國 立 交 通 大 學

電信工程研究所

碩 士 論 文

使用韻律信息之中文自發性語音辨認

A Prosody-Assisted Mandarin Spontaneous Speech Recognition

研究生：黃仰駿

指導教授：陳信宏 博士

中 華 民 國 一 百 零 三 年 八 月

使用韻律信息之中文自發性語音辨認

A Prosody-Assisted Mandarin Spontaneous Speech Recognition

研 究 生：黃仰駿　　Student：Yang-Chun Huang

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

August 2014

Hsinchu, Taiwan, Republic of China

中 華 民 國 一 百 零 三 年 八 月

使用韻律信息之中文自發性語音辨認

研 究 生：黃仰駿　　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

近年來朗讀式語音辨識已有相當不錯的效能，但自發性語音辨認卻因為語速較快、語法不規則、語流不流暢等原因仍舊困難，本論文探討中文自發性語音辨認，研究重點在語言模型的建立及加入韻律信息的辨認過程。在語言模型建立上，考慮語者說話猶豫時所使用的感嘆詞及無意義的慣用插語，並利用語言模型調適來解決文字語料不足及文法語流特性和朗讀語音不同的問題，以建立一套自發性語言模型；在辨認過程上，使用兩階段辨認來加入韻律信息協助辨認，首先在第一階段辨認使用傳統聲學模型及 bigram 語言模型產生一個 word lattice，接著在第二階段辨認先擴展語言模型為 factored 語言模型，再加入韻律邊界停頓資訊與音節韻律狀態資訊，經過重新評分後得到一條最佳路徑，並同時解碼出相關資訊。使用中研院 MCDC 語料作實驗，獲得詞、字及音節的辨識率分別為 58.29%、64.94%及 68.89%，較傳統只使用第一階段辨認的作法絕對辨認率改善了 4.43%、4.6%及 3.06%。經辨認結果分析發現，對於正常語流而言，加入韻律信息能夠改善搶詞及聲調辨認錯誤；但對於不正常語流來說，改善的效能非常有限。

A Prosody-Assisted Mandarin Spontaneous-Speech Recognition

Student：Yang-Chun Huang Advisor：Dr. Sin-Horng Chen

Institute of Communication Engineering
National Chiao Tung University

Abstract

In recent years, the Mandarin read-speech recognition technology is quite mature. However, it is still difficult for spontaneous speech recognition due to high speaking rate and the existence of disfluent speech events. This thesis discusses Mandarin spontaneous speech recognition, focusing on language model establishment and the process of prosody-assisted recognition. In the language model establishment, two particular words of particle and marker are added to the vocabulary to model the disfluency phenomena of spontaneous speech. Besides, language model adaptation is employed to solve the problem of the insufficiency of texts of spontaneous speech. In recognition, a two-stage recognition process to incorporate prosodic information is adopted. In the first stage, an acoustic model and a bigram language model is used to generate a word lattice. Then, in the second stage the word lattice is firstly extended to replace the bigram LM with a factorized LM. Then, break-related models and prosodic state-related models of a hierarchical prosodic model are sequentially added to rescore all searching paths in order to find the best recognized word sequence. Experimental results on the Academia Sinica MCDC corpus showed that word, character and base-syllable accuracy rates of 58.29%, 64.94% and 68.89% were achieved. They were better than the results of the baseline system by 4.43%, 4.6% and 3.06%, respectively. By error analysis we find that prosodic information is useful in resolving word segmentation ambiguity and tone pattern confusion for fluent speech part, while it is less effective for disfluent part.

致謝

終於要畢業了，在這兩年的日子裡，吃了不知道幾餐的ㄅㄨㄣ，咪了不知道幾次。首先感謝陳信宏老師，在百忙之中還是經常關心我的研究進度；感謝王逸如老師，給了我許多研究上的方法與建議，也許我不是個適合做研究的研究生，但是從中還是學到了許多；也感謝許多學長的幫助，讓我在研究上能夠順利許多。

感謝 707 實驗室裡的每個人，很會推導公式的實驗室一哥王柏(?)、關於小熊的茂隆(沒吧大哥)、提供 FHM 配咖啡的仲毛、從大學就一起打拼的阿璋、和我一起做辨識的小鋒、美麗翹秘書 Jean，還有每位有潛力的學弟妹，因為有你們每個人，我的實驗室生活變得更多采多姿。謝謝交大電研所有給力的隊友，也要謝謝每個給予我鼓勵的朋友們，很開心有你們大家。

最後我要感謝我的父母和女朋友，多少次掙扎著想要放棄，如果沒有你們的支持，我想不會有今天的成果。以此論文獻給所有幫助過我的人，謝謝你們！

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究動機

當語音辨認(Automatic Speech Recognition, ASR)系統已有相當程度的技術，對於朗讀式語音辨認效能相當好時，人們開始研究更接近日常生活的自發性語音辨認，而在外國的研究方面，自發性語音也有一定的辨識率，但中文自發性語音辨識系統卻仍然沒有很好的效能，辨認率與朗讀式(Read Speech)有一段差距，造成這樣的原因主要是自發性語音說話速度(speaking rate)較快、語者在說話時未經大腦良好的規劃，使得語音型態發生改變、說話聲音大小不定、語流不流暢、語料不足等。

上述的原因可能造成自發性語音發生一些特殊的語音現象，像是語速較快時，語者為了要節省發音時所需的力氣，造成發音變異(pronunciation variation)或是音節合併(syllable contraction)現象，對於人腦也許可以辨識出來，但是對於機器辨識上卻是一大挑戰；另外，因為說話時未經大腦良好規劃，語者常出現遲疑(hesitation)、詞語修補(repair)、重複(repetition)…等現象，都會造成語流不順暢(disfluency)；而在文法結構上，常會產生不具意義的感嘆詞(particle)、語者慣用的語助詞(marker)，以上這些特殊現象都會造成辨認上的困難，如果我們能有效解決這些自發性語音特殊現象對語音辨認造成的問題，相信自發性語音辨識效能會有大幅的提升。

在辨認系統上，聲學模型(Acoustic Model, AM)因為發音變異和音節合併這些現象必須做修正；語言模型(Language Model, LM)方面，因為難有大量的自發性文字語料，所以必須利用調適(adaptation)來得到適合的語言模型；而韻律模型方面，也必須考慮特殊單元和語流不流暢而需重新設計。傳統上的語音辨識系統是由聲學模型加入語言模型做辨識，本研究希望藉由韻律模型的加入來提升辨識率。

1.2 研究方向

在本論文中，希望先建立一套傳統的自發性語音辨識系統，並試著利用韻律模型的協助來提升辨識效能，辨認系統架構如圖 1.1 所示。首先，聲學模型是基於隱藏式馬可夫模型(Hidden Markov Model, HMM)訓練出 Context-Dependent 的三連音素模型(tri-phone model)；在語言模型方面，先利用大量文字資料訓練出朗讀式語音之語言模型，再利用語言模型 MAP 調適法中的 Model Interpolation 來將背景語言模型調適成合乎自發性語音之語言模型；另外，韻律模型是針對自發性語音特性進行設計。辨識過程是採用兩個階段辨識，在第一階段首先利用聲學模型加入語言模型進行辨識來產生一個 word lattice，第二階段辨識則是在 word lattice 上加入韻律模型分數，再將此三種模型機率分數做權重結合重新計算分數，以決定出最有可能的辨識結果。

