$$g(x_1^T, w_1^S, w'^{S'}_1) = \log[P(w_1^S)^{\lambda} P(x_1^T \mid w_1^S)] - \log[P(w'^{S'}_1)^{\lambda} P(x_1^T \mid w'^{S'}_1)] \tag{4.13}$$

現在假設 $P(w_1^S \mid x_1^T)$ 可以拆成 M 個不同的模型，$P_j(w_1^S \mid x_1^T), j = 1, ..., M$，那麼這些模型可以 log-linearly 組合如下：

$$P^{\prod}_{\{\Lambda\}}(x_1^T \mid w_1^S) = \exp\{\log C(\Lambda) + \sum_{j=1}^{M} \lambda_j \log P_j(x_1^T \mid w_1^S)\} \tag{4.14}$$

$\Lambda = (\lambda_1, ..., \lambda_M)^{tr}$ 可以看做是模型 P_j 分數結合時的權重，$C(\Lambda)$ 則是 normalization factor。所以依據 decision error rate，我們要從 discriminant function 找出一組最好的權重 $\boldsymbol{\Lambda}$，而 discriminant function 因為 $P(w_1^S \mid x_1^T)$ 變成可以拆解成 M 個不同的模型，所以改寫為：

$$g(x_1^T, w_1^S, w'^{S'}_1) = \sum_{j=1}^{M} \lambda_j (\log P_j(w_1^S \mid x_1^T) - \log P_j(w'^{S'}_1 \mid x_1^T)) \tag{4.15}$$

接下來，我們會定義一個 smooth misclassification function $\ell(x_n, k_{n0}, \Lambda)$，再利用 Generalized Probabilistic Descent (GDP) algorithm[13]求出各個模型的權重 $\boldsymbol{\Lambda}$，這部分也是本論文實際去

求得權重的作法。

首先對使用到的符號作定義：

- 詞串 w_1^S 表示成 class k，每一個句子 x_1^T 表示為 observation x.
- 訓練資料用 $(x_n, k_{nr}), n=1,...,N, r=0,...,K$, 其中 N 為句子數目，k_{n0} 為 observation x_n 的標準答案，$k_{nr}, r=1,...,K$ 互為彼此的競爭者(即 K-best list)
- $LD(k_{nr}, k_{n0})$ 為 Levenshtein-distance 也就是 hypothesis k_{nr} 的錯誤數量(插入、刪除、替換等錯誤)

下面(4.16)式為訓練語料(or held-out data)的 smoothed empirical error rate $L(\Lambda)$

$$L(\Lambda) = \frac{1}{N}\sum_{n=1}^{N} \ell(x_n, k_{n0}, \Lambda) \tag{4.16}$$

其中

$$\ell(x_n, k_{n0}, \Lambda)^{-1} = 1 + A \cdot \left(\frac{1}{K}\sum_{r=1}^{K} e^{\left\{ -\eta LD(k_{nr}, k_{n0}) \log \frac{p^{\Pi}_{\{\Lambda\}}(k_{n0}|x_n)}{p^{\Pi}_{\{\Lambda\}}(k_{nr}|x_n)} \right\}} \right)^{-\frac{B}{\eta}} \tag{4.17}$$

其中 A>0, B>0, η>0 作適當的調整。

最後隨下列遞迴架構於式(4.18)可以計算出權重 λ_j with stepsize ε: For j=1,...,M

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

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

$$\Lambda^{(I+1)} = (\lambda_1^{(I+1)}, ..., \lambda_M^{(I+1)})^{tr} \tag{4.18}$$

根據(4.18)式，我們可以看見 λ_j 在一次又一次的遞迴中是決定於 discriminant function $\log\left(\frac{p_j(k_{n0}|x_n)}{p_j(k_{nr}|x_n)} \right)$ 的權重和。

第五章 實驗結果與討論

本章將介紹本論文所做的實驗，第一節是 joint syntax language model 相關實驗，從建模型到加入語音辨認系統的 rescoring，第二節加入階層式韻律模型的實驗，韻律模型建立的情況與 rescoring 後的結果，第三節是實驗討論。

5.1 Joint Syntax Model 實驗

5.1.1 訓練 joint syntax model

本研究的 factor POS model 與 factor PM model 是使用 SRILM toolkit[14]所建構而成的，其建立流程如下圖 5.1 所示：

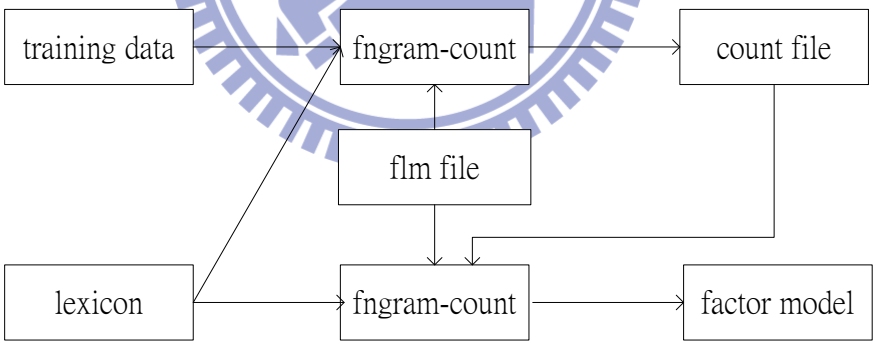


圖 5.1：FLM 訓練架構流程圖

- **Data preparation:**

訓練 factor model 時，準備的訓練語料必須包含所有我們有用到的資訊，譬如我們所訓練的 POS model 因為還給了 word 跟 PM 資訊，所以將訓練語料輸出成 W-word:T-POS:P-PM 這樣的格式，本研究使用了 46 類 POS，並依據標點符號特性將 PM 分成 4 類，分別為逗號，頓號，其他(如句號、分號、問號等…)，及沒有標點符號 4 類。

flm 訓練格式如下所示：

Original data:

