國立交通大學

電信工程研究所

碩士論文

使用韻律訊息於建立聲學模型之中文語音辨認

Incorporating Prosody Information in Acoustic Modeling for Mandarin Speech Recognition

研 究 生：邱子軒

指導教授：陳信宏　教授

中 華 民 國 一百零一 年 三 月 二十二 日

使用韻律訊息於建立聲學模型之中文語音辨認

Incorporating Prosody Information in Acoustic Modeling for Mandarin Speech Recognition

研 究 生： 邱子軒　　Student：Tzu-Hsuan Chiu

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 of Science

in Communication Engineering

March 2012

Hsinchu, Republic of China

中 華 民 國　一百零一 年 三 月 二十二 日

使用韻律訊息於建立聲學模型之中文語音辨認

研 究 生：邱子軒　　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

摘要

本研究探討如何使用韻律訊息於聲學模型(acoustic model, AM)之建立，用於中文語音辨認。本研究在訓練聲學模型時，將傳統前後文相關(context dependent)的 tri-phone HMM 拓展至在音節邊界時，同時考慮韻律停頓(prosodic break)的影響。其中韻律停頓分為四種強度，用以表示音節間不同的緊密接合程度，並採用分類回歸決策樹(Classification and Regression Trees, CART)建立一個與前後文及韻律停頓相關的聲學模型。在辨認時分為兩個階段，在第一階段只利用聲學模型進行音節的辨認產生音節圖(syllable lattice)，且含有韻律停頓的資訊。在第二階段，針對音節圖配合詞典並輔以韻律停頓的資訊進行構詞，將其轉為詞圖(word lattice)，最後再結合語言模型(language model, LM)重新計分(rescoring)，實現詞的辨認。使用 TCC300 語料庫之實驗結果顯示本方法較傳統之 tri-phone HMM 有較好的辨認率。

Incorporating Prosody Information in Acoustic Modeling for Mandarin Speech Recognition

Student：Tzu-Hsuan Chiu　　Advisor：Dr. Sin-Horng Chen

Institute of Communication Engineering
National Chiao Tung University

Abstract

The thesis presents a study on introducing prosody information to acoustic modeling for Mandarin speech recognition. Its idea is to extend the conventional context-dependent (CD) tri-phone HMM modeling approach to further consider the dependency of phone model on the break type of nearby inter-syllable boundary. Four break types are considered, including major break, minor break, normal non-break, and tightly-coupled non-break. In the training phase, prosody- and phonetic-dependent phone models are constructed by using Classification and Regression Trees (CART) Algorithm. In the test phase, a two-stage recognition approach is adopted. In the first stage, we use the acoustic models to generate a syllable lattice which contains prosodic break information. In the second stage, we first construct a word lattice from the syllable lattice by constructing all possible words using a lexicon with the help of prosodic information, and then find the best output word sequence by rescoring using a trigram language model. Experimental results on the TCC300 database showed that the proposed method slightly outperformed the conventional method using tri-phone acoustic models.

誌謝

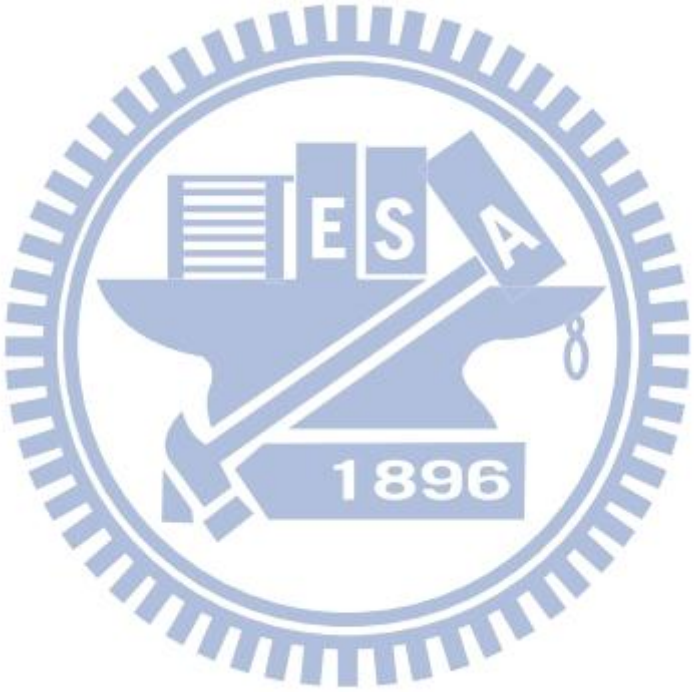
首先誠摯的感謝陳信宏老師在忙於學校的行政事務之餘，仍能關心我的研究，不時與我討論，引領我的研究方向。感謝王逸如老師教導我許多研究的基本觀念，如何抓住重點學習以呈現自己的報告，並改變自己的學習態度。

本論文得以完成也要感謝廖元甫老師對於我的研究細節給了許多的建議與協助，使我的研究更順利，論文更加完整。

此外，也感謝性獸學長，從我還是專題生時就給予了我許多不只是研究的知識，也告訴我不管以後從事的是哪個領域的工作，最重要的是保持一顆有學習熱忱的心，如此才能不斷的進步。感謝智合學長在我剛踏入研究時，耐心與我討論，給予我許多意見與協助，使我能迅速的熟悉並投入研究。感謝政賢學長對於我研究的提點，在我迷惘時能讓我更清楚下一步該怎麼做。感謝希群學長、輝哥學長及阿德學長對於我研究的鼓勵。感謝文良、豆腐、銘傑、大胖、小蝦、智障及啟全學長，不管是在研究上的指點或是你們的幽默風趣。當然也要感謝從大一就認識的企鵝，謝謝你平時給予我許多的協助，也感謝 707 的另外三顆輪子林俊翰、陳睿詮及謝喬華，研究少不了你們，玩樂更少不了你們。感謝純情可愛的翹秘書靖觀，善良天真的的蔡昌祐，教我 battle 的雅婷，有商業頭腦的昂星及籃球很厲害的維陽，謝謝你們的陪伴。接著要感謝，幫我買早餐的仲銘、胸肌很大塊的奕勳、體貼的婉君、聰明的子睿及神秘的良基等學弟妹，也祝你們明年能順利畢業。

最後要感謝我的家人及朋友，是你們對我的支持與信任，才讓我能堅持下去完成我的碩士論文。

目錄



表目錄

圖目錄

第一章 緒論

1.1 研究動機

韻律在口語中扮演著很重要的角色，能幫助人們辨認每一個詞以及整段句子的結構。所謂韻律就是連續語音中一種跨區段(supra-segmental)的特徵，如重音(stress)、停頓(pause)及語調模式(intonational pattern)等。多數與韻律相關的研究多探討能量(energy)、音高軌跡變化(pitch)及語音或停頓時間(pause duration)的長短；然而，有相關的研究指出，韻律也會影響發音聲學參數(phonetic-acoustic feature)。因此在建立聲學模型時，若能將韻律當作一個影響因子，所得到的聲學模型其模型分布會更集中，降低與其他模型的混淆度。同時，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)甚至到句子(sentence)，因此我們也可以藉由建立一個與韻律相依的聲學模型(prosody-dependent acoustic model, PD-AM)來偵測何處為詞的邊界，刪除符合構詞規則但是韻律不合理的詞，亦可偵測何處為句子的邊界以縮小辨認的搜尋網路。甚至可以利用韻律將詞綴併入形成韻律詞(prosodic word)，減少一字詞的錯誤，增加詞的辨識率。

1.2 文獻回顧

早在 Lee [1]在提出建立與前後文相關的音素模型時，就指出當訓練語料量夠多時，就應該考慮音節的位置(syllable position)或重音(accent)等影響因子。Ostendorf 等人[2]-[3]採用 Switchboard 對話語料庫，藉由分析決策樹(decision tree)來探討韻律對於聲學模型的影響，發現不同的停頓指數(break indices)的確對發音聲學參數有所影響。然而，由於缺乏豐富的韻律標記語料，並沒有進行詞的辨認實驗，無法評估其對辨認是否有幫助。Ostendorf 等人[4]提出使用韻律在語音辨

認，藉由引進韻律參數及 word-based 語言參數做動態發音模型(dynamic pronunciation model)，發現韻律參數對於動態發音預測有少許的改善。Chen 等人[5]使用了兩種韻律參數，分別為語調片語邊界(intonational phrase boundary)和音高重音(pitch accent)，用以建立與韻律相依(prosody-dependent)的音素模型，研究結果對於在 Boston University Radio News Corpus (BU-RNC)的詞錯誤率(WER)改善了 10.8%。 Ni 等人[6]以自動標記韻律的方式針對 863 corpus 標記出有無停頓(break)及重音(stress)，進而建立一個韻律相依的音調音節聲學模型(tonal syllable acoustic model)，實驗結果顯示對於音調音節的正確率有顯著的改善。黃等人[7]利用音高(pitch)、能量(intensity)及語音長度(duration)等韻律參數，以 polynomial regression model 產生可變參數的隱藏式馬可夫模型(variable-parameter HMM)，建立 PD-AM 針對傳統辨認器所產生的 N 條最佳(N-best)的辨認結果做重新評分，實驗結果顯示 WER 改善了 0.47%。

1.3 研究方向

本研究採用[8]所提出之階層式韻律架構，並經過修改，將韻律停頓(prosodic break)分為四類，把傳統考慮前後文相關(context dependent) 的 tri-phone HMM 拓展至在音節邊界時，同時考慮韻律停頓的影響。由分類回歸決策樹(CART)的分析，觀察韻律停頓對於發音聲學參數的重要性，同時藉此提供下層音節資訊無法提供之韻律階層架構的影響，使分群的結果更好。由於在訓練時考量了韻律停頓的影響，因此期許其效能可以超越傳統的聲學模型；此外，本研究也注重其解碼出之韻律停頓標記是否能幫助偵測呼吸群組句的邊界。藉由聲學模型辨認產生之音節圖(syllable lattice)，在構詞時搭配解碼出之韻律停頓的資訊，期望提高詞的辨認率。

1.4 章節概要說明

本論文一共分為五章，其各章節內容分配如下：

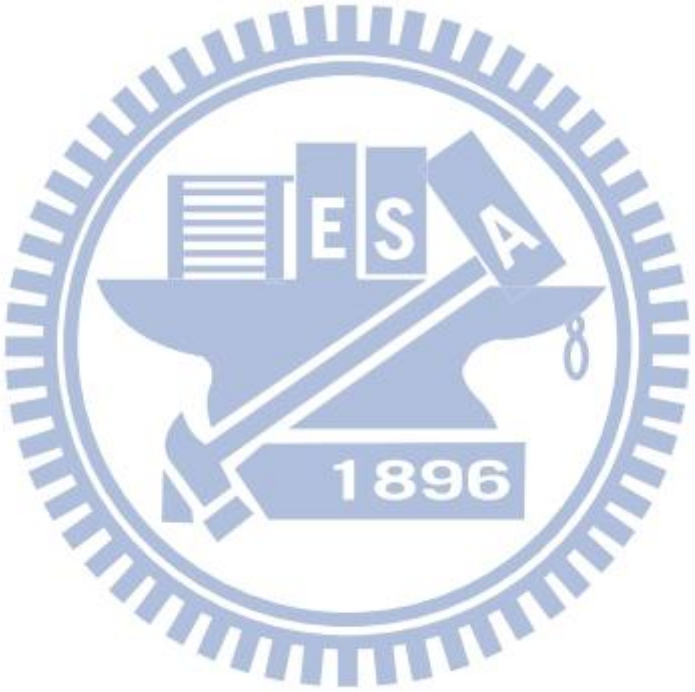
第一章：緒論。

第二章：與韻律停頓相依之聲學模型。

第三章：音節辨認之實驗結果及分析。

第四章：以有限狀態機實現大詞彙語音辨認。

第五章：結論與未來展望。



第二章 與韻律停頓相依之聲學模型

本章介紹本研究中所提出的與韻律停頓相依之聲學模型(PD-AM)。模型是使用 TCC-300 語料庫建立，並以[9]所提出之“非監督式中文語音韻律標記及韻律模式”(Unsupervised Joint Prosody Labeling and Modeling Algorithm, PLM)獲得的韻律停頓標記同時搭配分類回歸決策樹(CART)來訓練聲學模型，以隱藏式馬可夫模型(Hidden Markov Model, HMM)呈現。2.1 節簡介實驗所使用之語料庫；2.2 節介紹中文韻律模型；2.3 節介紹本論文聲學模型之訓練設定及流程； 2.4 節則對聲學模型的訓練結果做分析；2.5 節會進一步探討是否構成音節的所有音素皆會受到音節前、後之韻律停頓的影響。

2.1 語料庫簡介

本研究是使用 TCC-300 麥克風語音資料庫[10]，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