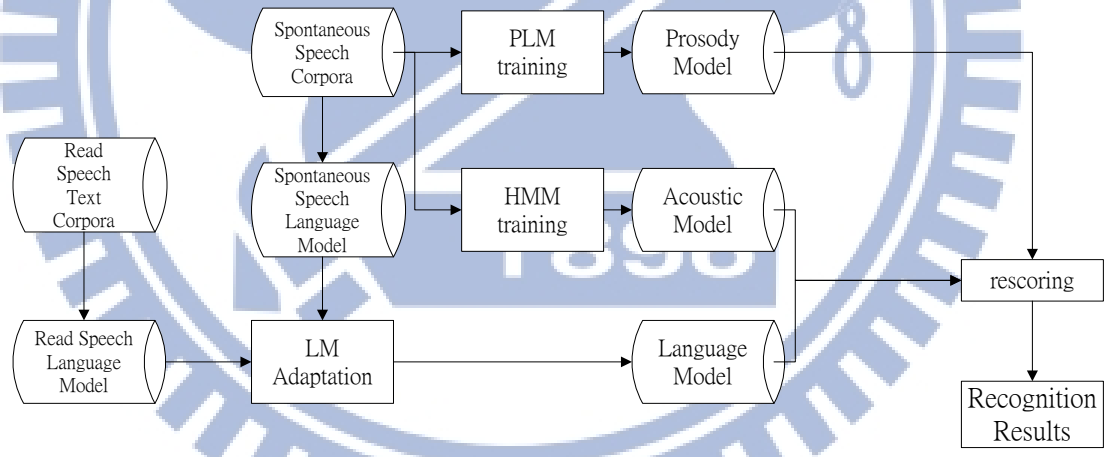


圖 1.1：訓練與辨識系統架構圖

1.3 文獻回顧

在自發性語音中，因為常發生許多特殊現象，使得聲學模型難以有效建立，像是發音變異導致聲學模型混淆而辨識效能下降，這個問題可以從辨認字典(lexicon)中加入可能之發音變異【1】，或是改進底層之聲學模型來解決，文獻【2】中利用決策樹(decision tree)的方式來決定額外訓練的發音變異聲學模型，接著使用 state tying 以及 mixture tying 來得到較好的聲學模型。而自發性語音中語者因為重複或是遲疑現象造成不流暢語流，【3】提出偵測重複現象的相關研究，考慮單一重複和多重重複，並將這些重複單元加入辨識字典中，對於偵測重複現象有不錯的效果；Zgank 等人【4】則將 filled pause 加入聲學模型中，並且利用 Interpolation 調適朗讀式語料與自發性語料來訓練語言模型，自發性語音文字語料量少一直是語言模型難以有效建立的原因，口語對話因為回答對方或是猶豫時常有無意義的 particle 的產生，【5】考慮 filler prediction model，預估這些 particle 產生的位置；Ng and Ostendorf【6】則是利用從網路上收集文字資料，再與朗讀式語言模型作調適結合；【7】研究中也利用 MAP 及 Class-based Deleted Interpolation Smoothing 方法調適出適合自發性語音的語言模型。

在韻律模型幫助辨認之研究方面，【8】用 event-based 的方式增加語音辨認效能，利用韻律參數建立一個偵測事件的模型，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(word lattice)重新計算分數；【9】則是利用韻律和語言結構之間的關係建立一套韻律相關之語言模型(prosody-dependent language model)，並且利用韻律邊界資訊來建立韻律相關之聲學模型(prosody-dependent acoustic model)【10】。

1.4 章節概要

本論文共分為六章，各章節內容分配如下：

第一章　緒論：說明研究動機及研究方向。

第二章　漢語口語對話語料庫介紹：介紹本研究實驗所使用之自發性語音語料庫、自發性語音之特性及現象分析、初步統計以及後處理。

第三章　辨識模型介紹：說明本研究所使用之聲學模型、語言模型、韻律模型之建立及調適。

第四章　語音辨識系統架構：說明加入韻律訊息於 two-stage 語音辨認系統之架構及實作、鑑別式模型組合。

第五章　實驗結果與分析：韻律訊息加入辨認後之實驗結果及分析。

第六章　結論與未來展望。

第二章 漢語口語對話語料庫介紹

2.1 語料庫介紹

本研究使用中央研究院語言學研究所提供一個完整的自發性語音語料庫－現代漢語口語對話語料庫(Mandarin Conversational Dialogue Corpus，MCDC)作為研究素材。

現代漢語口語對話語料庫是由中央研究院語言學研究所曾淑娟博士等人 2000~2002 年間所錄製，語者是由台北市民隨機抽樣，並依據 16~25 歲、26~35 歲以及 36~45 歲三大年齡層，選取 60 位語者(37 位女性、23 位男性)，共錄製 30 段對話，但其中有轉寫的僅有 8 段對話，因此拿 8 段對話當作語料，其中有 16 位語者(9 位女性、7 位男性)兩兩互相交談，總長度 496 分鐘。

2.1.1 音檔格式說明

MCDC 音檔錄製使用 Audio Technica ATM 33a 手持式麥克風，以取樣頻率 44.1kHz 將兩位語者的聲音分錄於左右聲道，再利用軟體 Cool Edit Pro，將它們分割成小的雙聲道音檔，依長度約三分鐘找到一個清楚可辨的停頓切開，其簡介如表 2.1 所示。在本研究中將每組對話語料之左右聲道抽取，並轉換為兩個單聲道之音檔，分別為對話中兩位語者之語料，並且將其取樣頻率下降至 16kHz，再利用每一段落相對應之開始及結束時間作切割，經由以上處理後產生 7,085 個音段(turn)，扣除一些只有非正常語音以及一些聲音過小之音檔後剩下 5,900 個音段(turn)將作為本研究所使用之語料。

表 2.1：MCDC 語料庫對話主題與語者對照表

對話序號	長度 (分鐘)	發音人	聲道 (L/R)	語者編號	對話主題
mcdc-01	61	MISC-08-male-25	R	01R	工作、休閒活動、經濟、開車
		MISC-07-female-29	L	01L	
mcdc-02	63	MISC-10-male-35	R	02R	休閒活動、經濟、工作、性別、政治
		MISC-09-female-37	L	02L	
mcdc-03	61	MISC-12-female-17	R	03R	家庭、學校、購物、生涯規劃、明星
		MISC-11-female-16	L	03L	
mcdc-05	63	MISC-15-male-40	L	05L	工作、家庭、社會階級、保險、歷史、省籍情結、名人
		MISC-16-female-46	R	05R	
mcdc-09	66	MISC-23-female-30	R	09R	工作、旅行、生活態度、環保、健康
		MISC-24-female-35	L	09L	
mcdc-10	54	MISC-26-male-23	R	10R	電影、政治、軍隊、捷運、學校、經濟
		MISC-25-male-35	L	10L	
mcdc-25	55	MISC-57-male-43	L	25L	交通、工作、小孩、旅行、電腦、管理
		MISC-58-female-45	R	25R	
mcdc-26	46	MISC-60-male-24	R	26R	工作、求職、家庭、車禍、休閒活動、學英文、婚姻、軍隊
		MISC-59-female-37	L	26L	

2.1.2 語料標記格式說明

本研究的 MCDC 文字語料標記是使用中央研究院語言學研究所釋出之版本，所採用的標注格式是一種標籤式的語言格式，有點類似於 XML 語法，標記段落上大致以對話中語者轉換處為一個段落 (Sub-turn) 作轉寫，轉寫內容包括：對應之音檔名稱、語者代號、段落的開始及結束時間、語音之文字轉寫以及其發音相對之漢語拼音，文字以及漢語拼音的轉寫包括語言及非語言部分，非語言部分主要是標記非人類產生聲音以及人類所產生但不是語音的聲音，例如：咳嗽聲、笑聲、呼吸聲等。如圖 2.1 及表 2.2 所示，為一個段落之文字轉寫範例及說明。