這份向海外報導中華民國各方面動態的光華畫報，終於在中華民國六十五年開始的時候，與各位讀者見面了。

flm formated data:

W-這份:T-DM:P-NONE W-向:T-P:P-NONE W-海外:T-Nc:P-NONE W-報導:T-VE:P-NONE W-中華民國:T-Nc:P-NONE W-各:T-Nes:P-NONE W-方面:T-Na:P-NONE W-動態:T-Na:P-NONE W-的:T-DE:P-NONE W-光華:T-Na:P-NONE W-畫報:T-Na:P-COM W-終於:T-D:P-NONE W-在:T-P:P- NONE W-中華民國:T-Nc:P-NONE W-六十:T-Neu:P-NONE W-五年:T-Nd:P- NONE W-開始:T-VH:P-NONE W-的:T-DE:P-NONE W-時候:T-Na:P-COM W-與:T-P:P-NONE W-各位:T-Nh:P-NONE W-讀者:T-Na:P-NONE W-見面:T-VA:P-NONE W-了:T-T:P-OTH

準備好 flm 格式的語料後，整個訓練的架構如圖 5.1，主要分成兩個步驟：

- 第 1 步：從訓練語料中產生 fngram count file。

此步驟主要是統計所有我們設定的組合，從最上層完整的資訊到退化的最底層，出現在訓練語料中的次數。

- 第 2 步：從步驟 1.產生的 fngram count file 訓練出 factor model。

要注意的是這兩個步驟中都需匯入一個 flm 檔案，flm 檔案主要是設定每一層的 back off node 中的所考慮的 factor，我們在訓練 factor POS、PM model 時皆給定一條 fixed back off path(圖 4.2、圖 4.3)：

最後，factor POS, PM model 的 perplexity 效能評估於表 5.1、5.2 所示，可以觀察到每加入一種不同的 factor，perplexity 就能夠往下降，表示每一個 factor 都有助於我們預估 POS 或 PM。

表 5.1：factor POS model 的 perplexity

POS model with different factors	perplexity	
	ppl	ppl1
$P(POS_i \mid W_i)$	1.31192	1.32864
$P(POS_i \mid W_i, POS_{i-1})$	1.25764	1.27115
$P(POS_i \mid W_i, POS_{i-1}, PM_{i-1})$	1.25395	1.26725

表 5.2：factor PM model 的 perplexity

PM model with different factors	perplexity	
	ppl	ppl1
$P(PM_i \mid PM_{i-1}^{i-2})$	2.2676	2.35584
$P(PM_i \mid W_i, POS_i)$	1.4451	1.47012
$P(PM_i \mid W_i, POS_i, PM_{i-1})$	1.42995	1.45399

5.1.2 Joint syntax model rescoring

由於我們是在第二個階段進行 Rescoring，所以我們必須要知道在哪個步驟能夠掌握哪些參數，代入模型計算分數。第一階段產生了 top N 詞串後，進入第二階段前，我們的 observation 只有詞串，此時根據 word 資訊加入 joint syntax model 分數，因為 POS 跟 PM 資

訊在此階段對我們來說為 hidden，所以必須根據 word 展開對應的 POS 及 PM node 如圖 5.2：

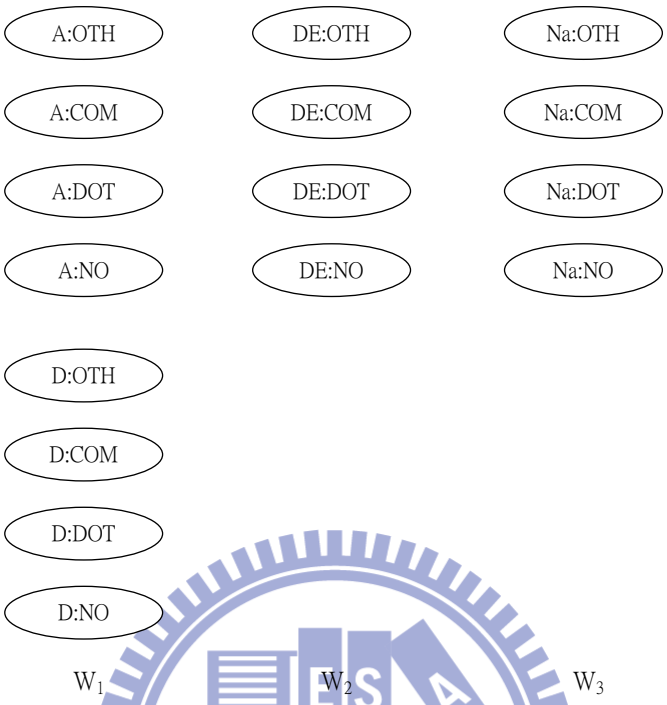


圖 5.2：詞展開對應的 POS 及 PM node 示意圖

依據詞串展開的 search space，node 跟 node 之間每一條 arc 上都可以看到目前跟下一個 node 上記錄的 POS 及 PM 資訊，再加上原來 Word 的資訊，可以依此求出所需的語言特徵參數 **L** 如圖 5.3，所以 acoustic-prosodic model 與 break-syntax model 所需對應的語言參數便能在 arc 上求得，再經決策樹分類後走到最後的 leaf node。