依據表 2.1，本研究考慮到韻律學習(prosody study)，所使用之語料庫只包含其中的 183 位語者，內容以長句為主。本研究會對該語料庫分為訓練語料及測試語料，其中訓練語料有 164 位語者共 962 的長句音檔，音檔長度共約 8.3 小時，詞總數量達 61534，共 106955 個音節；測試語料的部分包含 19 位語者共 226 個長句音檔，總長度約 2 小時，詞總數量為 14993，共 26357 個音節。

2.2 中文語音韻律模型

當人利用語音溝通時，除了話語本身的詞意會影響對方接收到的語意之外，說話時音調的抑揚頓挫及音量的高低起伏等，皆會影響，而這些語音上的變化我們稱為韻律變化。本節先介紹中文語音韻律階層式架構及本研究對其所做的修改；接著說明韻律標記及模型的訓練方法。

2.2.1 中文語音韻律階層式架構

根據韻律相關的研究[11]發現，語音的韻律結構呈現階層式架構，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，如圖 2.1 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)，及韻律組句(prosodic phrase group)。因中文一個音節一字的特性，最底層的韻律單元維音節；韻律詞則是由一個或多個詞所組成的詞組；韻律短語由一個或多個韻律詞組成；呼吸組層次代表一個有音高及音長明顯變化的段落；韻律組距則由連續的呼吸組構成。這整體架構統稱「階層式多短語韻律句群(Hierarchical Prosodic Phrase Grouping，HPG)」架構[12]。

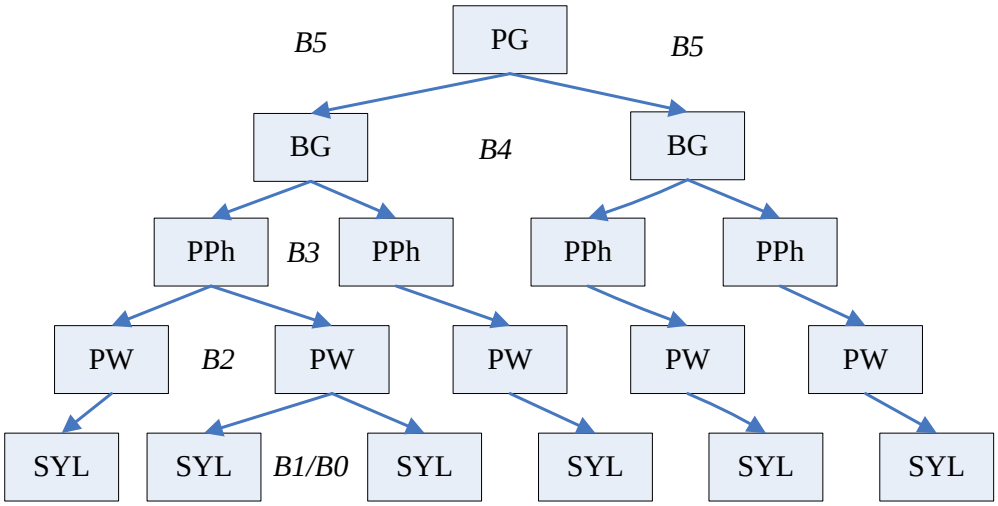


圖 2.1：中文語音韻律之階層式架構概念

這裡我們將使用韻律停頓標記來區分階層式韻律架構中的各層韻律組成份子，如上圖 2.1 所示。其中 B0 和 B1 都是 SYL 的邊界，差別在於 B0 表示的是 tightly-coupling syllable juncture，而 B1 表示的是 normal syllable boundary，通常在 B0 或 B1 的邊界不具有明顯停頓。而 B2 是韻律詞邊界，B3 及 B4 則分別代表韻律短語及呼吸組的邊界，和 B2 比較起來會有個明顯的停頓，至於 B5 代表一個完整的段落結束。

本研究以 HPG 架構為基礎並對其作進一步修改。首先我們將 B1 細分為 B1-1、B1-2，其中 B1-1 代表 normal syllable boundary，不具有明顯停頓，B1-2 代表詞內的音節邊界有較明顯的停頓。B2 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)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。現在我們共有 8 種韻律邊界停頓(break type) **B**={B0, B1-1,B1-2, B2-1, B2-2, B2-3, B3, B4}。

然而，在以這樣的設定下所做的聲學模型實驗，包含決策樹的分析及辨認出的音節圖(syllable lattice)，發現並非這八種 break type 對於發音聲學參數皆有很大的影響，因此在建立聲學模型時，我們將其合併成四大類，**B**={BT0、BT1、

BT2、BT3}。其中 BT0={B0}，連音現象較強；BT1={B1-1,B2-1,B2-3}，不具明顯之停頓；BT2={B1-2,B2-2}，含有可察覺之短停頓；BT3={B3,B4}則有明顯的長停頓。這四類韻律停頓其停頓時長(pause duration)之分布直條圖如圖 2.2 所示。

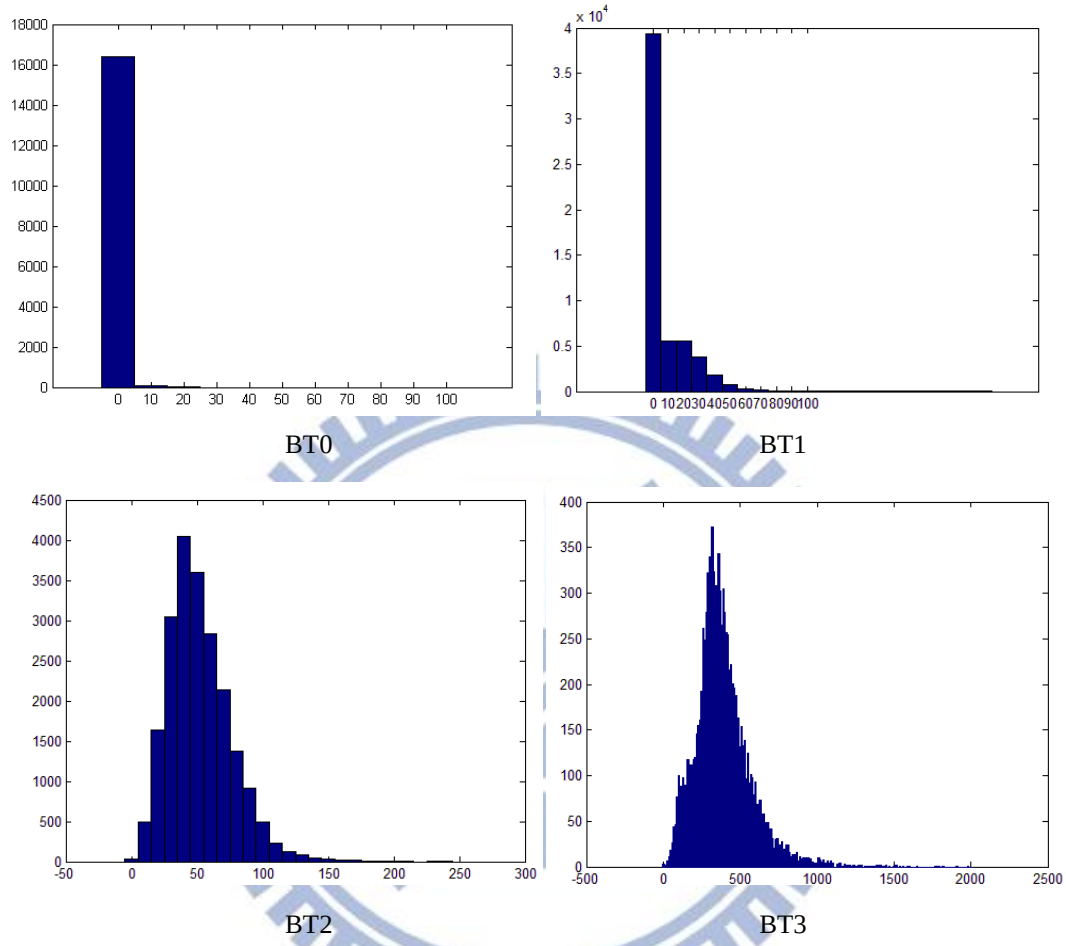


圖 2.2：各種韻律停頓其 pause duration 分布圖。橫軸單位為毫秒(ms)，縱軸單位為次數(count)

2.2.2 韻律標記及模型訓練方法

非監督式中文語音韻律標記及韻律模式(PLM) [9]依據 maximum likelihood(ML)法則來預估韻律模型的參數，並對訓練的語句做韻律標記，經由一連串最佳化的程序直到收斂。整個演算法分為兩個部份，分別為初始化和疊代。初始化會對所有訓練語句做初始的韻律標記，以及預估韻律模型參數的初始值；疊代則事先對所有語句定義一個 likelihood function，接著利用一個多重步驟的疊

代程序，重復更新韻律標記及韻律模型參數。依此訓練方式得到的四大類韻律停頓標記將會在訓練聲學模型時使用。

2.3 以韻律停頓標記輔助訓練聲學模型

2.3.1 發音聲學參數抽取

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，本研究將使用的參數為 MFCC (Mel-Frequency Ceptral Coefficients，梅爾倒頻譜參數)，以 32 毫秒之漢明窗(Hamming Window)且每位移 10 毫秒做為一筆資料，其成分包含 12 維 MFCC 加上 1 維能量共 13 維，並取其一階變量(delta term)和二階變量(delta-delta term)，但單純的能量在參數中較為缺乏鑑別性，因此去除能量係數，共 38 為做為本研究之發音聲學參數。系統相關設定如表 2.2 所示。

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

2.3.2 聲學模型建立流程

由於使用 Flat Start 在語句長時容易發生切割位置錯誤之情況，訓練得到的聲學模型較差，因此本研究利用右相關聲/韻母模型(Right-context-dependent Initial/Final Model, RCD)對訓練語料做音節的切割，再根據音節的切割位置訓練單音素(mono-phone)聲學模型。以單音素聲學模型為初始模型，接著訓練跨音節

三連音素模型(Cross-word Tri-phone Model)，每個音素採用 3 個由左至右(left-to-right)的狀態(state)表示。

有別於傳統僅考慮前後音素影響的三連音素模型，本研究會對音節邊界之音素額外考量 PLM 標記之韻律停頓對其造成的影響。訓練時文本(text)由音節層級(syllable level)展開至音素層級(phone level)的方式將如圖 2.3 所示，僅位於音節邊界的音素會受到相鄰之韻律停頓的影響。

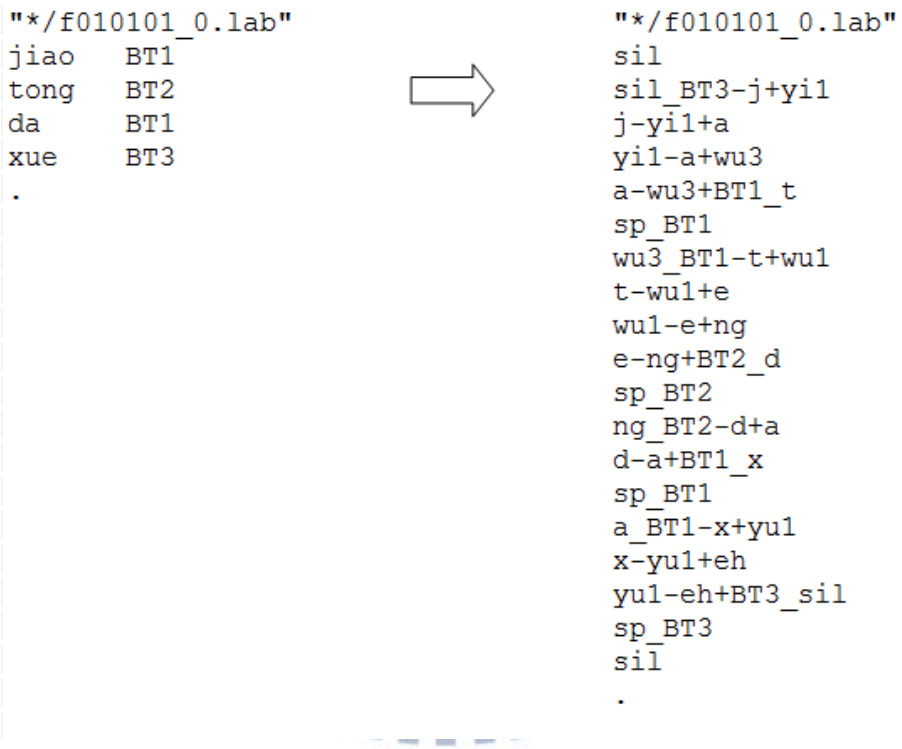


圖 2.3：訓練過程文本由音節層級展開至音素層級示意圖

short pause(sp)的 HMM 將依照 break type 個別訓練，分別為 sp_BT0、sp_BT1、sp_BT2 及 sp_BT3，其結構如圖 2.4 所示。sp_BT0 採用 1 個狀態，可以跳過(skip)且不能停留(non-recurring)，以達到其 tightly-coupled 的特性；sp_BT1 採用 1 個狀態，可以跳過且可以停留，以達到其不停頓或有極短暫停頓的特性；sp_BT2 採用 3 個狀態，其第 2 個狀態可以跳過且可以停留，以達到其短暫停頓的特性；sp_BT3 採用 3 個狀態，以達到其長停頓的特性。