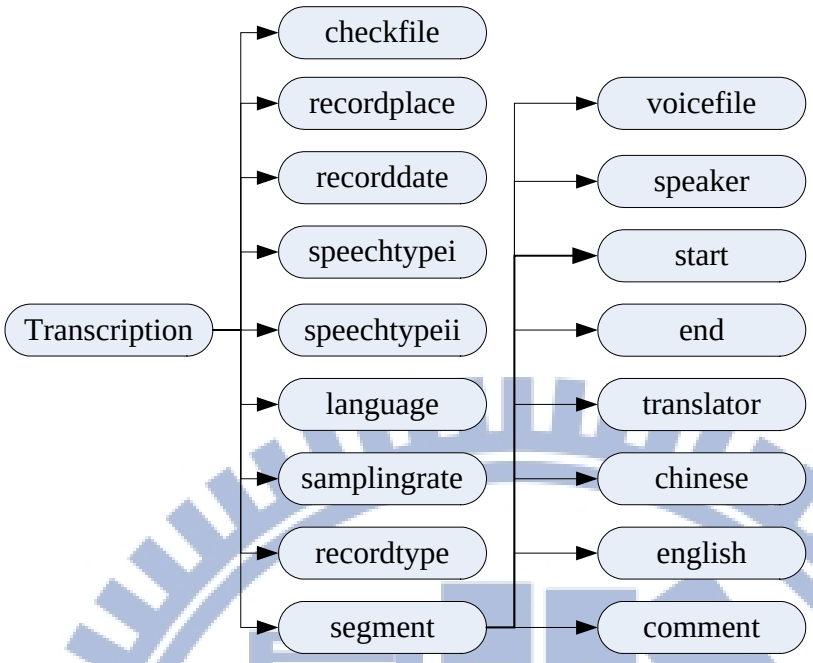


圖 2.1：標籤式的語言格式

表 2.2：文字轉寫範例

(段落開始)
<voicefile>D:\MCDC\stereo_01\mcdc-01-01.wav (音檔名稱)
<speaker>MISC-07-female-29 (發音人)
<start>020976 (音檔開始時間)
<end>025360 (音檔結束時間)
<translator>Fen (文字轉寫者)
<chinese> (中文轉寫內容)
O 我在一家公關公司上班 (unrecognizable non-speech sound)
</chinese>
<english> (漢語拼音轉寫內容)
O wo3 zai4 yi4 jia1 gong1 guan1 gong1 si1 shang4 ban1 (unrecognizable non-speech sound)
</english>
<comment> (註解標示)
</comment>
(段落結尾)

2.2 自發性語音之特性

自發性語音與朗讀式語音最大的不同在於自發性語音不是經過事先設計好的，所以說話時常常會伴隨著因大腦思考或情緒變化而產生一些無法預期的聲音或發生在語言學中較詞層次(word level)更為上層之行為，因而導致發音的完整性、文法結構、語速快慢，都與朗讀式語音有很大的差異。基於上述各種情況，致使自發性語音產生許多特有的現象，以下我們將針對 MCDC 語料庫中常出現的特性作介紹。

2.2.1 感嘆詞 (particle)

不具標準語意的感嘆詞，其語用成份居多如回應或同意。語流中出現的感嘆詞可以分成下列四類：

(1). 有相對應國字的感嘆詞，例如：A(啊)、AI(哎/唉)、BA(吧)、LA(啦)、MA(嗎、嘛)、O(喔/噢/哦)、E(呃)、NO(喏)、WA(哇)、YA(呀)、YOU(呦)

(2). 無相對應國字的感嘆詞，例如：EI、HEN

(3). 源於台語的感嘆詞，例如：HAN、HEIN、HO

(4). 其他的感嘆詞(Fillers)，例如：MHM、MHMHM

2.2.2 語助詞 (marker)

說話者本身在語流中慣用的插入語，這些習慣插語有其基本詞彙意義。但在語流中習慣插語已不保有其原有的完整語意，而較具語用功能。例如，作用於口語中說話者意欲保有其說話權且又需緩衝時間去思索組織其想說的句子，此時習慣插語 NA 便常被使用。語流中常出現的語助詞包括有：NE(那)、NA(那)、ZHE(這)、GE(個)、SHEN(什)、ME(麼)。

2.2.3 無法或難以辨識的語音

無法或難以辨識的語音主要可分為：

(1). 無法辨識的語音(unrecognizable speech sound)：標記員確定此為人類所發出之語音但無法辨認是何字何意何音

(2). 不確定字/音(uncertain)：

i. 可猜測出大概的語音內容，但無法百分之百確定

ii. 無法根據語意猜測出對應字詞，但可清楚記錄出其發音

2.2.4 非語音聲 (Non-Speech sounds)

在口語對話語料庫中常常會有一些非語音的聲音出現，非語言部分可分為人類所產生之副語言現象(para-linguistic)或非語言現象(non-linguistic)。

一般的非語音聲確定是由人所發出來的即稱為副語言現象，例如：笑聲、咳嗽聲、吞口水聲等等；而非語音且確定不是由人所發出來的則稱為非語言現象，例如：背景的雨聲、敲擊到麥克風聲等等。

2.2.5 語流中斷

在本研究中關注之語流中斷主要有沉默(silence)、停頓(pause)或短停頓(short break)，為語者在語流中因話題銜接不上或自身所產生之沉默。

2.2.6 非流暢現象 (disfluency)

不流暢的語音為自發性語音中一個重要特性，在本研究中關注之詞語修補主要有重覆(repetition)、詞語更正(repair)、部分重覆(restart)以及更正插語(editing term)，重覆是指完整地重覆詞語一次以上；詞語更正為說話者覺得說出的話不適當，立即更正說話內容；而部分重覆則是說話者重新說出這個句子且重覆詞語的片段，與完整的詞語重覆不同。更正插語是出現在被更正詞語與更正詞語之間，或是出現在完整重覆或部分重覆中，兩個重覆詞語之間。本研究定義詞語修補中斷點(IP)為被更正詞語與更正後詞語間之停頓點，或完整重覆或部分重覆中的兩個重覆詞語間之停頓點，本研究在文字轉寫中將詞語修補中斷點標記成「*」。以下為幾種非流暢現象範例：

- 重覆範例： 昨天卡卡表現的(普通)*(普通)
- 詞語更正範例： 今晚世足賽是(烏拉圭)*[EN](巴拉圭)對日本
- 部份重覆範例： (今)*(今天)晚上是冠軍賽

2.3 自發性語音現象分析

2.3.1 Particle 和 Marker 於句中位置分析

marker

- 大部分出現在句首或句中

	GE	ME	NA	NE	SHEN	ZHE
國字	個	麼	那	那	什	這
注音	ㄍㄜ	ㄇㄛ	ㄋㄚ	ㄋㄜ	ㄕㄜ	ㄓㄜ
Tone	5	5	4	4	2	4

EX: [那] 賴 先 生 呢 ？

我 就 說 [那] 你 們 真 的 會 去 嗎 ？

不 是 [那 個 什 麼] 大 稻 埕 啊 ！

particle

- 只出現在句尾

	BA		LA		MA		NA	NE	YA
國字	吧	吧	啦	啦	嗎	嘛	哪	呢	呀
注音	ㄅㄚ	ㄅㄚ	ㄌㄚ	ㄌㄚ	ㄇㄚ	ㄇㄚ	ㄋㄚ	ㄋㄜ	ㄧㄚ
Tone	5	1	5	1	1	5	1	1	5