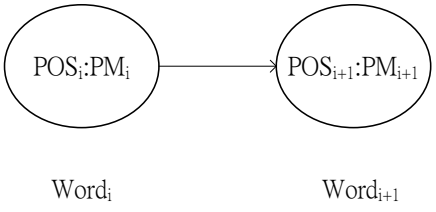


圖 5.3：從 arc 上抽取語言參數示意圖

接著我們根據第一階段所產生的 top N 詞串利用 HTK 做強迫對齊找出對應切割位置，如此便可求出詞串上對應的 acoustic-prosodic feature(pd_n 、 ed_n 、 pj_n 、 dl_n 和 df_n)，將這些參數做跟訓練韻律模型時一樣的正規化處理後，最後將這些 acoustic-prosodic feature 丟入 leaf node 的結果中計算各種 break type 的分數。break type 的決定沒有 dependent on 前一個音節的 break type，所以在搜尋最佳路徑時無須將 7 種 break type 展開，直接選擇 break type 中分數最高者加入做競爭即可。

因為先前斷成短句的原因，在搜尋最佳路徑時也會分成兩階段做處理：

- 短句段落內部最佳路徑搜尋：

在此我們使用 Viterbi algorithm，展開 POS 及 PM 搜尋空間，找出最佳路徑，且每條詞串只保留一條最佳路徑。

- 考慮跨短句段落的情況，將短句串接：

短句第一個詞上算出的 LM 分數為 $P(W_1|<s>)$，為給定前一個詞是 sentence start 的情況下計算出的 bi-gram LM 分數，這個分數除了斷開的第一段短句段落外，其餘的皆不正確，所以我們根據前一個短句中最後一個詞，將目前短句上的第一個詞的 LM 分數換掉。每一個短句段落經過第一步後都有自己的分數，跨短句的動作就等於是在短句與短句之間再做一次 Viterbi search，最終輸出即是長文的辨識結果。

表 5.3：Word 辨識率於 joint syntax model 的實驗

	Sub	**Del**	**Ins**	**Accuracy**
bigram-LM	3587	690	510	68.07%
Joint syntax model	3104	478	607	72.05%

表 5.4：Character 辨識率於 joint syntax model 的實驗

	Sub	Del	Ins	Accuracy
bigram-LM	5672	481	172	76.12%
Joint syntax model	4925	449	155	79.12%

表 5.5：Syllable 辨識率於 joint syntax model 的實驗

	Sub	Del	Ins	Accuracy
bigram-LM	3323	486	177	84.95%
Joint syntax model	3040	459	165	86.16%

5.2 階層式韻律模型的實驗

本論文在 TCC300 語料庫上的韻律停頓(break type)標記系統是依據[7]決定的，如第三章所述敘，它用來表達韻律詞組、韻律片語、韻律句組等邊界，我們依據這樣的韻律標記來訓練階層式韻律模型(break syntax model 與 break acoustic model)，將在接下來的小節作說明訓練方法。

5.2.1 訓練 break syntax model

break syntax model 經 CART 演算法利用一個已經設計好的問題集(Appendix)，依據語言參數將不同的韻律邊界停頓作分類得到一顆決策樹，如圖 5.1，決策樹中的每一個終止節點(terminal node)將有每一類韻律停頓的機率，我們也可以藉由中間非終止節點(nonterminal node)所問到的問題加以分析語言參數問題的重要程度。