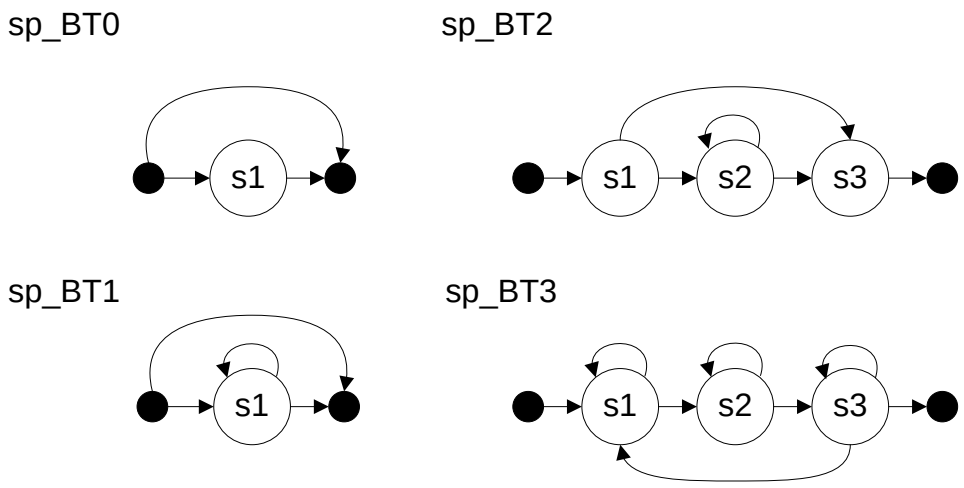


圖 2.4：short pause models 結構示意圖。其中黑色的狀態為 null state

由於訓練語料中不可能包含所有三連音素與韻律停頓之組合，加上某些組合出現的次數過少，其資料並不可靠，因此採用分類回歸決策樹(Classification and Regression Trees, CART)做為參數分享之方法，並使用語言學及韻律停頓相關之問題做為問題集，詳細的問題集如附錄所示。本研究採用的標音系統中共有 38 個音素，因此總共有 114 顆樹。與韻律相依之聲學模型(PD-AM)的訓練流程如圖 2.5 所示。

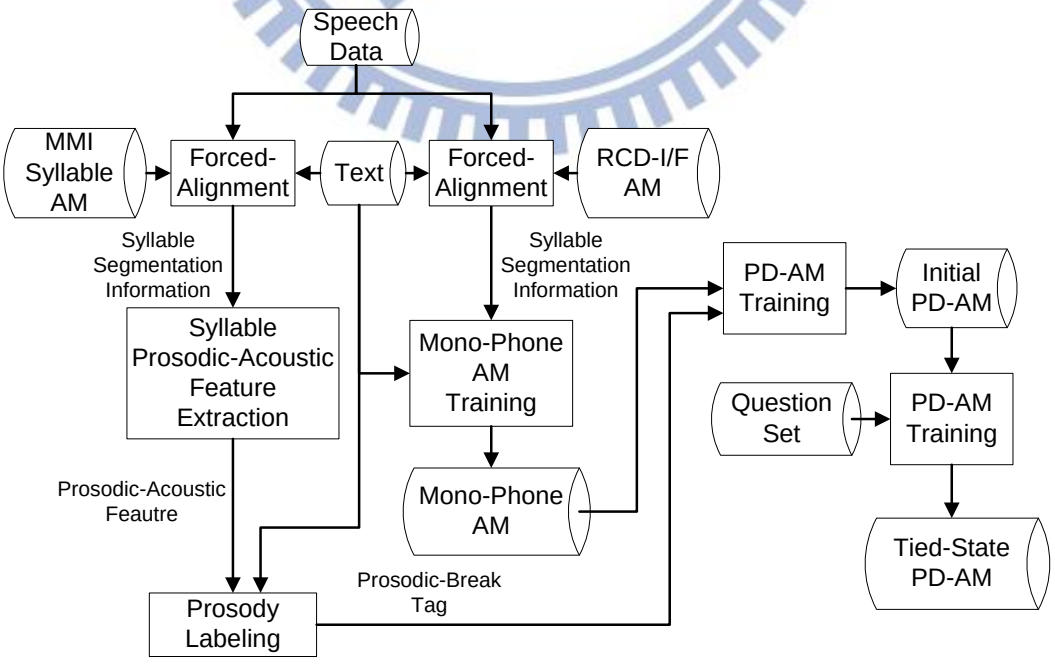


圖 2.5：與韻律相依之聲學模型訓練流程

2.4 模型訓練結果分析

本研究著重在韻律停頓訊息對於發音聲學參數之影響，因此本節的分析將著重在頻譜模型與其決策樹生長的情形，同時也會觀察不同韻律停頓下相同模型之 MFCC 分布的差異。最後會比較 PD-AM 和傳統不與韻律相依之聲學模型 (prosody-independent acoustic model, PI-AM)其訓練過程中每個 frame 平均之 log-likelihood 的演進。

2.4.1 決策樹分析

與前後文相關之模型的好壞與決策樹分裂的情形密切相關，若問題不恰當則會產生不好的分群，子節點得到的資料分布也容易使接續的問題變得不合理，如此得到的聲學模型其效果必定不好。在決策樹中，越上層被問到的問題代表其對整體資料的分群影響越大，能將資料分成差異明顯的兩群，降低模型彼此之間的混淆度。

本研究認為韻律停頓對於頻譜會造成顯著影響，因此期望與韻律停頓相關之問題可在上層就被問出。由表 2.3 的統計結果發現，有高達 63%的決策樹在前三層就已經出現韻律停頓相關問題，決策樹的平均階層為 7.03。這也證明韻律停頓對於發音聲學參數確實有影響，如此一來我們將可藉由韻律停頓標記使決策樹分群結果中之每個節點的模型分布更為集中，所得到的聲學模型效能更好。

表 2.3：音素單元決策樹中，韻律停頓相關問題第一次出現之位置統計

層數	1	2	3	> 3
樹的數量	14	32	26	42
累計樹的數量	14	46	72	114

若進一步觀察，可發現前三層沒有出現韻律停頓相關問題的模型為:

1. 母音的第 1 個狀態。
2. 子音的第 2 個狀態。
3. 子音的第 3 個狀態。

而這也非常符合我們的認知，音節中間的狀態的確較不容易受到韻律停頓的影響，這也是為什麼我們僅對音節邊界的因素考慮韻律停頓的影響。

此外，本研究也認為當韻律停頓屬於長停頓時(BT3)時，就不應該跨過音節去考慮前、後音素的影響，因此期許由決策樹的結構中，可觀察出韻律停頓能有效區隔連音效應的強、弱。因此，我們對決策樹做統計，在 230 個與 BT3 相關的問題中，回答「是」的這群在接下來的問題只出現過 14 個跨音節與前、後音素相關的問題，證明了長停頓類別確實能區分音節間連音效應的強、弱。我們更可藉由此現象，在辨認時，只在兩個 BT3 的區間做局部性的辨認，縮小辨認產生的音節圖(syllable lattice)，加速辨認的時間。

進一步分析，在所有的決策樹中，每顆樹的前 1/3 問題與韻律停頓相關的以 BT3 最多，如表 2.4 所示。由於明顯的停頓可區分連音效應，而連音效應對於發音聲學參數又有的極大的影響，因此與 BT3 相關的韻律問題出現較多次也是合理的。

表 2.4：音素單元決策樹其前 1/3 問題中，與韻律停頓相關之問題的數量統計

韻律停頓	BT0	BT1	BT2	BT3
數量	6	32	6	99

2.4.2 不同韻律停頓對 MFCC 造成之影響

由前一節的分析得知，不僅韻律停頓確實會影響頻譜，不同的韻律停頓對頻譜也會有不同的影響。根據表 2.4 的分析，足見 BT3 對於發音聲學參數造成的影

響更為明顯，因此若將相同的 tri-phone 在 BT3 和其他韻律停頓下的 MFCC 分布畫出，應該也可看出受到 BT3 影響的 MFCC 分布會與受到其他韻律停頓影響的分布不同。

為了驗證此想法，首先利用訓練好的 PD-AM 對訓練語料做強制對齊(forced-alignment)，並抽出我們欲觀察模型的 MFCC。在此將以 a-ng+d 第 3 個狀態其 MFCC 為例子，因為其模型在決策樹的結構中與 BT3 相關的問題出現在第二層，說明 BT3 對其發音聲學參數的影響很大。將 a-ng+d 第 3 個狀態其 MFCC 分為兩組，一組為受到 BT3 的影響，一組為受到其他韻律停頓的影響，接著將其 MFCC 的第一維及第二維畫出，如圖 2.6 所示。

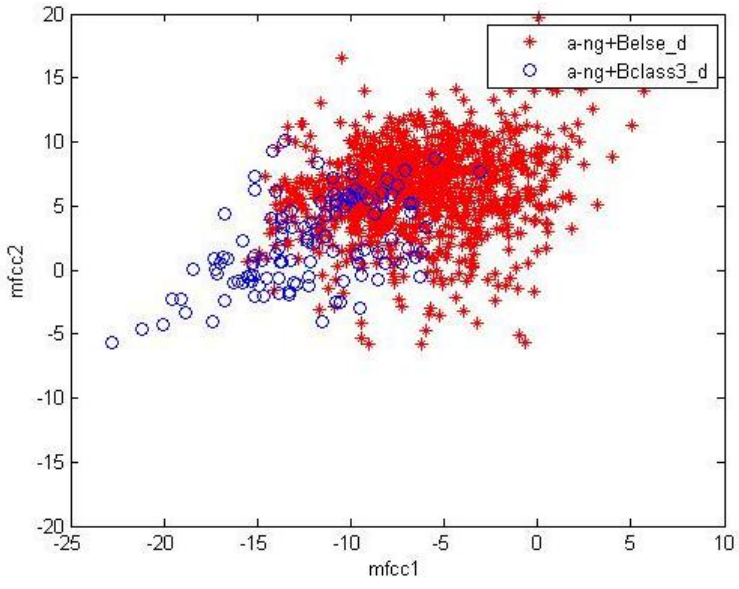


圖 2.6：a-ng+d 第 3 個狀態在 BT3 和其他韻律停頓條件下其 MFCC 之分布。圓形空心的點為受 BT3 影響之 MFCC；星狀實心的點為受其他韻律停頓影響之 MFCC

由圖 2.6 可看出，相較於其他韻律停頓，BT3 對於 MFCC 的影響的確不太相同，使其分布成兩群，也證明了不同的韻律停頓確實會對發音聲學參數有不同的影響。

2.4.3 Log- Likelihood 演進圖

為了證明在引入了韻律停頓標記以後，所得到的聲學模型較傳統僅考慮前後文相關的聲學模型來得佳，因此比較兩個模型在相同的訓練過程時平均每個音框(frame)的 log-likelihood 變化。

由於 PD-AM 多了韻律停頓的資訊，因此在使用決策樹做分群時，我們讓 PD-AM 的樹長得更深，共有 1849 個子節點，整體的高斯混合數(Gaussian mixture number)達 29584；傳統不與韻律相依之聲學模型(prosody-independent acoustic model, PI-AM)則有 1595 個子節點，整體的高斯混合數達 28710；兩者的參數量大致上是相等的。

圖 2.7 為兩個模型在訓練過程時平均每個音框(frame)的 log-likelihood 變化。其中每個階段(stage)所代表的意義如下:

Stage 1: 以單音素模型做為初始化之 PD-AM/PI-AM，1 mixture

Stage 2: 以 EM-Algorithm 重新評估 PD-AM/PI-AM 之 HMM 參數，1 mixture

Stage 3: 以決策樹來建立 tied-state 之 PD-AM/PI-AM，1 mixture

Stage 4: 增加 PD-AM/PI-AM 的 mixture 至 2 mixtures，且以 EM-Algorithm 重新評估 PD-AM/PI-AM 之 HMM 參數

Stage 5: 增加 PD-AM/PI-AM 的 mixture 至 4 mixtures，且以 EM-Algorithm 重新評估 PD-AM/PI-AM 之 HMM 參數

Stage 6: 增加 PD-AM/PI-AM 的 mixture 至 8 mixtures，且以 EM-Algorithm 重新評估 PD-AM/PI-AM 之 HMM 參數