EX: 應 該 這 種 比 較 能 夠 讓 大 家 接 受 [吧] ！

- 只單獨出現

	E	EI	HAN	HEIN	HEN	YA	AI
國字	呃	無相對應國字	源於台語	源於台語	無相對應國字	ya	唉
注音	ㄜ	ㄟ	ㄏㄚ	ㄏㄟ	ㄏㄣ	ㄧㄚ	ㄞ
Tone	5	2	2	3	3	4	1

EX: EI！你的輪胎跟原廠不一樣。

➢ 出現在句尾 or 單獨出現

	A	EI	HO		NO	O		WA	YOU
國字	啊	無相對應國字	源於台語	源於台語	喏	哦	喔	哇	喲
注音	ㄚ	ㄟ	ㄏㄡ	ㄏㄡ	ㄋㄨㄛ	ㄛ	ㄛ	ㄨㄚ	ㄧㄡ
Tone	1	1	4	3	4	2	1	5	1

EX: 句尾：我們每個司機都會知道[哇]！

單獨：那我一看，[哇]！好吧！

Filler

填補詞 (Fillers)，本身不具任何語意，也不具句法功能，通常是說話者正在思考接下來所要講的句子時，用來填補對話空白，也可以使說話者保留說話權。

	MHM	MHMHM
國字	Fillers	Fillers
注音	ㄣ	ㄣㄣ
Tone		

2.3.2 自發性語音特有現象

特殊單元

難以辨認的語音

uncertain	6333	有標記 tone 的資訊，但 syllable type 有些不在 411 base-syllables 中
unrecognizable speech sound	226	

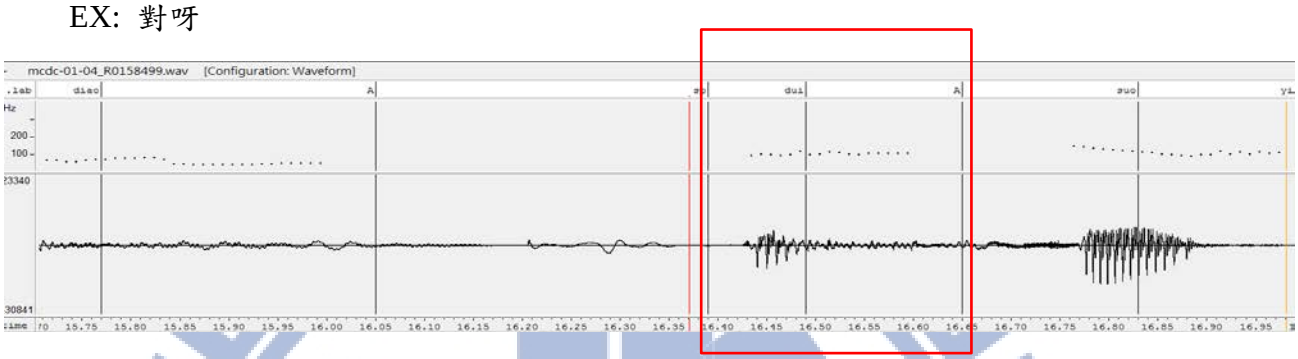
Code-Switch

English	135
Japan	15
MinNan	283

特殊現象

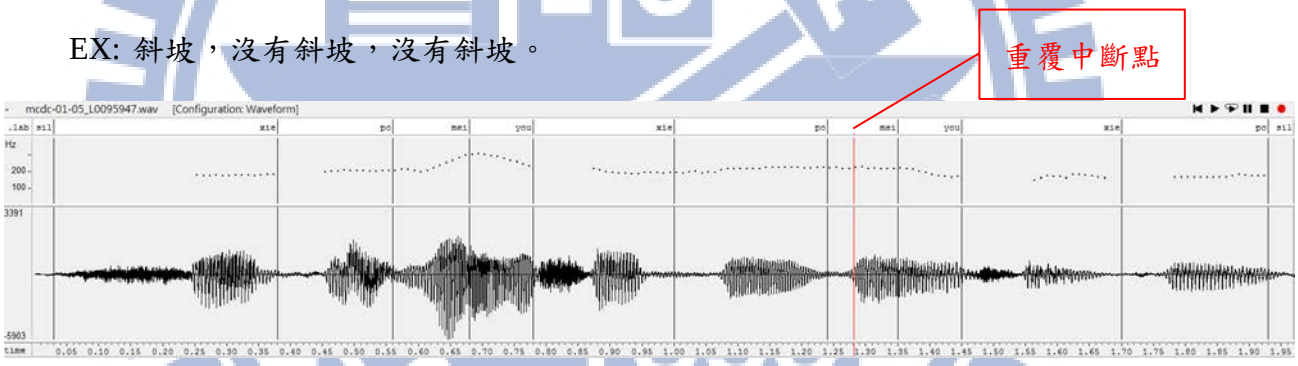
syllable contraction

EX: 對呀

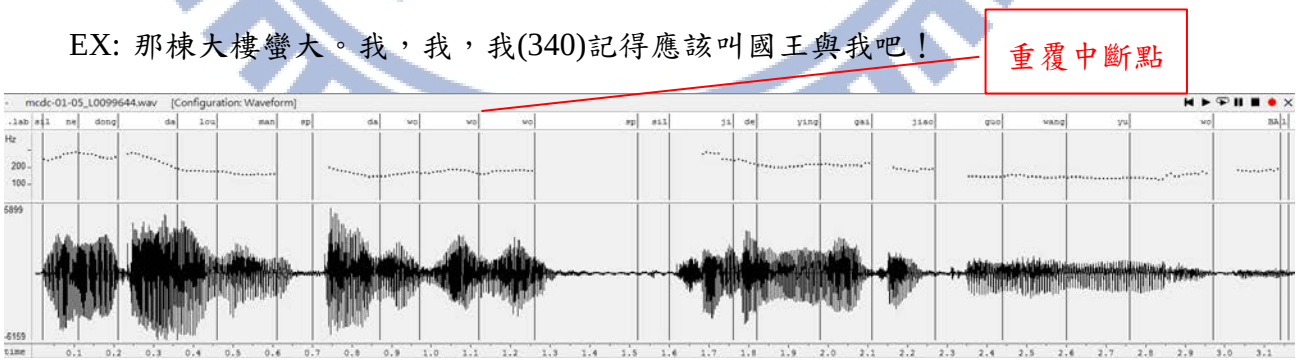


repeat

EX: 斜坡，沒有斜坡，沒有斜坡。

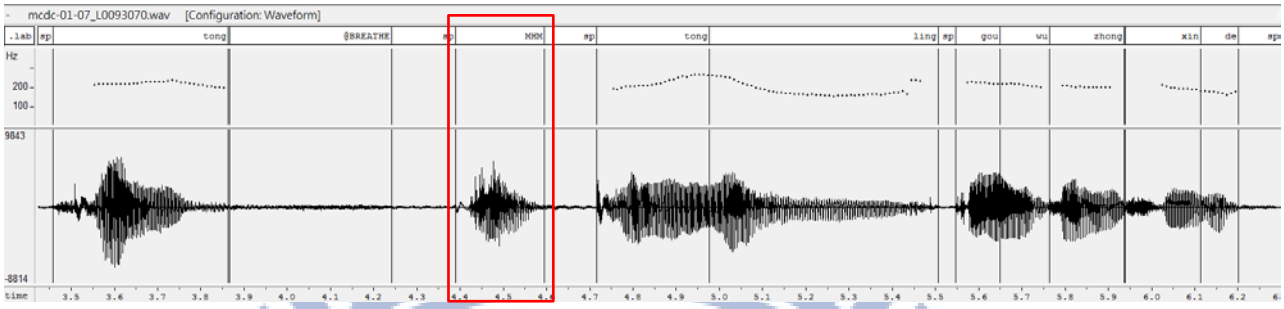


EX: 那棟大樓蠻大。我，我，我(340)記得應該叫國王與我吧！



更正插語

EX: 忠孝東路那個統， {@呼吸聲}MHM！統領購物中心的背後。



2.4 MCDC 語料庫初步統計

音檔處理後將每段對話切割成若干 Sub-turn，統計語料庫中所有 Sub-turn 之音節數分佈如圖 2.2 所示。其中音節個數大於 100 的 Sub-turn 有 325 個。