下面是 CART 訓練的一些設定：

使用了 TCC300 中約 106955 個音節來訓練韻律模型

1. Minimum number of sample in a leaf node < 700
2. Relative likelihood gain < 0.001

圖 5.4 中的每一個 bar chart 代表 break types ，從左到右分別為 B0, B1, B2-1, B3, B4, B2-2, B2-3。

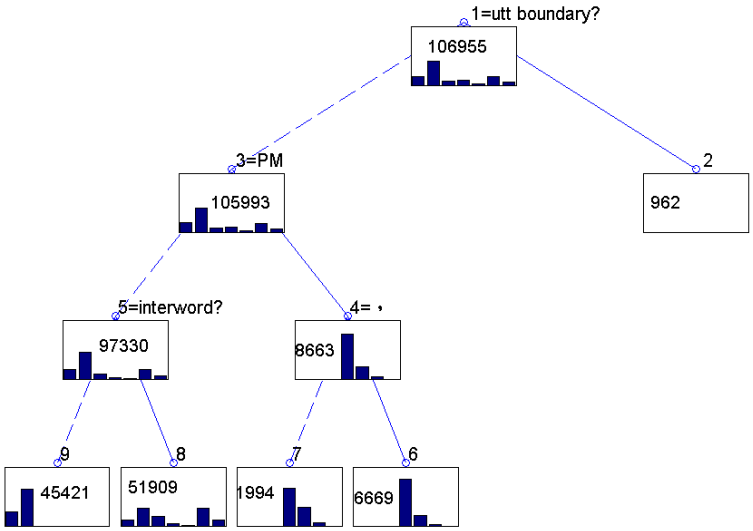


圖 5.4：break-syntax 決策樹開始於根節點

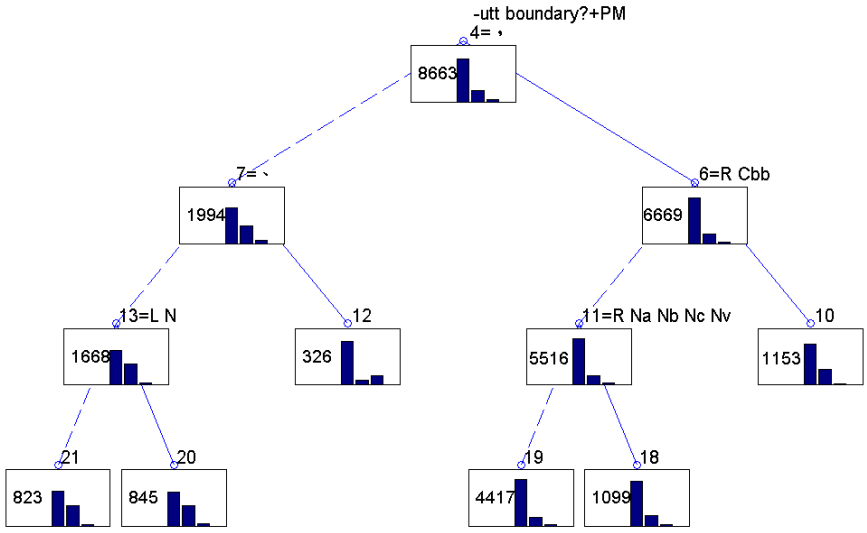


圖 5.5：break-syntax 決策樹開始於節點 4

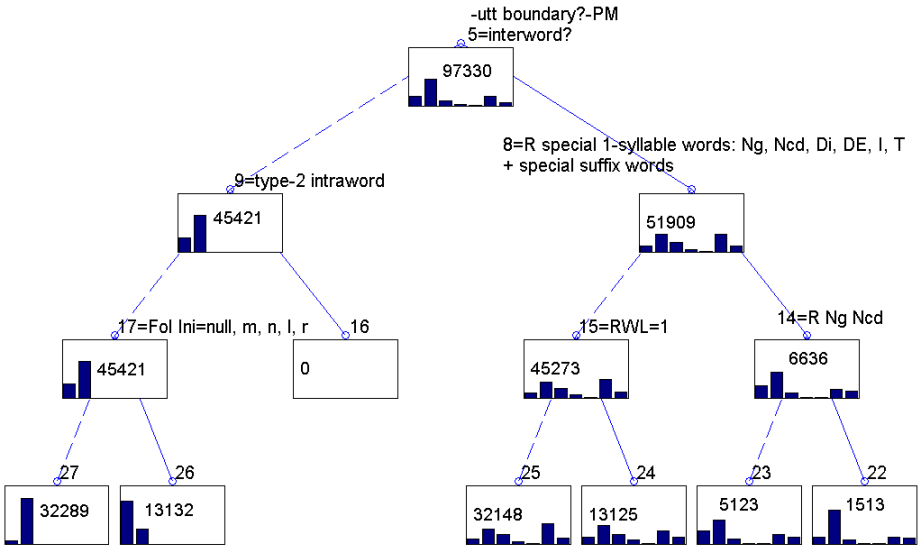


圖 5.6：break-syntax 決策樹開始於節點 5

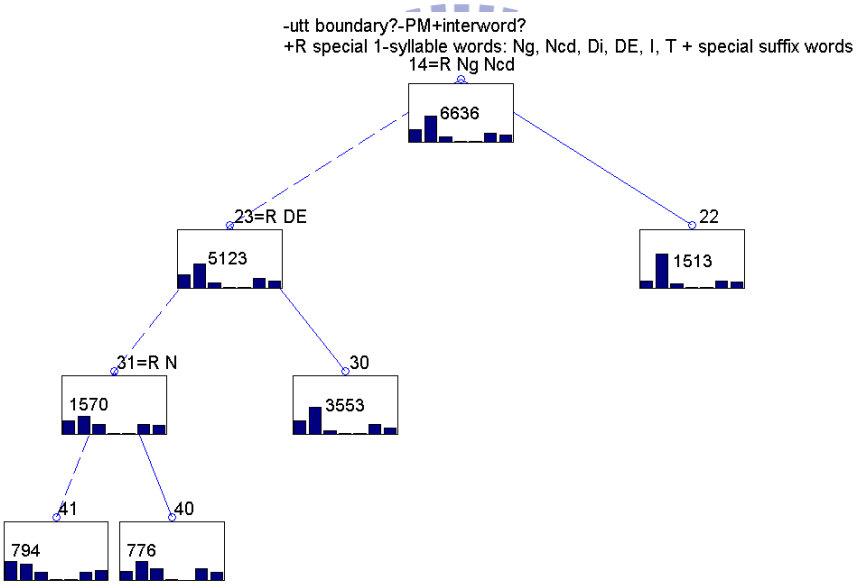


圖 5.7：break-syntax 決策樹開始於節點 14

由圖 5.6 中可以觀察到「 右邊特殊一字詞」是一個很重要的問題，所以才會在樹中前十個問題就被問題到，而這個語言參數又與我們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，因此預測此 break syntax model 將會有助於提升中文語音辨認系統的效能。