Stage 7: 增加 PD-AM/PI-AM 的 mixture 個別至 16 mixtures 和 18 個 mixtures，且以 EM-Algorithm 重新評估 PD-AM/PI-AM 之 HMM 參數

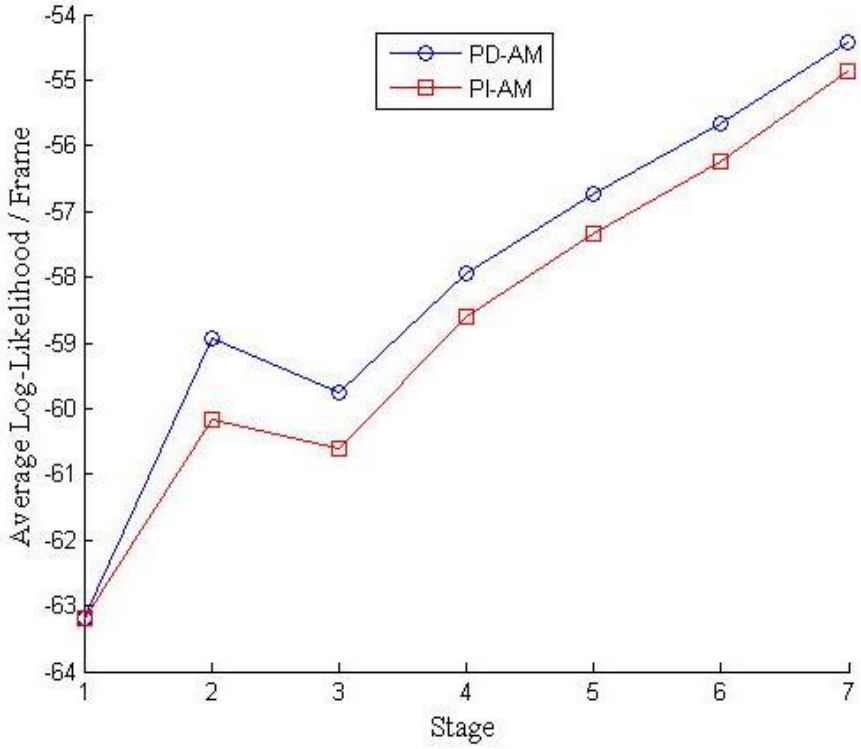


圖 2.7：PD-AM 及 PI-AM 訓練過程之 log-likelihood/frame 演進

由此圖可看出 PD-AM 在訓練的過程中其平均每個音框的 log-likelihood 演進較 PI-AM 來的好，說明在加入韻律停頓的資訊後，PD-AM 能較 PI-AM 學到更好的模型分布。

2.5 韻律停頓對於音節中所有音素影響之探討

本研究認為韻律停頓對於頻譜會造成影響，且直觀的認定音節邊界之音素受到其相鄰之韻律停頓的影響應是最為明顯的。音節內之音素，如介音，並不太會受到跨音素之韻律停頓的影響。子音及韻尾也較不會分別受到音節後及音節前之韻律停頓的影響。然而，我們在此仍然訓練了一種聲學模型進行了實驗以驗證我們的想法。在 2.5.1 節中，將介紹此聲學模型是如何建立的；在 2.5.2 節中將由決策樹分析該模型。

2.5.1 模型之建立

2.4 節中訓練的聲學模型，僅有音節邊界的音素會受到相鄰之韻律停頓的影響；在此，要建立一個聲學模型是構成音節所有的音素皆會受到音節前、後之韻律停頓之影響。其文本由音節層級展開至音素層級的展開的方式將如圖 2.8 所示，接著一樣照圖 2.5 的流程來訓練。

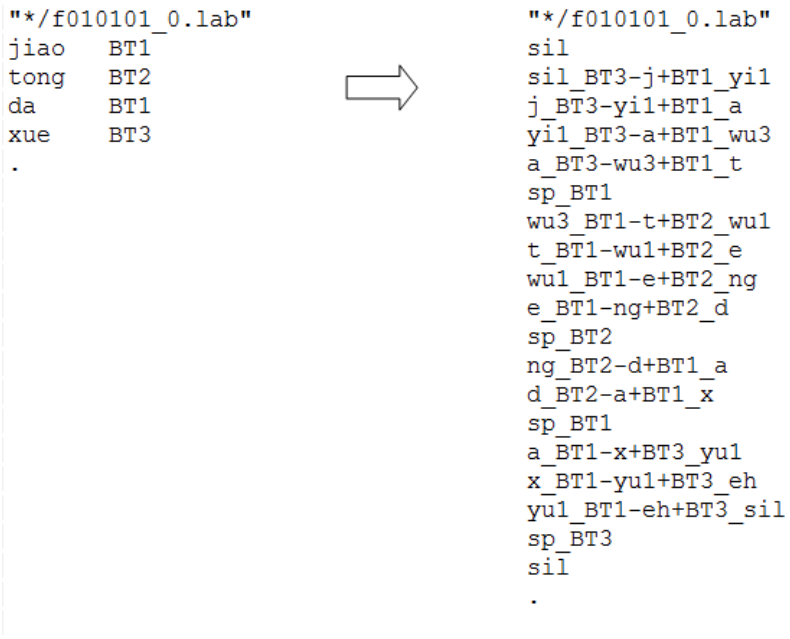


圖 2.8：訓練過程文本由音節層級展開至音素層級示意圖

2.5.2 模型分析

在 2.4 節中曾提及，在決策樹中越上層被問到的問題代表其對整體資料的分群影響越大。因此，我們將觀察該模型以下幾點:

1. 子音的決策樹中，其前三分之一的問題中與音節後之韻律停頓相關之個數。
2. 韻尾的決策樹中，其前三分之一的問題中與音節前之韻律停頓相關之個數。
3. 介音的決策樹中，其前三分之一的問題中與音節前、後之韻律停頓相關之問題個數。

若是欲觀察的韻律停頓相關問題並沒有在訓練模型的決策樹上層就出現，則證明該韻律停頓對於該音素的影響微乎其微。

首先，關於第 1 點，本研究認為相較於子音受到音節後之韻律停頓的影響，其受到相鄰音素或是音節前之韻律停頓的影響更明顯。由表 2.5 的統計結果，在多達 21 個子音(63 顆決策樹)中，其前三分之一的問題裡，僅一個問題是與音節後之韻律停頓相關的問題，也證明子音是不需要跨過後面相鄰的音素去考慮音節後之韻律停頓的影響。

表 2.5：子音的決策樹中，其前三分之一的問題與音節後之韻律停頓相關之個數

狀態	問題個數
第 1 個狀態	0
第 2 個狀態	1
第 3 個狀態	0

關於第 2 點，本研究認為相較於韻尾受到音節前之韻律停頓的影響，其受到相鄰音素或是音節後之韻律停頓的影響更明顯。由表 2.6 的統計結果，在 4 個韻尾(分別為 en, ng, yi3, wu3，共 12 顆決策樹)中，其前三分之一的問題裡，沒有任何問題是與音節前之韻律停頓相關的，證明韻尾是不需要跨過前面相鄰的音素去考慮音節前之韻律停頓的影響。

表 2.6：韻尾的決策樹中，其前三分之一的問題與音節前之韻律停頓相關之個數

狀態	問題個數
第 1 個狀態	0
第 2 個狀態	0
第 3 個狀態	0

關於第 3 點，本研究認為相較於介音受到音節之前、後之韻律停頓的影響，其受到相鄰音素的影響更明顯。由表 2.7 的統計結果，3 個介音(分別為 yi1, wu1, yu1)各自狀態的決策樹中，只有 2 個狀態其前三分之一的問題裡，各自有 1 個問題是與音節前、後之韻律停頓相關的。進一步觀察，在表 2.7 中，yi1 的第 3 個狀態出現的韻律停頓問題是音節後之韻律停頓問題，出現在該顆樹共 30 個問題中的第 9 個，因此對其資料分群的影響力並不大；wu1 的第 3 個狀態出現的韻律停頓問題則出現在該顆樹共 30 個問題中的第 8 個，因此也對其資料分群的影響力並不大。證明介音應該是不需要考慮音節前、後之韻律停頓的影響。

表 2.7：介音的決策樹中，其前三分之一的問題中與音節前、後之韻律停頓相關之問題個數

狀態	問題個數
yi1 第 1 個狀態	0
yi1 第 2 個狀態	0
yi1 第 3 個狀態	1
wu1 第 1 個狀態	0
wu1 第 2 個狀態	0
wu1 第 3 個狀態	1
yu1 第 1 個狀態	0
yu1 第 2 個狀態	0
yu1 第 3 個狀態	0

由此證明，音素並不太會受到跨音素的韻律停頓影響，因此在建立聲學模型時，僅需使音節邊界的音素受到相鄰之韻律停頓之影響即可。

第三章 音節辨認之實驗結果及分析

本研究進行大詞彙語音辨認分為兩個階段，第一階段先利用聲學模型進行音節的辨認產生音節圖(syllable lattice)，因此本章將使用與韻律停頓相依之聲學模型(PD-AM)進行音節辨認的實驗，其流程如圖 3.1 所示。在 3.1 節中，會介紹詞典(lexicon)的建立方式；3.2 節，會針對音節辨認時的 grammar 做說明；3.3 節則介紹辨認時，模型的路徑如何展開以構成音節；3.4 節則比較 PD-AM 與 PI-AM 之音節辨認率(recognition rate)及涵蓋率(coverage rate)；3.5 節分析 PD-AM 解碼出之 BT3 是否可幫助辨認呼吸群組句的邊界。

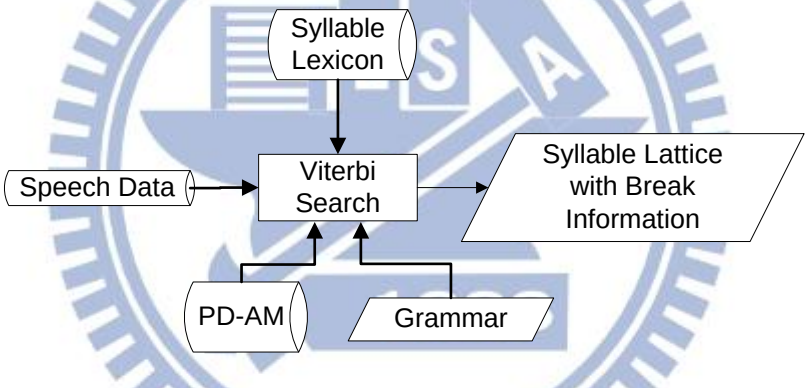


圖 3.1：音節辨認之系統架構圖

3.1 詞典之建立

詞典包含了中文的 411 個音節，及其對應的音素序列(phone sequence)，並針對四種韻律停頓做展開，使辨認的結果能輸出每個音節後的韻律停頓資訊。展開後的詞典共有 1644 個辨認單元。表 3.1 以兩個音節做為詞典建立的範例。

表 3.1：詞典建立之範例

詞典單元(*SYL_BT*)	P1	P2	P3	P4
yin_BT0	yi	e	en_BT0	
yin_BT1	yi	e	en_BT1	
yin_BT2	yi	e	en_BT2	
yin_BT3	yi	e	en_BT3	
bian_BT0	b	yi	a	en_BT0
bian_BT1	b	yi	a	en_BT1
bian_BT2	b	yi	a	en_BT2
bian_BT3	b	yi	a	en_BT3
		⋮		

3.2 Grammar 設計

根據相關研究[11]，語音的韻律呈現階層式的架構，為了保持語音之流暢性，音節間的韻律停頓其出現機率是不相等的，例如連續出現兩個音節其後皆為長停頓(BT3)的可能性極低，因此在辨認時，應把韻律停頓相接之機率考量近來，刪除一些機率較低的路徑，縮小辨認的空間。藉由訓練語料中 PLM 標記出之韻律停頓，統計出韻律停頓相接之機率，以 bi-gram 的方式結合至辨認的網路中，如圖 3.2 所示。

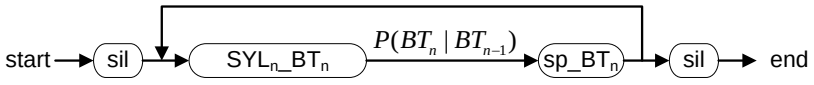


圖 3.2：音節辨認之 grammar

3.3 辨認網路之模型展開

本研究的聲學模型採用跨音節之三連音素，在音節邊界時同時考慮韻律停頓之影響。圖 3.3 介紹在使用 PD-AM 辨認音節序列“yu-yin-bian-ren”(語音辨認)時，如何由音節展開至 tri-phone model 之示意圖。4 個音節對應的音素序列為“yu, yi, e, en, b, yi, a, en, r, e, en”，每個音節後之韻律停頓分別為“BT1, BT2, BT1, BT3”。