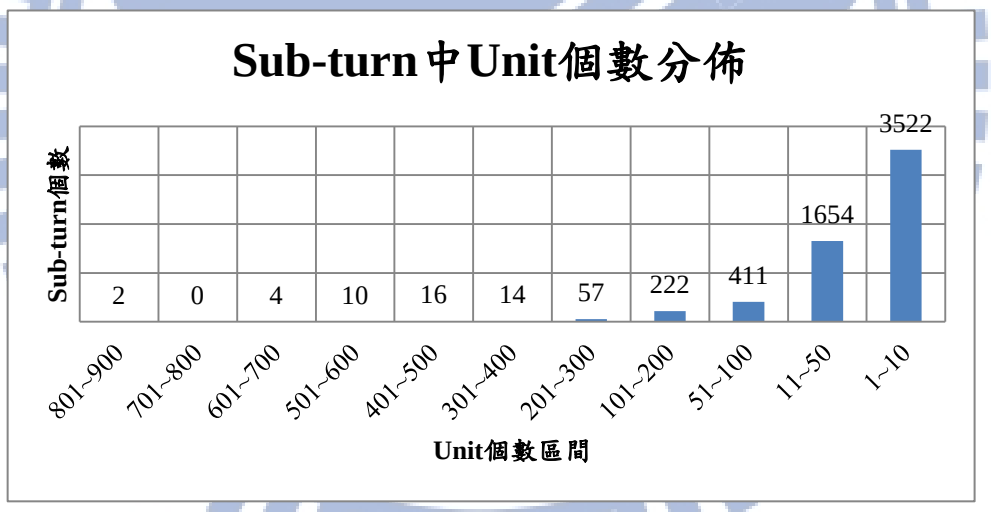


圖 2.2：語料庫中每個 Sub-turn 之音節數分佈圖

2.5 語料庫之後處理

由於中央研究院語言學研究所釋出之 release 版本中並無標記某些聲學現象及語言資訊，但這些資訊在語音處理中相當重要，因此在本節介紹本研究將對語料庫聲學以及語言學資訊的處理方法。另外，針對文字語料庫標注錯誤的部分進行更正。

2.5.1 斷詞與相關統計

由於中研院提供的文字轉寫並沒有斷詞和詞性(Part of Speech, POS)的相關標記，因此本研究先利用國立交通大學語音處理實驗室之斷詞器對 MCDC 之文字轉寫進行斷詞及標詞性，詞性標記以中研院的 46 類詞性為標準(參見附錄一)。值得注意的是，MCDC 的文字轉寫中包含了許多音節間(如：笑聲、呼吸聲)或是特殊現象的標記，為了不受這些標記的影響，斷詞時會先移除這些標記等斷完後再加回，另外對於感嘆詞和語助詞的部分，會先將其轉成相對應國字再放入斷詞器，以避免其被視為外來語(Foreign word, FW)。

2.5.2 切除音檔頭尾過長的靜音

中研院提供的語料庫，雖然是使用雙聲道的方式將交談的兩語者的聲音用左右聲道分開錄製，但交談的過程中，麥克風仍舊會收錄到另一位語者的聲音。如果另一位語者的聲音與原聲道的語者出現重疊的話，會進一步形成 cross talk 的現象。此外在中研院提供的 MCDC 文字轉寫中有標註每一語段(Sub-Turn)的開始與結束時間，但某些語段所標註的開始與結束時間預留了過長的靜音(silence)，並且這些在語段頭尾預留的靜音往往不屬於純靜音，而含有從另一位語者收錄過來的聲音。由於這些聲音在文字轉寫上是標為靜音，但實際辨認時仍舊能夠辨認出另一位語者的說話內容，進而導致嚴重的插入型錯誤(insertion error)。由於這些現象並不是我們研究的主要內容，因此針對這些頭尾過長的靜音，我們利用切割位置找出頭尾長度大於 500 毫秒且能量大於平均值的靜音予以切除。

2.5.3 修正音節切割位置

要訓練出好的聲學模型及韻律模型首先要有良好的音節切割位置，但目前中研院提供的 MCDC 語料並不包括音節切割資訊。為了得到較準確的切割資訊，本研究首先使用【11】建立之自發性語音聲學模型，利用 HTK【12】以強迫對齊的方式對每一個音

檔作音節之切割，得到每一個音節的音節長度以及停頓時長，接著檢查一些較明顯音節切割錯誤的位置，以人工的方式進行修正。較明顯音節切割錯誤的情況如下：

1. 詞內邊界(intra-word)中出現超過 0.1 秒的停頓。
2. 容易出現極端音節長度的切割位置，如：音節合併現象、拖長音現象。
3. 停頓時長區間內的能量高於平均值的停頓
4. 發生不流暢現象的音節切割位置

第三章 辨識模型介紹

第三章介紹語音辨認所使用的模型，3.1 節為聲學模型介紹，本研究使用【14】訓練之自發性語音聲學模型，它是基於隱藏式馬可夫模型(Hidden Markov Model, HMM)，而所使用的工具為英國劍橋大學開發的 HTK (HMM Tool Kit) 【12】軟體；3.2 節為語言模型介紹，首先會利用 SRILM toolkit【13】訓練出朗讀式語言模型，再利用調適方法訓練出適合自發性語音的語言模型；3.3 節介紹本研究所使用【14】所設計的韻律模型，考慮自發性語音的特殊現象，希望藉由此模型對自發性語音辨認有所幫助。

3.1 聲學模型

本研究是使用【14】所訓練的聲學模型，其採用多語者辨識方式，也就是訓練和測試語料皆有相同的語者，但是句子不同，由 MCDC 內 16 名語者中，選取每位語者約 9/10 的語句做訓練語料，如表 3.1，其餘的 1/10 做為測試語料如表 3.2。

表 3.1：訓練語料統計

	正常語音	難以辨認的語音	語助詞和感嘆詞	Code-Switch	非語音
字數	105,800	6,857	10,198	494	5,053
比例	82.40%	5.34%	7.94%	0.38%	3.94%
總字數	128,402 (89.99%)				
音檔數	5,750				
時間長度	7.967 (hours)				
語者數	16				

表 3.2：測試語料統計

	正常語音	難以辨認的語音	語助詞和感嘆詞	Code-Switch	非語音
字數	12,048	750	756	72	652
比例	84.38%	5.25%	5.29%	0.50%	4.57%
總字數	14,278 (10.01%)				
音檔數	150				
時間長度	59.79 (minutes)				
語者數	16				

表 3.3 為訓練時相關參數抽取的基本設定,所使用的聲學模型為音節內 context-dependent 的三連音素模型 (tri-phone model)，表 3.4 為各個 HMM 模型之設定

表 3.3：參數抽取基本設定

Sampling rate	16*kHz*
Frame size	32*ms*
Frame shift	10*ms*
No. Filter Bank	24

表 3.4：HMM 模型之設定

模型種類	模型個數	State 數目
triphone	517	3
sil	1	3
sp	1	1
filler	1	5
Eng	1	5
Particle	24	5
Paralinguistic	6	5
@SPEECH_SOUND	1	5