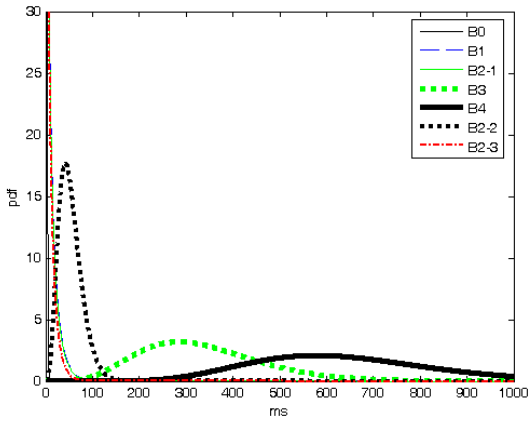
5.2.2 訓練 break acoustic model

在這部分的實驗，式子(4.8)中 $P(pd_n, ed_n, pj_n, dl_n, df_n \mid B_n, \mathbf{l}_n)$ 是經由 CART 演算法推導出來，其結點的分類標準是依據最大概似函數增(maximum likelihood gain)，CART 演算法利用一個已經設計好的問題集，依據不同的韻律邊界停頓，同時將所有的 pd_n、ed_n、pj_n、dl_n 和 df_n 做好分類。其中 pd_n 以 gamma distribution 來 fit，而 ed_n、pj_n、dl_n 和 df_n 則以 normal distribution 來 fit，因此 $P(pd_n, ed_n, pj_n, dl_n, df_n \mid B_n, \mathbf{l}_n)$ 會是一個 gamma distribution 和四個 normal distribution 的乘積。所以(4.8)式變成下式(5.1)

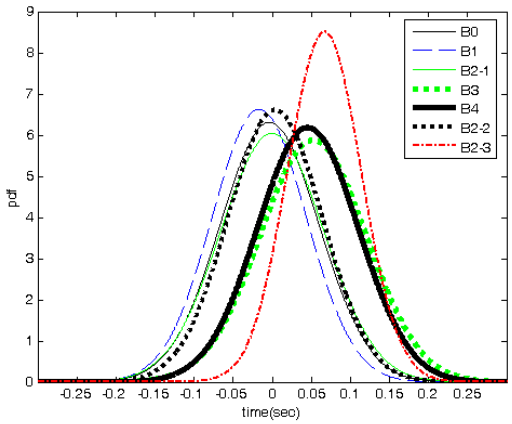
$$
\begin{aligned}
& P(\mathbf{pd},\mathbf{ed},\mathbf{pj},\mathbf{dl},\mathbf{df}|\mathbf{B},\mathbf{L}) \\
& \approx \prod_{n=1}^{N-1} P(pd_n, ed_n, pj_n, dl_n, df_n | B_n, \mathrm{L}_n) \\
& = \prod_{n=1}^{N-1} g(pd_n; \alpha_{B_n,\mathbf{l}_n}, \beta_{B_n,\mathbf{l}_n}) N(ed_n; \mu_{B_n,\mathbf{l}_n}, \sigma^2_{B_n,\mathbf{l}_n}) \\
& \cdot N(pj_n; \mu_{B_n,\mathbf{l}_n}, \sigma^2_{B_n,\mathbf{l}_n}) N(dl_n; \mu_{B_n,\mathbf{l}_n}, \sigma^2_{B_n,\mathbf{l}_n}) N(df_n; \mu_{B_n,\mathbf{l}_n}, \sigma^2_{B_n,\mathbf{l}_n})
\end{aligned}
\tag{5.1}
$$

我們將各種停頓標記之下，決策樹根節點中各個參數的 pdf 分布畫出，即不考慮語言猜數 $\mathbf{l}$ 各停頓標記之參數分布，如圖 5.8。從圖 5.8(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 pdf 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 5.8(b)、5.8(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。

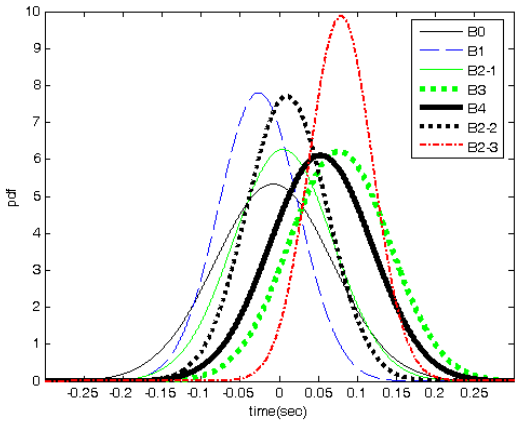
從圖 5.8(d)可看出 break type 在 energy dip 上分布的情況，對照停頓時長分布來看，確實在有比較長停頓時長的音節邊界如 B2-2、B3、B4 來看，energy dip 會比停頓時長較短的邊界還來的低。圖 5.8(e)是正規化音節間基頻差的分布，除了 B3、B4，B2-1 也表現出有明顯基頻差，B0、B1、B2-3 在正規化音節間基頻差則沒有大的鑑別度。



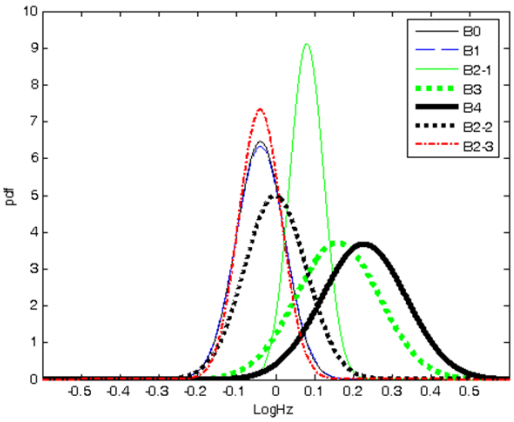
(a)



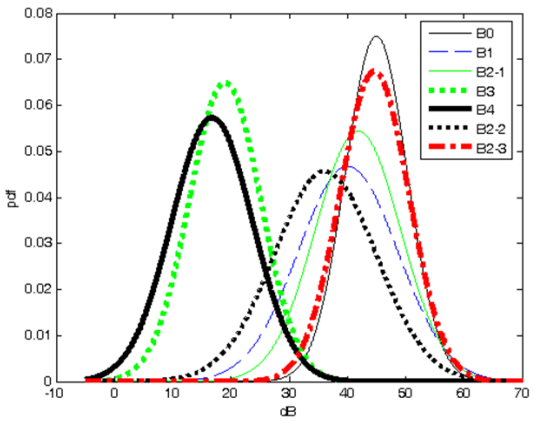
(b)