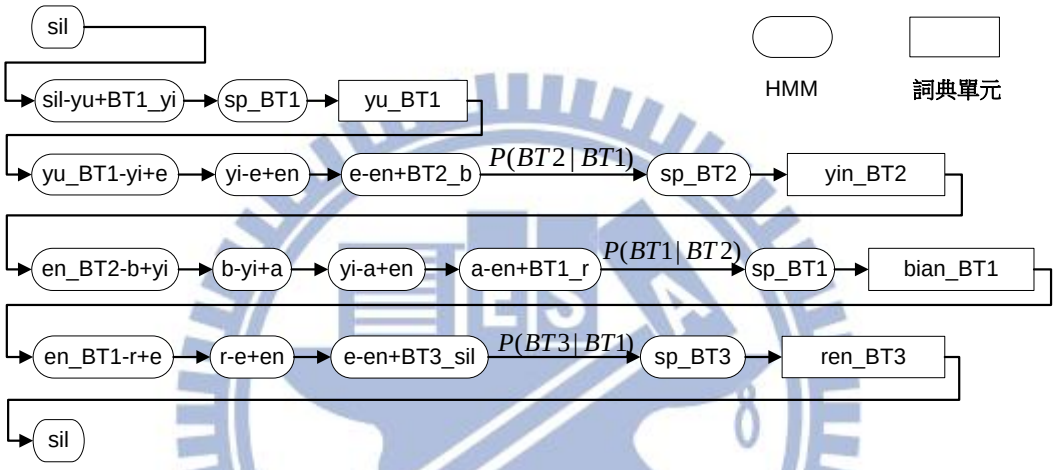


圖 3.3：音節辨認之模型展開示意圖

3.4 音節辨認率及涵蓋率之比較

本實驗所使用的辨認工具為 HTK3.4[13]，針對 TCC300 語料庫中的 226 段長句分別用 PD-AM 及 PI-AM 進行音節的辨認，實驗的結果如表 3.2。如表所示，PD-AM 之音節的辨識率可達 68.20%，較 PI-AM 的 67.55%來得高；不僅如此，insertion 及 deletion error 也下降了。

表 3.2：不同聲學模型之中文音節辨認實驗結果

Model	Hit	Sub.	Ins.	Del.	Total	Recognition Rate
PI-AM	18235	7889	354	348	26472	67.55%
PD-AM	18364	7790	311	318	26472	68.20%

進一步分析測試語料中出現頻率較高，且辨識率改善超過 6.5%的音節，如表 3.3 所示。

表 3.3：測試語料中出現頻率較高且辨識率改善較多之音節

音節	改善之辨識率	出現次數
da	6.86%	175
ta	9.94%	161
zhe	7.64%	157
zhen	11.11%	72
yi	7.04%	824
mu	8.00%	75
xie	9.21%	76
liu	8.79%	91
xian	8.43%	166
ling	10.39%	77
ming	10.00%	80
wang	7.62%	105
yue	9.72%	72

由此表可看出，辨識率改善較多之音節不乏含鼻音韻尾及爆破音之音節，而這也符合語言學上的知識。因為在發鼻音韻尾時，若其後沒有停頓會破壞其發音方式；而在發爆破音時，因嘴唇需先緊閉，前面也需要停頓。因此，在加入韻律停頓後，能將訓練語料中之鼻音韻尾或爆破音有效區分有無停頓的兩群，使其不互相汙染，進而訓練出較好的聲學模型，改善辨識率。

此外，我們也比較的 PD-AM 和 PI-AM 在不同的音節圖大小下(以 arc 的數量

計算)的音節涵蓋率，如圖 3.4。由圖可看出， PD-AM 仍略優於 PI-AM。

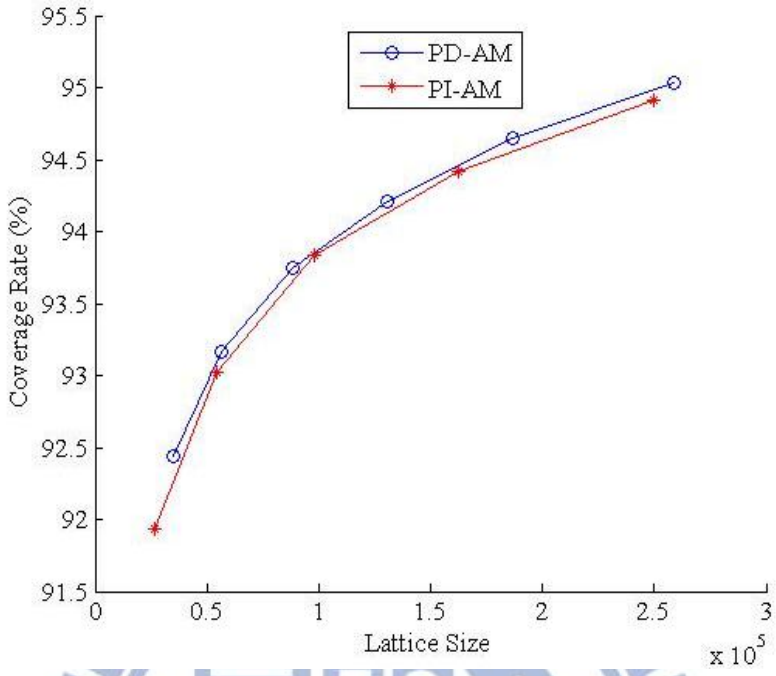


圖 3.4：PD-AM 與 PI-AM 在不同的音節圖大小下，音節涵蓋率之比較

雖然整體的效能是有改善的，但其幅度並不大。由決策樹之分析中，可發現出現在上層之韻律停頓問題以 BT3 為主，其餘之韻律停頓對於頻譜之影響似乎並不明顯。這可能是 PLM 標記出之韻律停頓在同一類別中仍有少部分是 outlier，如標記為 BT1 的仍有少部分的停頓時間過長，標記為 BT2 的有少部分的停頓時間過短，甚至是不可察覺之停頓，使模型在訓練時產生混淆。也有可能是停頓不明顯之韻律停頓對頻譜的影響幾乎是相同的，在訓練模型時，只需將其標記為 BT3 和非 BT3 的兩群。

3.5 呼吸群組句邊界辨認之分析

本研究加入韻律停頓資訊至訓練聲學模型除了期望能使聲學模型的效能變好以外，也期望其解碼出的 BT3，能幫助判斷何處為句子的邊界，縮小辨認的搜尋空間。因此在本節中，將會討論 PD-AM 在進行音節辨認時，所解碼出的 BT3

是否能幫助辨認呼吸群組句的邊界。

假設以停頓時間超過 200ms 的做為呼吸群組句的邊界，其出現數量及在這之中 PD-AM 的能解碼出 BT3 的次數如表 3.4 所示。由表 3.4 可觀察出有 97.4% 停頓時間超過 200ms 的停頓皆能解碼出 BT3，證明 PD-AM 解碼出之 BT3 的確可幫助辨認句子的邊界，可用其將長句切割成數個短句，縮小在辨認時展開之搜尋網路。

表 3.4：呼吸群組其句尾能解碼出 BT3 之數量比較

	呼吸群組句	句尾為 BT3
數量	2212	2154

第四章 以有限狀態機實現大詞彙語音辨認

在使用與韻律停頓相依的聲學模型(PD-AM)辨認產生音節圖(syllable lattice)後，接著我們將使用加權有限狀態轉換機中的組合演算法對音節圖做構詞，將其變為詞圖(word lattice)，最後再整合 n-gram 的語言模型對詞圖重新計分(rescoring)，實現第二階段的詞辨認。因此在本章節中，將會探討以下幾點: 4.1 節介紹有限狀態機及組合演算法；4.2 節將簡介語言模型之架構及語料庫；4.3 節則介紹如何以有限狀態機整合音節圖及詞典進行構詞，最後再整合語言模型，實現完整的大詞彙語音辨認；4.4 節則會呈現實驗的結果並加以分析。

4.1 有限狀態機及組合演算法之簡介

4.1.1 加權有限狀態轉換機

加權有限狀態轉換機(Weighted Finite-State Transducer, WFST)可視為一個具方向性之狀態(state)轉移圖。在狀態轉移(transition)時，會帶有輸入字元(input symbol)、輸出字元(output symbol)與權重(weight)。圖 4.1 提供一範例，其中粗線圈代表初始狀態(initial state)，雙線圈表示終止狀態(final state)，轉移邊上的三個數字由左至右分別代表輸入字元、輸出字元及權重。

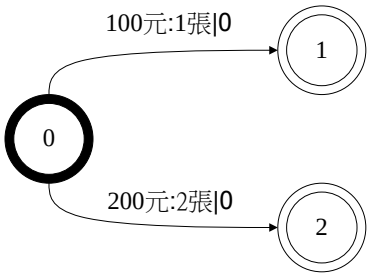


圖 4.1：售票機之加權有限狀態轉換機

每個 WFST 基本上由六個元素(Q, I, F,Σ,Δ,δ)所構成，其中：

1. Q：所有狀態的集合。在圖 4.1 中，Q={0, 1, 2}，含初始狀態與結束狀態共 3 個狀態。
2. I：初始狀態，表示有限狀態機的唯一的初始狀態。在圖 4.1 中，I={0}。
3. F：終止狀態，表示有限狀態機結束的狀態，至少要含有一個以上的終止狀態。在圖 4.1 中， F={1, 2}。
4. Σ：所有可接受的輸入字元集合。在圖 4.1 中，Σ={100 元, 200 元}。
5. Δ：所有的輸出字元集合。在圖 4.1 中，Δ={1 張, 2 張}。
6. δ：轉移函式。表示某來源狀態接受輸入字元後，會轉移到哪一個目標狀態。

以下將對一些專有名詞做深層之解釋：

1. 狀態：

WFST 含有有限數量個狀態，這些狀態中必須有一個初始狀態與一個以上的結束狀態。一開始由初始狀態出發，接受輸入字元序列後，經過一連串的狀態轉移，當最後一個轉移完成後，若停留在終止狀態，表示此條路徑是可接受(accept)的；反之則拒絕輸出(reject)。

2. 轉移：

狀態與狀態間的轉移由上述的轉移函式 δ 所定義，每個轉移需帶有來源狀態 s(t)、目的狀態 d(t)、輸入字元 i(t)、輸出字元 o(t)與權重 w(t)，其中 t 代表不同的轉移。描述一個轉移時寫作(I:O|W)，I 表示 input symbol、O 表示 output symbol、W 表示 weight。

3. 空轉移：

我們允許轉移上的輸入與輸出字元為 ε (epsilon)。當輸入字元為 ε 時，表示該轉移不需要輸入字元就可以轉移到下一個狀態；當輸出字元為 ε 時，表示經過此轉移並不會有輸出字元。在設計 WFST 時，會藉由空轉移來表示圖形的特性。

4. 權重：

除了在轉移上會帶有權重之外，每個結束狀態也可以再賦予權重。在設計描述語音辨識所用之 WFST 時，一般會對各模型之分數取 negative nature log，因此權重可視為經過此轉移所需付出的代價(cost)，因此在尋找最佳路徑時可視為搜尋累積權重或代價最小的路徑。

4.1.2 組合演算法

WFST 的一個最大特性是可以將不同層次的 WFST 進行組合演算法(Compose Algorithm)操作，將其整合以得到一個最終的 WFST，本研究也是依此演算法將音節圖與詞典的 WFST 整合，形成一個詞圖。

給定兩個加權有限狀態轉換機 A 與 B，將 A 的輸出字元與 B 的輸入字元結合，進而將 A、B 整合成一個新的加權有限狀態轉換機 C，寫作 C=A。B。每個 C 的狀態、轉移都是由 A 跟 B 的狀態與轉移所組成，其所攜帶的權重則是由 A 跟 B 相加而得，並且只留下可成功走完的路徑。圖 4.2 及 4.3 提供一範例。

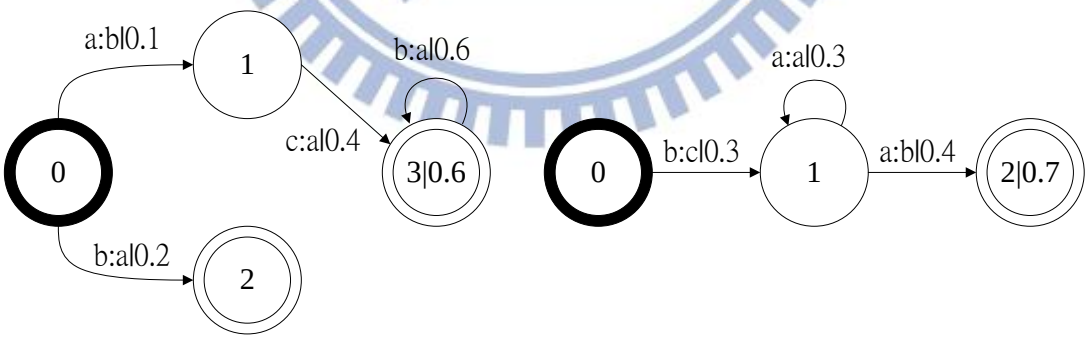


圖 4.2：加權有限狀態轉換機 A(左)與 B(右)