3.2 語言模型

由於所有語言都有其獨特的文法規則，因此我們可針對此規則性來求得一個機率模型，一般稱此為語言模型(Language Model, LM)。在語音辨認時，除了聲學模型外，若能加入語言模型的參考，通常能大幅提升辨識系統的效能。本節中首先介紹朗讀式語言模型的建立，接著介紹本研究所使用的語言模型調適方法，最後介紹語言模型之效能分析及涵蓋率與辨識率比較。

3.2.1 朗讀式語言模型之建立

文字語料庫簡介

用於訓練朗讀式語言模型的文字資料庫共有以下來源：

1.) 光華雜誌(Sinorama)：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：為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)：它是一套由中研院收集，內容包含多種主題，以語言分析研究為目的的資料庫。

4.) Chinese Gigaword：由 Linguistic Data Consortium (LDC)整合發行，內容包含台灣中央社、北京新華社等國際新聞。

5.) 維基百科語料(Wiki) ：維基百科為領域廣泛且資訊較為新，可以使語言模型更加多元，資料庫增加。

- 總詞數為 4.4 億，文章篇數為 8771 篇。

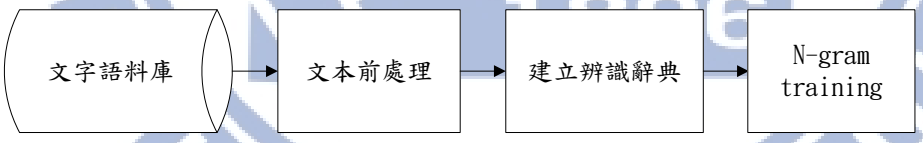


圖 3.1：語言模型訓練流程

上圖 3.1 為語言模型訓練流程，我們有了文字語料庫後，需先對語料庫的文章進行前處理，其中大致上分為 CRF 斷詞、文字正規化...等等，將文章中會影響辨識效能的內容移除或修改，在文本前處理後，再經由選詞將常出現、較重要的詞收錄在辭典內以便訓練出語言模型，本研究是利用蘇仲銘【15】論文中王逸如博士所提出的選詞方法，依據詞頻與出現文章篇數評估一個詞彙在文章中分布的均勻程度，共選入了 60,000 個詞彙，而六萬詞詞典在 MCDC 文字語料 99179 個詞中只出現 5177 個詞，可以看出口語對

話中的使用的詞不多，且都不艱深。表 3.5 及表 3.6 為感嘆詞及口語常用詞在朗讀式語料和 MCDC 語料中所佔的比例，由此可見兩語料用詞上的差異。

表 3.5：感嘆詞在兩語料庫中所佔比例

	朗讀式語料	MCDC 語料
啊	0.0029%	1.98%
喔	0.0007%	1.26%
啦	0.0016%	0.37%
嘛	0.0009%	0.28%
吧	0.0042%	0.21%

表 3.6：常用詞在兩語料庫中所佔比例

	朗讀式語料	MCDC 語料
然後	0.01%	0.72%
因為	0.074%	0.65%
所以	0.034%	0.47%
可是	0.008%	0.45%
其實	0.013%	0.42%

本研究訓練語言模型使用的軟體為 SRILM【13】，並利用退化平滑法(back off)及使用 Good-Turing discounting 建立而成一個 tri-gram 語言模型，其數學式如(3.1)式所示。若定義訓練語料中詞串出現的次數門檻k，則可將詞串分為出現次數高於門檻值、出現次數低於門檻值及從未出現三種。則參數可表示為下式：

$$P(w_i|w_{i-n+2},\dots,w_{i-1})$$

$$= \begin{cases} \alpha(w_{i-n+1}\cdots w_{i-1})P(w_i|w_{i-n+2}\cdots w_{i-1}) & ,Count(w_{i-n+1},\dots,w_i)=0 \\ d_\alpha \cdot \frac{Count(w_{i-n+1},\dots,w_i)}{Count(w_{i-n+1},\dots,w_{i-1})} & ,1 \le Count(w_{i-n+1},\dots,w_i) \le k \\ \frac{Count(w_{i-n+1},\dots,w_i)}{Count(w_{i-n+1},\dots,w_{i-1})} & ,Count(w_{i-n+1},\dots,w_i) > k \end{cases} \quad (3.1)$$

訓練出來的朗讀式語言模型其 perplexity 為 66.501，OOV rate 為 2.87%，對 MCDC 測試語料 perplexity 為 733.719，OOV rate 為 2.91%。

以下本研究將利用此朗讀式語言模型與自發性語言模型的進行調適。

3.2.2 語言模型之調適

在進行語言模型調適之前，必須先訓練出自發性語言模型，其中我們把 MCDC 中的感嘆詞替換回朗讀式語料也有的感嘆詞，例如：LA→啦、BA→吧，並根據 2.2 章自發性語音的特性，加入 particle 和 marker 兩詞於詞典中：

- particle：語者說話猶豫或回應對方時所發出的，其在語法上屬於較獨立的。
  (如：MHM、E、EI)
- marker：語者的習慣插語，已不保有其原來的語意
  (如：NA-GE、ZHE-GE、SHEN-ME)

語者在說話時不像朗讀式語音那麼順暢，常會有這兩類詞插入句子中，下圖 3.2 為語者詞語修補(repair)示意圖：

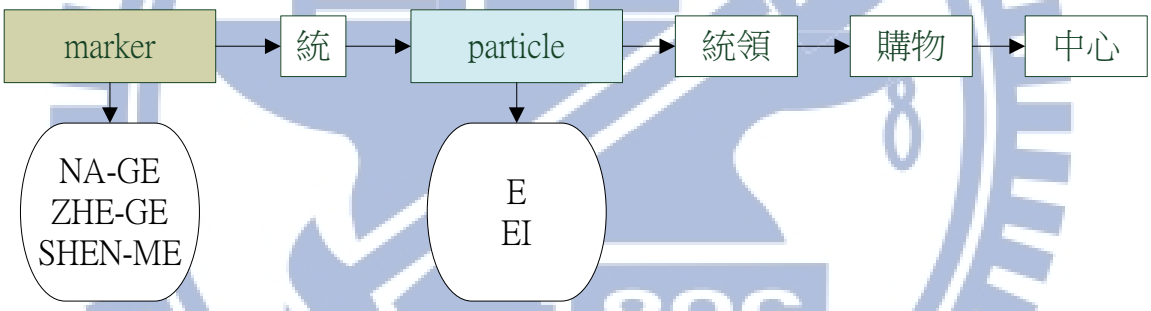


圖 3.2：詞語修補示意圖

在語言模型調適中，MAP(maximum a posteriori)調適法【16】是最普遍的方法，其主要概念是改變已知的模型參數分布使其與觀察資料的參數分布相符合，其數學式(3.2)如下：

$$X_{MAP} = \arg\max_x P(X|W) = \arg\max_x P(W|X)P(X) \qquad (3.2)$$

X:已知的參數模型分布(model parameter distribution)

W:有限的觀察資料(observation)

$P(W|X)$:觀察資料W的概似度(likelihood)

根據 MAP 調適法又可分為 count merging 和 model interpolation 兩種做法，本研究是使用 model interpolation 進行調適，其是將 out-of-domain model 與 in-domain model 進行內插，來得到我們要的語言模型，其數學式(3.3)如下：

$$P'(w_i|w_{i-n+1},\dots w_{i-1}) = \lambda P_{base}(w_i|w_{i-n+1},\dots,w_{i-1}) + (1-\lambda)P_{adapt}(w_i|w_{i-n+1},\dots,w_{i-1}) \quad (3.3)$$