(c)



(d)



(e)

圖 5.8：(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)音節間能量低點 (e)正規化基頻跳躍值之分布圖

5.2.3 階層式韻律模型的 rescoring

表 5.6：Word 辨識率於階層式韻律模型的實驗

	Sub	Del	Ins	Accuracy
bigram-LM	3587	690	510	68.07%
Joint syntax model	3104	478	607	72.05%
+Prosody model	2880	495	564	73.72%

表 5.7：Character 辨識率於階層式韻律模型的實驗

	Sub	Del	Ins	Accuracy
bigram-LM	5672	481	172	76.12%
Joint syntax model	4925	449	155	79.12%
+Prosody model	4527	480	103	80.70%

表 5.8：Syllable 辨識率於階層式韻律模型的實驗

	Sub	Del	Ins	Accuracy
bigram-LM	3323	486	177	84.95%
Joint syntax model	3040	459	165	86.16%
+Prosody model	2853	487	110	86.97%

可以觀察到階層式韻律模型比 joint syntax model 在 word, character, 和 syllable 的正確率分別提升了 1.67%、1.58%和 0.81%。

5.3 實驗討論

由上表 5.6~5.8 可以看出，在辨識系統中加入 Joint syntax model 後，辨識率與只用了 bigram 語言模型的結果有接近 4%的大幅提升，主要是因為在 Joint syntax model 中包含了 trigram 語言模型的資訊，至於其中的 POS 及 PM 模型可能對於提升辨認率的效果可能不大。

觀察比較 joint syntax model 及加入 prosody 的實驗辨認率後發現，在 Word 的辨認率上的提升，比 Character 和 Syllable 的部分都還來的要高，這是因為我們加入的 prosody model 是用來描述韻律邊界，最主要改善的地方是在詞邊界位置判斷的正確性，所以雖然也能夠使 Character 和 Syllable 層級的辨認率有所改善，但對於 Word 辨識結果會有較為顯著的影響。

表 5.9：標記詞類及標點符號正確率

	POS	PM
Joint syntax model	73.73%	82.96%
Prosody model rescored	74.12%	81.57%

表 5.9 為標記詞類及標點符號的正確率，POS 正確率會提升主要是因為加入韻律模型做 Rescoring 時，經 DMC 調整權重後，POS 模型的權重會增加，而且由於詞的辨認率也提升了，自然 POS 的標記結果也會對應提升，PM 模型的權重在加入 Prosody Model 調整後反而變低，所以標記結果比會 joint syntax model 系統稍降。

- **更正辨認錯誤的例子**

表 5.10 將辨認結果中有改善搶詞情況及一次詞修正的部分列出，第一欄是正確文本，第二欄為 baseline 系統(加入 joint syntax model)的辨識結果，第三欄則是加入韻律模型後辨識結果，並且將解碼出的 break type 標示在右。

表 5.10：實驗辨識結果

Ref.	**baseline.rec**	**rescored.rec**
重形	重新	重形(B2-3)
砂石車	小時	砂石車(B2-1)
之	車子	據(B1)
晤談	晤談	晤談(B1)
者	的	者(B1)
男子	南市	男子(B1)
單打	但	單打(B3)
弟子	到底	弟子(B3)
	是	
演繹出	遠處	電影處(B1)
純	城市	曾(B2-1)
自我	為	自我(B1)
情境	清淨	情境(B1)
牙醫師	牙醫師	牙醫師(B2-2)
公會	公會	公會(B1)
NULL	理事	理事長(B3)
理事長	張國政	或(B2-2)
郭振興	新	真心(B3)

證實為	證實	證實為(B2-1)
口腔癌	唯恐	口(B1)
NULL	將來	將來(B2-2)

	有	
有所	數萬	有所(B2-2)
關聯	人	關聯(B3)

李女	另一支	另一(B2-1)
指控	同形	指控(B1)
情事	市場	情事(B4)

	之	
只是	實	只是(B2-2)
證明	戰	證明(B1)

及	NULL	是(B2-2)
職員	此次	志願(B3)
之	院士	時(B1)
身分	身分	身分(B2-2)

應	螢光幕	應(B2-1)
廣泛	上	廣泛(B3)

由表 5.10 的結果分析，我們可以觀察階層式韻律模型可以救到一些搶詞的錯誤、一字詞的混

淆。

一字詞混淆的更正：

「晤談 者」能夠被更正回來的理由，是因為我們在決策樹中問了「特殊一字詞」的關係，「者」這個後詞綴就屬於特殊一字詞的一種。

搶詞錯誤的更正：

「理事 張國政」更正回「理事長(B3) 」，韻律模型在「張」後面插入一個 B3，所以張國政不該是一個詞。

第六章 結論與未來展望

6.1 結論

本研究提出一套新的辨識方法，整合階層式韻律模型於語音辨識當中，實驗結果顯示，在 Word 的辨認率上提升了 1.67%，在 Character 及 Syllable 上分別提升 1.58%及 0.81%，如圖 6.1 所示，並且的確改善了我們原先預期的一字詞混淆及搶詞類型錯誤，而 POS 標記正確率與基本系統相比也提升了 0.39%。