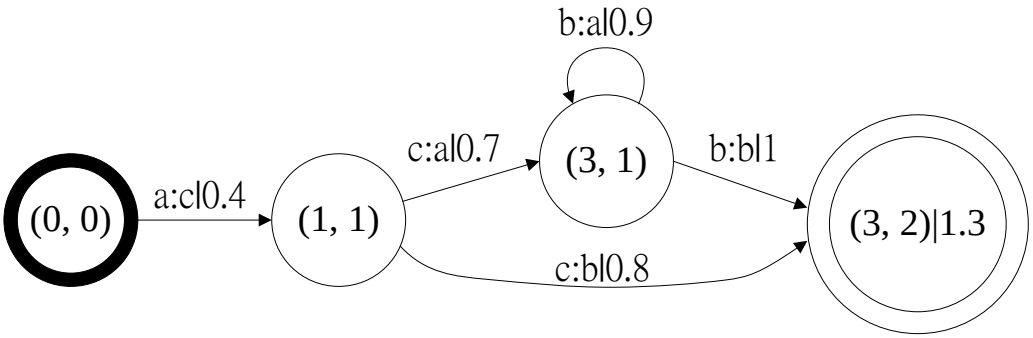


圖 4.3：加權有限狀態轉換機 C=A。B

執行組合演算法時，加權有限狀態轉換機 A 下方的轉移因其輸出字元 a 無法與 B 之輸入字元 b 相接，因此不會出現在加權有限狀態轉換機 C 中。此外，在 B 的狀態 1，有兩條轉移其輸入字元皆為 a，皆可以與 A 之狀態 1 其轉移之輸出字元 a 結合，因此在 C 出現了兩條路徑，而本研究也是利用此特性，將音節串轉化成所有可能構成的詞串，最終將音節圖轉變為詞圖。

4.2 語言模型之建立

所有的語言都有其文法規則，利用文法規則所建立出的機率模型稱為語言模型。在大詞彙連續語音辨認時，會利用語言模型，考慮前後詞彙的關連性，期望能使輸入的語音辨認出合理且有意義的詞串。在本節中，將會說明本論文採用的語言模型，其建立之資料來源及流程。

4.2.1 語言模型架構

一般在建立語言模型時，是以詞(word)做為基本單位，現假設有一詞串共有 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，則產生該詞串之機率可拆解如下：

$$
\begin{aligned}
&P(w_1, w_2, ...w_N) \\
&= P(w_1)P(w_2 \mid w_1)\cdots P(w_N \mid w_1, w_2, ..., w_{N-1}) \\
&= \prod_{i=1}^{N} P(w_i \mid w_1, w_2, ..., w_{i-1})
\end{aligned} \tag{4.1}
$$

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，因此我們以 n-gram 的機率形式去近似(4.1)式，如下所示：

$$
P(w_1, w_2, ...w_N) \cong \prod_{i=1}^{N} P(w_i \mid w_{i-n+1}, ..., w_{i-1}) \tag{4.2}
$$

其中

$$
P(w_i \mid w_{i-n+1}, ..., w_{i-1}) = \frac{Count(w_{i-n+1}, ..., w_i)}{Count(w_{i-n+1}, ..., w_{i-1})} \tag{4.3}
$$

$Count(\cdot)$ 表示詞串的出現次數。由於 n-gram 語言模型是統計式的模型，如果訓練語料中沒出現該詞語組合，就無法預估其機率；然而，一個詞串即使在訓練語料中沒有出現，並不代表在測試語料中不會出現，且若 $Count(w_{i-n+1}, ..., w_i)$ 很小時，所計算出的機率也是不準確的。為此，我們以後撤平滑化(back-off smoothing)來調整模型的機率分佈，使語言模型中所有的 n-gram 機率均能被良好的估計。機率預估式被改寫如下：

$$
\begin{aligned}
&P(w_i \mid w_{i-n+1}, ..., w_{i-1}) \\
&= \begin{cases}
a(w_{i-n+1}, ..., w_{i-1})P(w_i \mid w_{i-n+2}, ..., w_{i-1}) & , Count(w_{i-n+1}, ..., w_i) = 0 \\
d_a \cdot \dfrac{Count(w_{i-n+1}, ..., w_i)}{Count(w_{i-n+1}, ..., w_{i-1})} & , 1 \le Count(w_{i-n+1}, ..., w_i) \le k \\
\dfrac{Count(w_{i-n+1}, ..., w_i)}{Count(w_{i-n+1}, ..., w_{i-1})} & , Count(w_{i-n+1}, ..., w_i) > k
\end{cases}
\end{aligned} \tag{4.4}
$$

(4.4)式中 $a(w_{i-n+1}, ..., w_{i-1})$ 為經過正規化(normalization)的後撤係數，且需滿足下條件式：

$$
\sum_{w \in V} P(w_i = w \mid w_{i-n+1}, ..., w_{i-1}) = 1 \tag{4.5}
$$

觀察(4.4)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 的機率並乘上一個後撤係數，用以產生一個適當的機率值取代機率 0

的出現；*Count*(·)的值很小時，則將原始的 n-gram 機率乘上一個小於 1 的值d_a (Discount Coefficient Factor)來進行平滑，d_a是依據 Good-Turning discounting 所計算出的，並將扣除的機率值分給詞串沒有出現的 n-gram 機率使用。

4.2.2 語言模型語料庫

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，簡稱 NSS：

1. 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。

2. 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。

3. 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。

NSS 在經過 CRF[14]斷詞器斷詞以及文字正規化[15]的處理後，得到詞的總數量為 122,541,303 個，字數為 231,225,705。而要建立一個完善的語言模型，另一項重要關鍵便是詞典的選擇，由於受限於記憶體大小，僅能將較常出現的詞整理在詞典內提供建立訓練語言模型使用。在本研究中，詞典的選擇方式則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。

4.3 有限狀態機的整合

為了以有限狀態機實現第二階段的詞辨認，需先將帶有韻律停頓資訊的音節圖、詞典及語言模型轉化成加權有限狀態機的圖形，利用組合演算法結合音節圖及詞典以進行構詞，產生詞圖；接著，同樣利用組合演算法，詞圖再整合語言模型，對詞圖重新計分，最後再搜尋一條權重最小之路徑輸出辨認答案。整個系統架構如圖 4.4 所示。

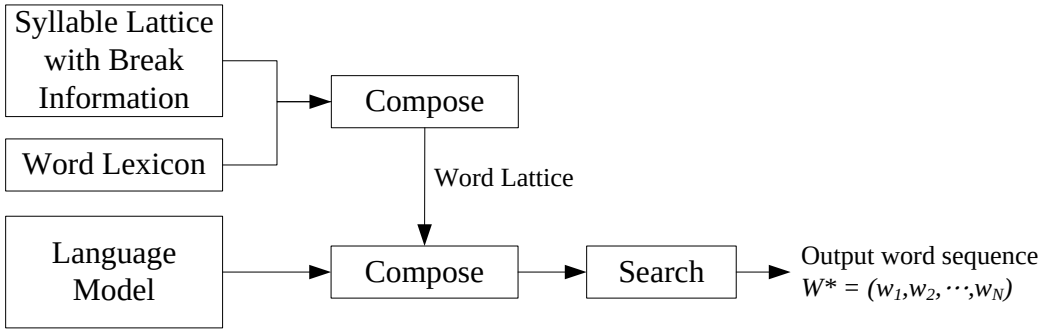


圖 4.4：對音節圖構詞實現詞的辨認之系統架構圖

在本實驗中進行組合演算法及搜尋權重最小之路徑時，是採用 Google Research and NYU's Institute 所開發之 OpenFst library[16]進行；在建立語言模型之 WFST 則採用 Idiap Research Institute 所開發之 Weighted Finite State Transducer Decoder-Juicer[17]。以下將針對各有限狀態機做進一步的說明。

4.3.1 帶有韻律停頓資訊的音節圖

由 HTK 產生出帶有韻律停頓資訊的音節圖(syllable lattice)，我們將其點(node)用 WFST 的狀態(state)表示，邊(arc)用轉移(transition)表示，轉移上的輸入及輸出字元皆對應到邊所攜帶含有韻律停頓資訊的音節，權重則是將邊上之聲學模型分數(log-likelihood)取負號。圖 4.5 提供一範例。

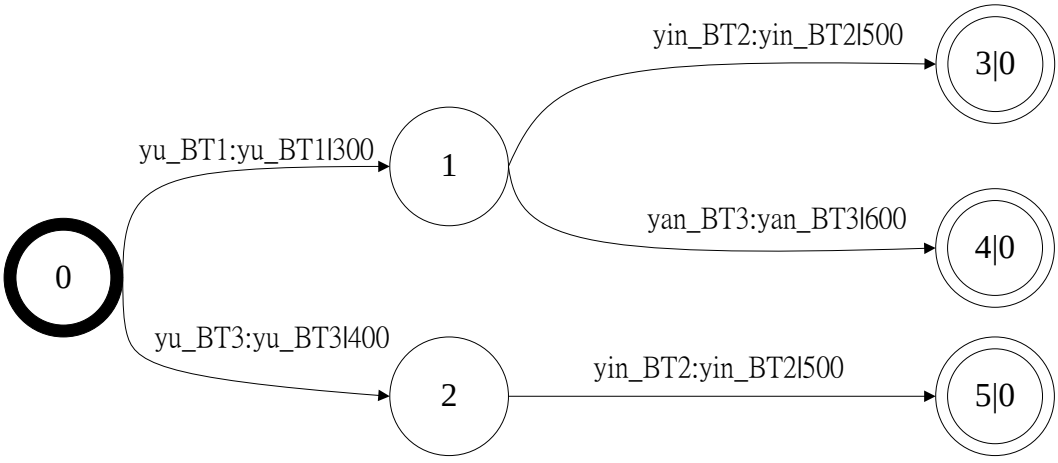


圖 4.5：音節圖之 WFST

4.3.2 詞典

藉由聲學模型解碼出音節圖以後，必須藉由詞典，同時搭配組合演算法將其展開成詞圖。在此，我們將依照詞頻挑選出的六萬詞詞典，依照韻律停頓的種類展開，其中 intra-word 有三種韻律停頓，分別為 BT0、BT1 及 BT2，inter-word 有四種韻律停頓，分別為 BT0、BT1、BT2 及 BT3。Intra-word 之所以不將 BT3 一併展開是因為詞內不應出現長停頓，如此一來稍後在音節圖上構詞時，並不會跨過長停頓做構詞，達到我們欲使用韻律停頓來判斷何處為詞的邊界之目的。表 4.1 以"語音"這個詞為詞典建立之範例。

表 4.1：詞典建立之範例

詞典單元	SYL1	SYL2
語_BT0 音_BT0	yu_BT0	yin_BT0
語_BT0 音_BT1	yu_BT0	yin_BT1
語_BT0 音_BT2	yu_BT0	yin_BT2
語_BT0 音_BT3	yu_BT0	yin_BT3
語_BT1 音_BT0	yu_BT1	yin_BT0
⋮	⋮	⋮
語_BT2 音_BT3	yu_BT2	yin_BT3

接著我們將詞典轉成 WFST 之格式，轉移上的輸入為詞對應的音節串，輸出則為詞。圖 4.6 為一範例。在該範例中包含了"語_BT1"，"語_BT3"，"音_BT2"，"語_BT1 音_BT2" 及 "語_BT1 言_BT3"等五個詞。終止狀態至起始狀態的空轉移，是為了在音節圖上每構完一個詞後，能繼續構詞。