P_{base}: probability for word sequence w_i of the baseline language model

P_{adapt}: probability for word sequence w_i of the adaptation language model

λ: linear interpolation coefficient, $0 \le \lambda \le 1$

權重λ也是根據 SRILM 上的估計方法，使用 EM algorithm，最大化調適模型與測試語料的概似度(likelihood)，經過迭代(iterative)7 次得到$\lambda = 0.690221$。

3.2.3 語言模型效能分析

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：

$$H = -\frac{1}{m}\log P(W = w_1, w_2, \dots, w_m) \quad (3.4)$$

上式為一個詞串$W = w_1, w_2, \dots, w_m$，對於每個新詞提供的平均資訊量(entropy)，經過適當的化簡而得。而混淆度可直接使用(3.4)式進一步定義為：

$$PPL = \exp(H) \quad (3.5)$$

若$P(W = w_1, w_2, \dots, w_m) = \prod_{i=1}^{m} P(w_i|w_1, w_2, \dots, w_{i-1})$則可發現混淆度就是$P(w_i|w_1, w_2, \dots, w_{i-1})$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的詞有較多選

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。利用(3.5)式來評估調適後的語言模型效能，如表 3.7 所示。

表 3.7：語言模型混淆度評估

語言模型	混淆度(*PPL*)
Read-Speech_tg-LM	733.719
Adapted_tg-LM	280.741

3.2.4 辨識率與詞涵蓋率比較

利用聲學模型對語音辨識將產生音節辨認資訊，若加入語言模型，辨識單元可由音節變為詞，同詞產生 word lattice，但為了與聲學模型比較，我們亦會將詞轉為字元及音節做比較。表 3.8 為本研究調適語言模型後之辨識率。其中 Free-gram-syllable-LM 即為 3.1 節聲學模型之辨認實驗，只有音節辨認率；加入 Read-Speech_tg-LM 之後的實驗開始有字元及詞的辨識率，在此值得注意的是，由於 MCDC 的測試語料中的辨識單元含有「particle」及「marker」，但是 Read-Speech_tg-LM 中並無這兩類的機率，為了比較公平性，在此只計算「一般詞」的辨識率。

表 3.8：加入語言模型之辨識率

case / Acc	Word	Character	Syllable
Free-gram-syllable-LM			48.10%
Read-Speech_tg-LM	38.86%	47.42%	57.20%
Adapted_tg-LM	53.44%	60.34%	65.83%

觀察表 3.8 可注意到在加入 Read-Speech_tg-LM 後音節辨識率上升了 9.1%，此乃因為有了文法的資訊，確實可以幫助中文辨識，但是 Read-Speech_tg-LM 中並沒有「particle」及「marker」之機率，會造成許多刪除型錯誤，另外，訓練語料類型也完全屬於朗讀式語料，在用詞與文法上皆有所不同；而經過調適後 Adapted_tg-LM 辨識率有了大幅改善，其中「particle」及「marker」辨識率分別為 51.03%及 30.56%，加入這兩類詞後能夠有效減少刪除型和取代型錯誤。

此外，在口語對話中有許多口語常用詞，這些詞以連接詞與副詞居多，下表列出一些口語常用詞的辨識率，由表 3.9 可以看出用 MAP 方法調適後更能有效的辨認出來。而表中「然後」、「因為」及「所以」的辨識率較其他詞低，觀察後發現，這些詞常被說話者當作句首的常用插入詞，通常會出現音節合併的現象，造成聲學模型較難建立，這也是自發性語音較難處理的問題。

表 3.9：口語對話常用詞之辨識率

詞(數量) / 模型	就是(132)	我們(88)	然後(115)	因為(71)	沒有(48)	所以(51)
Read-Speech_tg-LM	55.30%	75.00%	46.96%	43.66%	60.42%	43.14%
Adapted_tg-LM	74.24%	80.68%	64.35%	53.52%	72.91%	52.94%

接著，為了未來要加入韻律模型幫助改善辨識率，所以我們先將產生的 word lattice 與辨認標準答案做 Optimal Search 得到詞的涵蓋率(coverage rate)，來判斷韻律模型加入後可以改進的程度。

詞涵蓋率Accuracy的計算方式為(3.6)式

$$\text{Accuracy} = \frac{Match-Ins}{Match+Sub+Del} \tag{3.6}$$

而表 3.10 為朗讀式模型與調適後之語言模型的詞涵蓋率比較

表 3.10：詞涵蓋率比較

case	word coverage rate
Read-Speech_tg-LM	74.89%
Adapted_tg-LM	82.42%

表示調適後 Adapted_tg-LM 所產生的 lattice 較接近我們的正確答案，由詞辨識率 53.44% 和詞涵蓋率 82.42%可以看出，加入其他資訊來幫助辨識還有大幅的提升空間，期望加入韻律訊息來尋找 lattice 上之最佳路徑，以改善辨識率。

3.3 韻律模型

人類在說話時會有抑揚頓挫、高低起伏的表現，這些韻律訊息主要表現在語速(speaking rate)、停頓時長(pause duration)、因高軌跡(pitch contour)、音量大小(energy level)等因素上，而在自發性語音上，語者可能因為猶豫、遲疑造成音節拉長等特殊現象，本研究使用吳聲鋒論文【14】中所訓練的自發性語音韻律模型，其設計考慮了這些特殊單元於模型中。本節首先介紹中文語音韻律階層式架構，接著則介紹階層式韻律模型之設計。

3.3.1 中文語音韻律階層式架構

根據語言學家研究發現【17】，語音的韻律結構會呈現出階層式的架構，以五階層為中文語音韻律的架構，如下圖 3.3 所示：

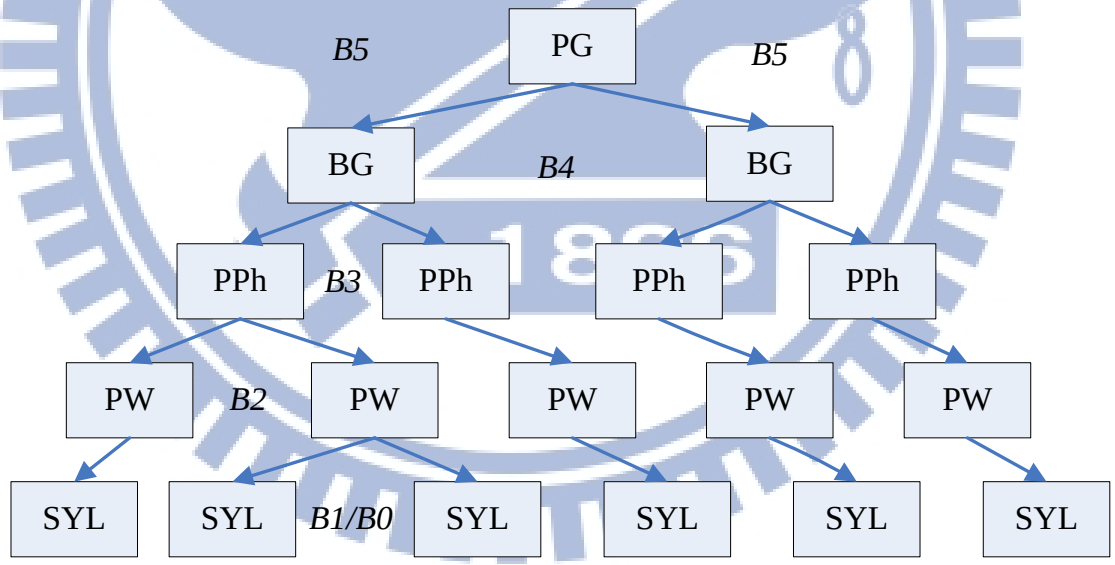


圖 3.3：中文語音韻律之階層式架構

圖中最底層為音節層次(syllable layer, SYL)，中文特性為一個音節一個字，故最底層的韻律單元為音節；向上發展依序為韻律詞層次(prosodic word layer, PW)，由雙音節或多音節所構成的詞組，此詞組通常在語法和語意上有緊密的關係；韻律短語層次(prosodic phrase layer, PPh)，由一個或多個韻律詞組成，結尾通常會帶有不明顯但可察覺