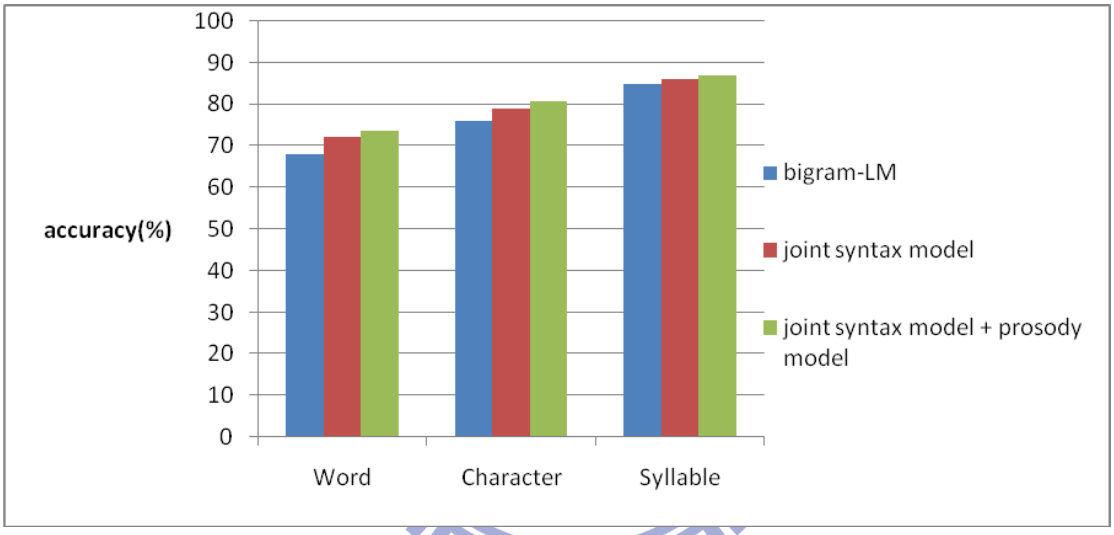


圖 6.1：各階層辨識率比較圖

6.2 未來展望

因為語音的內容是由音節與音節間的停頓所建構而成，經由本研究實驗結果已經看見音節間的停頓的韻律訊息帶來的效果，除了本研究考慮的與韻律停頓標記相關的模型外，還有與韻律狀態相關的韻律模型，能夠描述音節內的參數是如何受到韻律邊界停頓和韻律狀態及語言參數的影響，若是能將其也完整一併加入，相信對辨識率的改進還有提升的空間。並且若能將第一階段產生的 N-best list 改為輸出 lattice 形式，如此涵蓋率又能大幅提升，甚至修改二階段辨識系統架構，直接將韻律模型整合至第一階段的辨認系統上，應該能有更顯著的效果。

參考文獻

[1] L.R.Rabiner and B.H.Juang, ”Fundamental of speech Recognition,” NEW Jersey,Prentice-Hall,Inc.,1993

[2] M. Ostendorf, I. Shafran, and R. Bates, “Prosody models for conversational speech recognition,” in Proc. of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing 2003, pp. 147–154.

[3] A. Stolcke, E. Shriberg, D. Hakkani-Tür ,and G. Tür, “Modeling the prosody of hidden events for improved word recognition,” in Proc. of Eurospeech 1999, pp. 311-314.

[4] K. Chen and M. Hasegawa-Johnson.“Improving the robustness of prosody dependent language models based on prosody syntax dependence”. In Proceedings IEEE Workshop on Speech Recognition and Understanding, pp. 435–440, St. Thomas, U. S. Virgin Islands, 2003.

[5] 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 Transactions on Speech, Audio, Language Processing, Vol. 14, No. 1, pp. 232–245, 2006.

[6] 江振宇, “非監督式中文語音韻律標記及韻律模式”，國立交通大學博士論文，民國九十八年三月。

[7] Hidden Markov Model Toolkit (HTK) , http://htk.eng.cam.ac.uk

[8] F. Sha and F. Pereira. Shallow parsing with conditional random fields.

[9] 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。

[10] 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

[11] 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

[12] Jeff A. Bilmes and Katrin Kirchhoff. Factored language models and generalized parallel

backoff. In Proceedingsof HLT/NAACL, pages 4.6, 2003

[13] Beyerlein, P., “Discriminative model combination.” Proc. ICASSP 1998

[14] Stolcke, Andreas (2002): "SRILM - an extensible language modeling toolkit", In ICSLP-2002, 901-904

附錄一：決策樹問題

The question set

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:

1 Is the inter-syllable location an utterance boundary?

2 Is the inter-syllable location an interword?

3 Does a PM exist at the inter-syllable location?

4 Does a Major PM exist at the inter-syllable location?

5 Does a。exist at the inter-syllable location?

6 Does a，exist at the inter-syllable location?

7 Does a、exist at the inter-syllable location?

8 Does a．exist at the inter-syllable location?

9 Does a；exist at the inter-syllable location?

10 Does a：exist at the inter-syllable location?

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

12 Does a！exist at the inter-syllable location?

13 Does a（exist at the inter-syllable location?

14 Does a）exist at the inter-syllable location?

15 Is the the preceding special prefix words + special 1-syllable words: Ng, Ncd, Di, DE, I, T?

16 Is the POS of the preceding word A?

17 Is the POS of the preceding word C?

18 Is the POS of the preceding word D?

19 Is the POS of the preceding word N?

20 Is the POS of the preceding word I or T?

21 Is the POS of the preceding word P?

22 Is the POS of the preceding word V?

23 Is the POS of the preceding word DE?

24 Is the POS of the preceding word SHI?

25 Is the POS of the preceding word FW?

26 Is the POS of the preceding word DM?

27 Is the POS of the preceding word Da Di Dk D?

28 Is the POS of the preceding word Dfa?

29 Is the POS of the preceding word Dfb?

30 Is the POS of the preceding word Na Nb Nc Nv?

31 Is the POS of the preceding word Nd?

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

33 Is the POS of the preceding word Ng Ncd?

34 Is the POS of the preceding word Nh?

35 Is the POS of the preceding word VA VAC VG?

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

37 Is the POS of the preceding word VH VHC VI?

38 Is the POS of the preceding word V_2?

39 Is the POS of the preceding word Caa?

40 Is the POS of the preceding word Cab?

41 Is the POS of the preceding word Cba?

42 Is the POS of the preceding word Cbb?

43 Is the POS of the preceding word Da?

44 Is the POS of the preceding word Di?

45 Is the POS of the preceding word Dk?

46 Is the POS of the preceding word D?

47 Is the POS of the preceding word Na?

48 Is the POS of the preceding word Nb?

49 Is the POS of the preceding word Nc?

50 Is the POS of the preceding word Ncd?

51 Is the POS of the preceding word Neu?

52 Is the POS of the preceding word Nes?

53 Is the POS of the preceding word Nep?

54 Is the POS of the preceding word Neqa?

55 Is the POS of the preceding word Neqb?

56 Is the POS of the preceding word Nf?

57 Is the POS of the preceding word Ng?

58 Is the POS of the preceding word Nv?

59 Is the POS of the preceding word I?

60 Is the POS of the preceding word T?

61 Is the POS of the preceding word VA?

62 Is the POS of the preceding word VAC?

63 Is the POS of the preceding word VB?

64 Is the POS of the preceding word VC?

65 Is the POS of the preceding word VCL?

66 Is the POS of the preceding word VD?

67 Is the POS of the preceding word VE?

68 Is the POS of the preceding word VF?

69 Is the POS of the preceding word VG?

70 Is the POS of the preceding word VH?

71 Is the POS of the preceding word VHC?

72 Is the POS of the preceding word VI?

73 Is the POS of the preceding word VJ?

74 Is the POS of the preceding word VK?

75 Is the POS of the preceding word VL?

76 Is the length of the preceding word is 1?

77 Is the length of the preceding word is 2?

78 Is the length of the preceding word is 3?

79 Is the length of the preceding word is 4?

80 Is the length of the preceding word is 5?

81 Is the length of the preceding word is 6?

82 Is the length of the preceding word less than 2?

83 Is the length of the preceding word less than 3?

84 Is the length of the preceding word less than 4?

85 Is the length of the preceding word less than 5?

86 Is the length of the preceding word less than 6?

87 Is the following special 1-syllable words: Ng, Ncd, Di, DE, I, T + special suffix words?

88 Is the POS of the following word A?

89 Is the POS of the following word C?

90 Is the POS of the following word D?

91 Is the POS of the following word N?

92 Is the POS of the following word I or T?

93 Is the POS of the following word P?

94 Is the POS of the following word V?

95 Is the POS of the following word DE?

96 Is the POS of the following word SHI?

97 Is the POS of the following word FW?

98 Is the POS of the following word DM?

99 Is the POS of the following word Da Di Dk D?

100 Is the POS of the following word Dfa?

101 Is the POS of the following word Dfb?

102 Is the POS of the following word Na Nb Nc Nv?

103 Is the POS of the following word Nd?

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

105 Is the POS of the following word Ng Ncd?

106 Is the POS of the following word Nh?

107 Is the POS of the following word VA VAC VG?

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

109 Is the POS of the following word VH VHC VI?

110 Is the POS of the following word V_2?

111 Is the POS of the following word Caa?

112 Is the POS of the following word Cab?

113 Is the POS of the following word Cba?

114 Is the POS of the following word Cbb?

115 Is the POS of the following word Da?

116 Is the POS of the following word Di?

117 Is the POS of the following word Dk?

118 Is the POS of the following word D?

119 Is the POS of the following word Na?

120 Is the POS of the following word Nb?

121 Is the POS of the following word Nc?

122 Is the POS of the following word Ncd?

123 Is the POS of the following word Neu?

124 Is the POS of the following word Nes?

125 Is the POS of the following word Nep?

126 Is the POS of the following word Neqa?

127 Is the POS of the following word Neqb?

128 Is the POS of the following word Nf?

129 Is the POS of the following word Ng?

130 Is the POS of the following word Nv?

131 Is the POS of the following word I?

132 Is the POS of the following word T?

133 Is the POS of the following word VA?

134 Is the POS of the following word VAC?

135 Is the POS of the following word VB?

136 Is the POS of the following word VC?

137 Is the POS of the following word VCL?

138 Is the POS of the following word VD?

139 Is the POS of the following word VE?

140 Is the POS of the following word VF?

141 Is the POS of the following word VG?

142 Is the POS of the following word VH?

143 Is the POS of the following word VHC?

144 Is the POS of the following word VI?

145 Is the POS of the following word VJ?

146 Is the POS of the following word VK?

147 Is the POS of the following word VL?

148 Is the length of the following word is 1?

149 Is the length of the following word is 2?

150 Is the length of the following word is 3?

151 Is the length of the following word is 4?

152 Is the length of the following word is 5?

153 Is the length of the following word is 6?

154 Is the length of the following word less than 2?

155 Is the length of the following word less than 3?

156 Is the length of the following word less than 4?

157 Is the length of the following word less than 5?

158 Is the length of the following word less than 6?

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

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

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

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

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

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