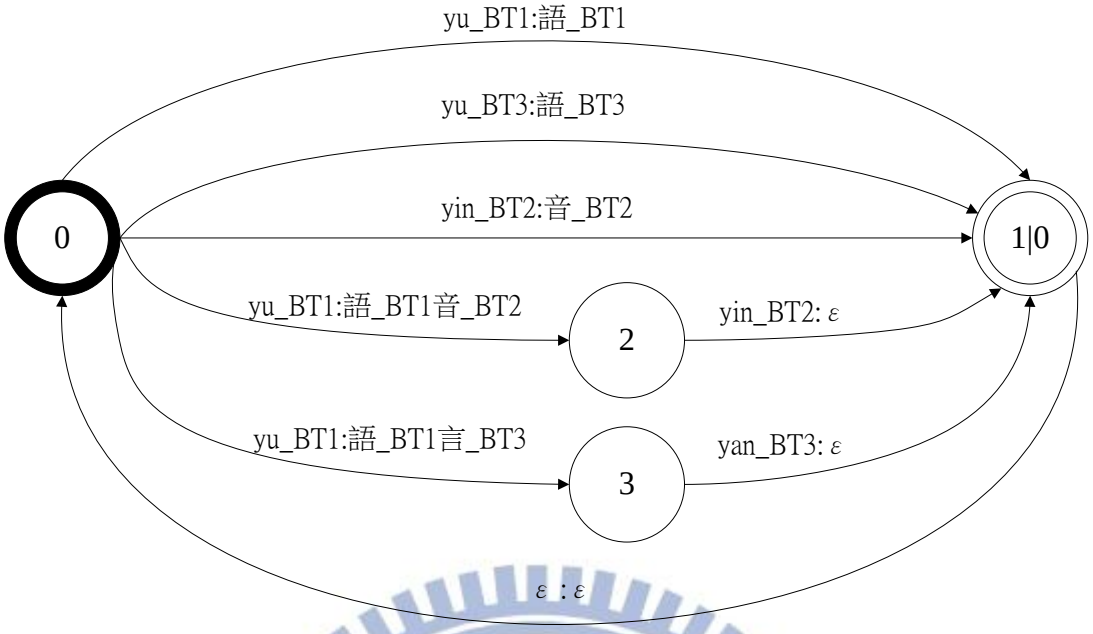


圖 4.6：詞典之 WFST

若以圖 4.5 為例子，其與圖 4.6 組合之結果將如圖 4.7 所示，轉移上的輸出字元即為能構成的詞。觀察圖 4.7 可發現，由於在詞典中並沒有定義“語_BT3 音_BT2”(因為 intra-word 不應出現長停頓 BT3)，使得圖 4.5 下面的路徑並無法產生“語音”這個二字詞，達到能以音節圖上的韻律停頓資訊來協助構詞。

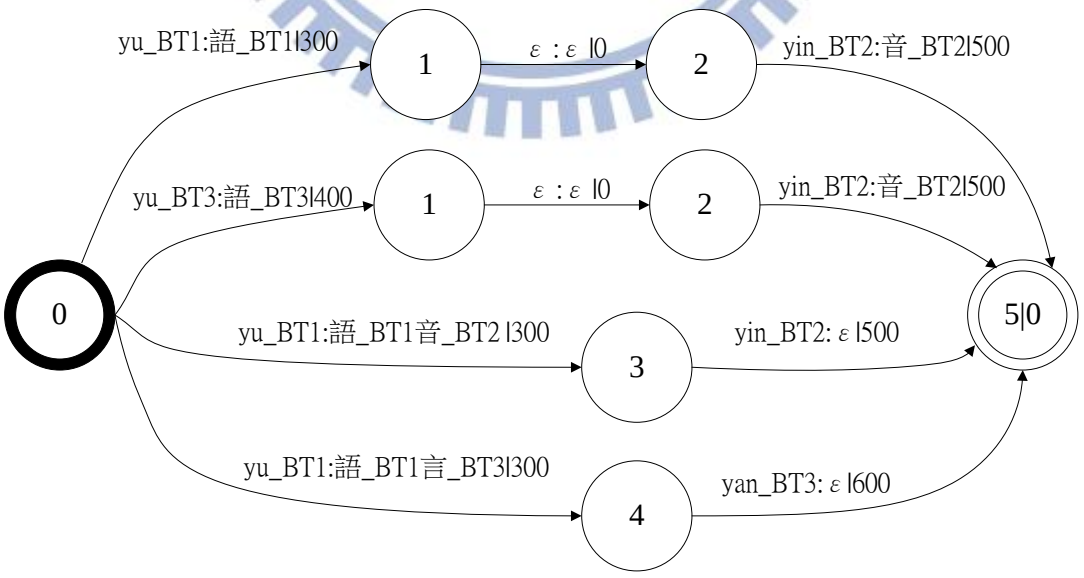


圖 4.7：以組合演算法整合圖 4.5 與圖 4.6 產生的 WFST

4.3.3 語言模型

在本研究中使用 n-gram 來描述語言模型。圖 4.8 提供一個 bi-gram 語言模型轉換為 WFST 圖形的範例。狀態內的詞代表其走過的詞，從圖中可以觀察到：當沒有有效的輸入時，就藉由空轉移走到狀態α，同時也帶上了一個後撤的分數，而由狀態α走到其他狀態所帶上的分數，就是已經後撤到 uni-gram 的分數。

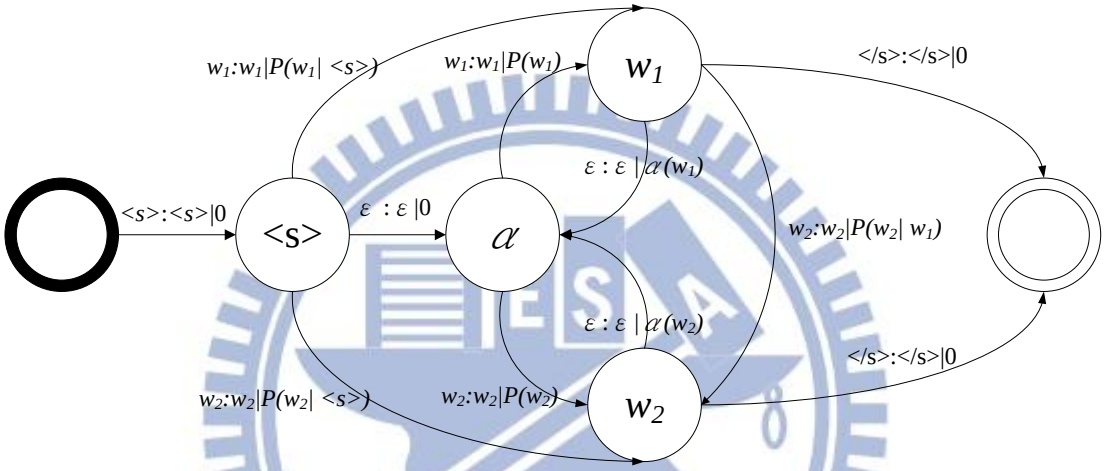


圖 4.8：雙連語言模型之 WFST

由於本研究並沒有建立與韻律相依的語言模型，因此在構詞完畢產生之 WFST 圖形的輸出字元上，會捨棄韻律停頓的資訊，使其能與語言模型進行組合演算法，但輸入字元仍能保有韻律停頓的資訊，可供往後結合[18]所提出之 Break Syntax Model 以及 Break Acoustic Model，引入如 pitch、energy 及 pause duration 等韻律參數。

4.4 大詞彙語音辨認實驗結果及分析

4.4.1 節將會比較本研究所提出之兩階段式辨認系統下，PD-AM 及 PI-AM 之辨認的結果；4.4.2 節則進一步比較 PD-AM 之兩階段式辨認與 PI-AM 之傳統一階段式辨認結果。

4.4.1 PD-AM 與 PI-AM 之兩階段式辨認結果比較

如圖 4.4 所示，我們分別以 PD-AM 與 PI-AM 辨認出音節涵蓋率相近之音節圖，進行構詞產生詞圖後，再結合 tri-gram 的語言模型輸出辨認詞串。其詞辨認結果如表 4.2 所示。

表 4.2：不同聲學模型搭配 tri-gram 語言模型之辨認結果

聲學模型	詞辨認率	詞涵蓋率	音節涵蓋率
PI-AM	70.08%	84.11%	94.82%
PD-AM	70.25%	83.26%	94.85%

由表 4.2 可發現， PD-AM 在詞的辨認率略優於 PI-AM。除了聲學模型的改善有助於其提高辨認率以外，PD-AM 產生之音節圖會利用解碼出的韻律停頓資訊 BT3，使其無法跨過 BT3 做構詞，達到以韻律停頓刪除音節符合構成詞典的詞但韻律不合理的搶詞狀況。表 4.3 列出了一些搶詞狀況的改善範例。其中 a)為正確文本，b)為 PI-AM 之詞辨認結果，c)為 PD-AM 之詞辨認結果，包含其解碼出之 BT3。

表 4.3：搶詞狀況的改善

a)	……改良 的 新 **賭法 既** 不 怕 作弊……
b)	……大量 的 刑度 **罰金** 不 怕 作弊……
c)	……大量 的 刑度 **法(BT3) 既** 不 怕 作弊……

a)	……協助 **他 防止** 危險 的 外人……
b)	……協助 **套房** 此 危險 的 外人……
c)	……協助 **他(BT3) 放置** 危險 的 外人……
a)	……吳桐潭 兩名 **要角 竟** 在……
b)	……巫統 他倆 民謠 **較勁** 再……
c)	……不動產 兩名 **要角(BT3) 竟** 在……
a)	……選手 **訓練站 但** 八月……
b)	……選手 去年 **炸彈** 八月……
c)	……選手 去年 **在(BT3) 但** 八月……

由表 4.2 也發現，PD-AM 產生之詞圖在詞的涵蓋率略低於 PI-AM 所產生的。其原因為測試語料中，有部分的語者在朗讀時因為遲疑而導致詞內出現長的停頓，使該時間點辨認出韻律停頓 BT3 無法構出正確的詞。此外，在停頓不明顯的時間點，音節圖上容易出現許多有 BT0 及 BT1 的相同音節，然而其聲學模型的差異並不大，在有限的 token 數下(每個 node 允許的來源 arc 數)，使下個時間點之音節無法與更多不同的音節來源相連結，進而影響詞的涵蓋率及辨認率。

4.4.2 PD-AM 之兩階段式辨認與 PI-AM 之一階段式辨認結果比較

由於 PD-AM 需將詞典依照韻律停頓做展開，詞的數量過多，而受限於記憶體的大小，使 PD-AM 並無法直接結合詞典、與語言模型結合進行詞的辨認，這也是本研究採用兩階段式辨認的主要原因；然而，PI-AM 則無此限制。若我們將 PI-AM 以一階段式辨認的方式結合詞典及 bi-gram 的語言模型產生詞圖後，再以 tri-gram 之語言模型重新評分，其詞辨認率及涵蓋率如表 4.4 所示。此處所用的辨認工具為 HTK3.4。

表 4.4：PI-AM 之一階段式詞辨認率及涵蓋率

詞辨認率	詞涵蓋率
76.13%	89.11%

若將表 4.2 之 PD-AM 兩階段式的辨認結果與其相比，可發現 PD-AM 的詞辨認率低了 5.88%，兩者的詞涵蓋率差了 5.85%，而這也是辨認率下降的主要原因。經檢查發現，以有限大小的音節圖進行構詞，詞涵蓋率降低的主要的原因為:

1. PD-AM 在進行兩階段式辨認時，僅靠聲學模型產生發音相近之音節圖，若是語者發音不清楚，單靠聲學模型無法補救，在該時間點就無法構出該詞。PI-AM 因採用一階段式辨認，直接結合聲學模型及語言模型的分數搜尋辨認網路，其好處在於能及早將語言模型的分數加入，產生聲學參數不符合但語言參數符合之正確的詞，表 4.5 提供了一些範例。其中 a)為正確文本，b)為 PI-AM 產生之詞圖中詞涵蓋率最高之路徑，c)為 PD-AM 產生之詞圖中詞涵蓋率最高之路徑。

表 4.5：PI-AM 與 PD-AM 產生之詞圖中最佳詞涵蓋率路徑之比較

a)	……影響 較 大 的 是 高中 **學生** 的 選手……
b)	……影響 較 大 的 是 高中 **學生** 的 選手……
c)	……影響 較 大 的 是 高中 **絕症** 的 選手……
a)	……**小** 西瓜 主要 產區 已 自 **雲林縣** 麥寮 附近……
b)	……**小** 西瓜 主要 產區 已 移師 **雲林縣** 麥寮 附近……
c)	……**腔** 西瓜 主要 產區 已 自 **嶺 林嫌** 麥寮 附近……
a)	……專 供 **殘障** 人士 批購 **彩券** 之 用……
b)	……專 供 **殘障** 人士 批購 **彩券** 之 用……
c)	……專 供 **拆帳** 人士 批購 **裁軍** 之 用……

a) ……三月份 國際 **奧會** 派出 了 **考察團** 赴 北京 指出……
b) ……三月份 國際 **奧會** 派出 了 **考察團** 赴 北京……
c) ……三月份 國際 **鑰 奧會** 派出 了 **考察 竄** 赴 北京……

2. PD-AM 在產生音節圖時，即使在某個時間點上能夠產生正確詞之音節，但因不同時間點之音節可相連的 arc 有限，使其無法構出正確的詞，這種情況在長詞更為明顯，表 4.6 提供了長詞無法構出之範例。其中 a)為正確文本，b)為 PI-AM 產生之詞圖中詞涵蓋率最高之路徑，c)為 PD-AM 產生之詞圖中詞涵蓋率最高之路徑。

表 4.6：PD-AM 之兩階段式辨認無法構出長詞之範例

a) 戶 與 戶 之間 的 區隔 也 逐漸 由 鐵門 鐵窗 所 取代 **近在咫尺** 的 鄰居
b) 戶 與 國 之間 的 區隔 也 逐漸 由 鐵門 鐵窗 所 取代 **近在咫尺** 的 鄰居
c) 戶 與 戶 之間 的 許可 也 逐漸 由 鐵門 鐵窗 所 取代 **近代 子嗣** 的 鄰居
a) 大 魚 小 魚 **一網打盡**
b) 大 魚 小 魚 **一網打盡**
c) 大 魚 小 魚 **力挽 發青**
a) 在 **暗潮洶湧** 的 桌協 理事長 改選 前夕
b) 在 **暗潮洶湧** 的 桌協 理事長 改選 前夕
c) 在 **叛逃 熊** 的 桌協 理事長 改選 前夕

若要提高詞的涵蓋率，需增加音節圖之 arc 數，提高音節之涵蓋率及增加不同時間點之音節可相連的路徑，辨識率也有機會再提高。但本實驗 PD-AM 所採用之音節圖在構完詞後已經相當的大，若再擴大音節圖的大小，構完詞之後的搜尋空間會變得更大。PD-AM 兩階段式辨認與 PI-AM 一階段式辨認之詞圖其平均每個測試音檔的 arc 數比較如表 4.7 所示。

表 4.7：PD-AM 兩階段式辨認與 PI-AM 一階段式辨認之詞圖的 arc 數比較

聲學模型	arc 數
PI-AM	26534
PD-AM	980625

因此，在沒有詞典及語言模型的協助下，在有限的音節圖大小下，無法展開所有可能的相近音，若是語者發音不清楚單靠聲學模型更無法補救。此外，觀察音節圖上也可發現相鄰之時間點容易產生出許多相同音節；在停頓不明顯的時間點上更容易出現許多有 BT0 及 BT1 的相同音節。這些因素都導致詞的涵蓋率無法隨著詞圖大小的增加而有效提高，進而降低本研究之兩階段式的詞辨認率。因此，如何以音節涵蓋率較高之音節圖進行構詞，但使其產生之詞圖大小不要過大，是本研究所提出之兩階段式辨認需要解決的問題。若能在構詞的同時結合語言模型做 beam search，將一些分數低之路徑砍掉，則可以由音節涵蓋率較高、arc 數較多的音節圖，進行構詞，如此一來其展開之詞圖就可以避免 arc 過多的問題。

第五章 結論與未來展望

5.1 結論

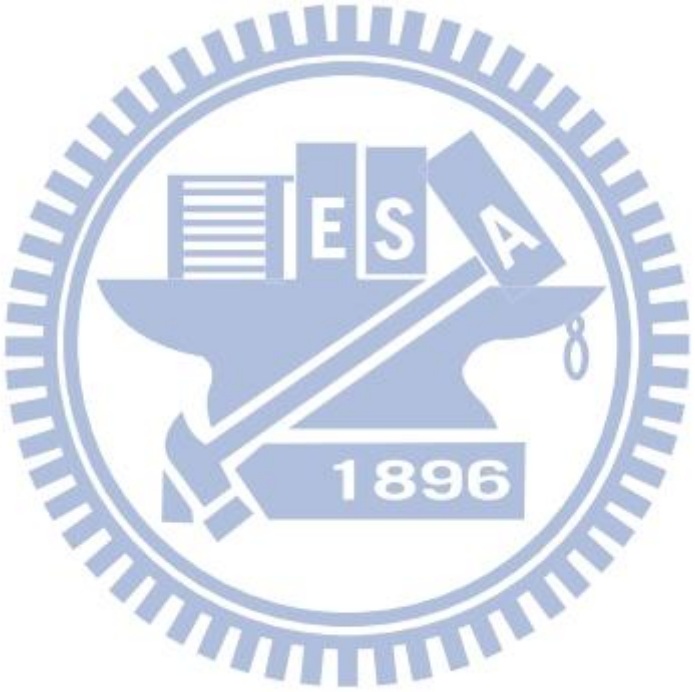
本研究利用 PLM 標記出的韻律停頓，將傳統的 tri-phone model 拓展至在音節邊節時，同時考慮韻律停頓的影響，建立一個與韻律相依之聲學模型(PD-AM)。由決策樹的分析發現韻律停頓對於發音聲學參數確實有影響，尤其又以長停頓(BT3)的影響最為明顯。本研究採用兩階段式的辨認系統，第一階段之音節辨認率由原本 67.55%改善至 68.20%，其解碼出之 BT3 也確實能幫助偵測呼吸群組句之邊界。在第二階段的詞辨認實驗中，除了聲學模型的改善有助於提高辨認率，其解碼出之 BT3 也能改善搶詞問題。

然而，其整體的改善幅度並不大，可能是 PLM 標記出之韻律停頓在同一類別中仍有少部分是 outlier，使模型在訓練時產生混淆。也有可能是停頓不明顯之韻律停頓對頻譜的影響幾乎是相同的，無需將其分為四大類。此外，與一階段式的辨認相比，其詞辨認率仍低了 5.88%，因為在沒有詞典及語言模型的協助下，為避免構詞產生之詞圖過大，在有限的音節圖大小無法展開所有可能的相近音，相鄰之時間點也容易產生出許多相同音節，在停頓不明顯的時間點上更容易出現許多有 BT0 及 BT1 的相同音節，這些因素都導致詞的涵蓋率無法隨著詞圖大小的增加而有效提高，進而降低辨認率。若能在構詞的同時結合語言模型做 beam search，將一些分數低之路徑砍掉，則可以由音節涵蓋率較高但是 arc 數較大的音節圖，進行構詞，如此一來其展開之詞圖就可以避免 arc 過多的問題。

5.2 未來展望

從本研究可延伸出下列幾個議題值得未來探討:第一，以一階段式的方式，

在 PD-AM 產生音節圖時，每當偵測出 BT3 即做局部性的辨認，縮小辨認的搜尋空間。第二，利用 PD-AM 解碼出之韻律停頓，引入如 pitch、energy 及 pause duration 等韻律參數，進一步提高詞的辨認率。第三，一字詞一直都是造成辨認率下降的主要原因，因此，我們可利用韻律停頓將詞綴併入形成韻律詞(prosodic word)，減少一字詞的錯誤。第四，目前本研究只對朗讀式語音作辨認，未來若能延展到自發性語音，相信可以將語音辨認更廣泛地應用在生活之中。



參考文獻

[1] K. F .Lee, “Context–Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition,” IEEE Trans. Speech Audio Process., vol38, no.4, Apr. 1990, pp. 599-609.

[2] I. Shafran, M. Ostendorf and R. Wright, “Prosody and phonetic variability: Lessons learned from acoustic model clustering”, in Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding, pp. 127-131, 2001.

[3] M. Ostendorf et al., “A prosodically labeled database of spontaneous speech,” Proc. of the ISCA Workshop on Prosody in Speech Recognition and Understanding, pp. 119-121, 2001.

[4] 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.

[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 Trans. on Audio, Speech and Language Processing, vol. 14 no.1, pp.232-245, January 2006.

[6] C. Ni, W. Liu, and B. Xu, “Using prosody to improve Mandarin automatic speech recognition,” in Proc. INTERSPEECH 2010, Makuhari, Japan, Sept. , pp 2690-2693.

[7] Jui-Ting Huang, Po-Sen Huang, Yoonsook Mo, Mark Hasegawa-Johnson, Jennifer Cole, “Prosody-Dependent Acoustic Modeling Using Variable-Parameter Hidden Markov Models,” in Proc. Speech Prosody 2010, Chicago, USA, Apr.

[8] Jyh-Her Yang, Ming-Chieh Liu, Hao-Hsiang Chang, Chen-Yu Chiang, Yih-Ru Wang, and Sin-Horng Chen, "Enriching mandarin speech recognition by incorporating a hierarchical prosody model," in Proc. ICASSP 2011, Prague, Czech, May, 2011, pp 5052-5055.

[9] 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.

[10] Mandarin microphone speech corpus – TCC300, http://www.aclclp.org.tw/use_mat.php#tcc300edu.

[11] Z. Sheng, J.-H. Tao, and D.-L. Jiang, "Chinese prosodic phrasing with extended features,"Proceedings of the IEEE ICASSP, Vol. 1, pp.492-495, 2002.

[12] 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, pp.284-309,2005.

[13] "HTK Web-Site", http://htk.eng.cam.ac.uk. Accessed 2009.

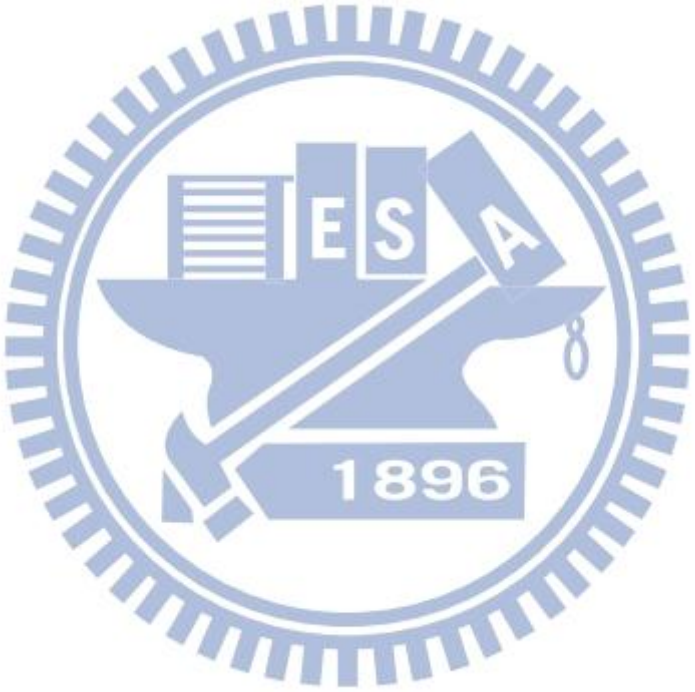
[14] F. Sha and F. Pereira, " Shallow Parsing with Conditional Random Fields",2003.

[15] 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。

[16] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri. OpenFst: A general and efficient weighted finite-state transducer library. In Proceedings of the 12th International Conference on Implementation and Application of Automata (CIAA 2007), Prague, Czech Republic, July 2007, volume 4783 of Lecture Notes in Computer Science, pages 11–23. Springer, Heidelberg, 2007.

[17] D. Moore, J. Dines, M. Magimai Doss, J. Vepa, O. Cheng, and T. Hain, “Juicer: A weighted finite state transducer speech decoder,” in Proc. MLMI (to appear), Washington DC, May 2006.

[18] 劉銘傑，“以韻律輔助之中文語音辨認系統之實現”，國立交通大學碩士論文，民國一百年七月。



附錄：決策樹之問題集

Prosodic break question

Bclass0	BT0
Bclass1	BT1
Bclass2	BT2
Bclass3	BT3,sil

Phonetic context question-聲母

國語聲母音素問題集 (依據發音方式及發音部位)	
清音	p, t, k, c, ch, q, b, d, g, z, zh, j, f, s, sh, x, h
濁音	r, m, n, l
送氣	p, t, k, c, ch, q
不送氣	b, d, g, z, zh, j
塞音	p, t, k, b, d, g
送氣塞音	p, t, k
不送氣塞音	b, d, g
擦音	f, s, sh, x, h, r
擦音清	f, s, sh, x, h
擦音濁	r
塞擦音	c, ch, q, z, zh, j
送氣塞擦音	c, ch, q
不送氣塞擦音	z, zh, j
鼻音	m, n

邊音	l
唇音	m, p, b, f
雙唇音	m, p, b
唇齒音	f
舌尖前音	z, c, s
舌尖中音	d, t, n, l
舌尖後音	zh, ch, sh, r
舌面音	j, q, x
舌根音	g, k, h
齦音	n, d, s, l
軟顎	g, ng, h
雙唇塞音	b, p
舌尖中塞音	d, t
舌根塞音	g, k
舌尖前塞擦音	z, c
舌尖後塞擦音	zh, ch
舌面塞擦音	j, q

Phonetic context question-韻母

國語韻母音素問題集 (依據舌面高低及舌位前後)	
舌面高	FNULL1, FNULL2, wu1, wu2, wu3, yi1, yi2, yi3, yu1, yu2
舌面半高	er, e, o
舌面半低	eh
舌面低	a
舌位前	a, eh, yi1, yi2, yi3, yu1, yu2

舌位中	FNULL1, FNULL2, er, e
舌位後	o, wu1, wu2, wu3
圓唇	yu1, yu2, wu1, wu2, wu3, o
圓唇舌面高	yu1, yu2, wu1, wu2, wu3
圓唇舌面半高	o
舌面高舌位前	yi1, yi2, yi3, yu1, yu2
韻尾鼻音	ng, en
閉央不圓唇母音	FNULL1, FMULL2