之停頓；呼吸組層次(breath group, BG)，代表一個有音高及音常明顯變化的段落；最上層則為韻律組句(prosodic phrase group)，由連續的呼吸組構成。這整體架構統稱「階層是多短語韻律句群(Hierarchical Prosodic Phrase Grouping，HPG)」架構【18】。

本研究所使用的韻律架構以 HPG 為基礎對其做修改，如圖 3.4 的四層韻律結構：

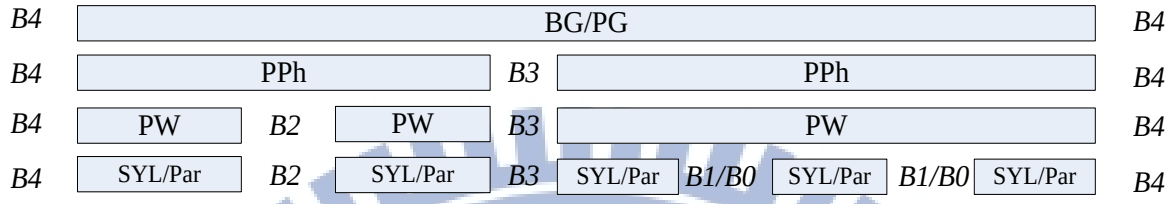


圖 3.4：本研究所使用的中文自發性語音韻律階層式架構

其中主要由七種韻律邊界停頓(break type) **B**={*B*0, *B*1, *B*2-1, *B*2-2, *B*2-3, *B*3, *B*4}來標記四種韻律單元：音節(SYL)、韻律詞(PW)、韻律短語(PPh)、呼吸組/韻律句組(BG/PG)，其對應關係如下表 3.11 所示。

表 3.11：韻律結構之停頓標記

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

另外，在自發性語音中會發生一些特殊現象造成語流不順暢，為了避免這些特殊現象影響其他正常語流之統計特性，我們依其特性分成兩類做為模型設計的依據：一類是可以融入流暢語流的 syllable-like unit；而另一類則歸類成屬於特殊韻律現象的 particular unit(共可分成 26 種類別)，分類方法如下表 3.12 所示：

表 3.12：syllable-like unit 與 particular unit 分類表

類別	分類方法
syllable-like unit	● base syllable ● 有相對應國字的 uncertain
Particular unit	● particle (MHM、E、EI、HAN、HEIN) ● marker (NA-GE、ZHE-GE、SHEN-ME) ● unrecognized speech sound ● Code-switch

3.3.2 階層式韻律模型設計

在本節所介紹之韻律模型主要是幫助語音辨認，自發性語音在韻律上的較不規則使得辨認上的困難，希望利用此模型來幫助搶詞等狀況的改善，本研究是根據【19】使用韻律訊息幫助語音辨認：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 ，如下式(3.7)：

$$\begin{aligned} \Lambda_l^*, \Lambda_p^*, \Upsilon_s^* &= \arg \max_{\Lambda_l, \Lambda_p, \Upsilon_s} P(\mathbf{W}, \mathbf{POS}, \mathbf{PM}, \mathbf{B}, \mathbf{P}, \Upsilon_s \mid \mathbf{X}_a, \mathbf{X}_p) \\ &= \arg \max_{\Lambda_l, \Lambda_p, \Upsilon_s} P(\mathbf{W}, \mathbf{POS}, \mathbf{PM}, \mathbf{B}, \mathbf{P}, \Upsilon_s, \mathbf{X}_a, \mathbf{X}_p) \end{aligned} \tag{3.7}$$

其中 $\mathbf{W} = \{w_1^M\}$ 是代表詞序列；$\mathbf{POS} = \{pos_1^M\}$ 是詞所對應到的詞性序列；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；$\mathbf{B} = \{B_1^N\}$ 則是韻律停頓標記序列，包含七種韻律停頓標記 $B_n \in \{B0, B1, B2\text{-}1, B2\text{-}2, B2\text{-}3, B3, B4\}$；$\mathbf{P} = \{\mathbf{p}, \mathbf{q}, \mathbf{r}\}$ 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(如：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，包含了音節音高軌跡**(sp)**、音節能量強度**(se)**及音節長度**(sd)**；**Y** 代表音節邊界參數，包含了音節間的停頓長度**(pd)**及音節間的能量低點**(ed)**；**Z** 代表音節間的 differential 參數，包含了正規化的音節內基頻差**(pj)**及經正規化過的音節長度拉長因子**(dl)**。在下表 3.13 中將對所有在(3.7)式中有包含到的韻律標記、聲學參數及語言參數做一個統整。

表 3.13：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 feature	**X**: syllable prosodic feature	**sp**: syllable pitch contour **sd**: syllable duration **se**: syllable energy level
	Y: inter-syllabic prosodic feature	**pd**: pause duration **ed**: energy-dip level
	Z: differential prosodic features	**pj**: normalized pitch jump **dl**: normalized duration lengthening factor
L: linguistic feature	**l**: reduced linguistic feature set **t**: syllable tone sequence **s**: base-syllable type **f**: final type	

以下根據(3.7)式提出下列五種假設，以方便韻律模型的設計：

假設一：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** 相依於鄰近的韻律停頓標記 **B**。

經由以上五種假設後，(3.7)式將會簡化成以下形式：

$$\Lambda_l^*, \Lambda_p^*, \Upsilon_s^* \approx \arg\max_{\Lambda_l, \Lambda_p, \Upsilon_s} \{P(\mathbf{X}_a, \Upsilon_s|\mathbf{W})P(\mathbf{W}, \mathbf{POS}, \mathbf{PM}) \cdot P(\mathbf{B}|\Lambda_l)P(\mathbf{P}|\mathbf{B})P(\mathbf{X}|\Upsilon_s, \Lambda_p, \Lambda_l)P(\mathbf{Y}, \mathbf{Z}|\Upsilon_s, \Lambda_p, \Lambda_l)\} \quad (3.8)$$

其中 $P(\mathbf{X}_a, \Upsilon_s|\mathbf{W})$ 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，這部分在後面 4.1.1 節中會做更詳細的介紹；$P(\mathbf{B}|\Lambda_l)$ 是代表停頓語法模型，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停頓

B 的模式，$P(\mathbf{P}|\mathbf{B})$ 稱為韻律狀態轉移模型，用來描述韻律狀態 P 的變化是如何受到韻律停頓 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、L 的影響而產生的變化；$P(\mathbf{Y},\mathbf{Z}|\Upsilon_s,\Lambda_p,\Lambda_l)$ 則為停頓聲學模型，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。以下我們將對【14】所設計的這四種韻律模型作進一步的介紹。

停頓語法模型

我們可以將停頓語法模型 $P(\mathbf{B}|\Lambda_l)$ 近似成以下式：

$$P(\mathbf{B}|\Lambda_l) \approx \prod_{n=1}^{N-1} P(B_n|L_n) \tag{3.9}$$

其中 $P(B_n|L_n)$ 是描述音節韻律停頓與相關的語言參數之間關係的模型，可經由分類樹與決策樹(CART)演算法堆導出來。

停頓聲學模型

我們將停頓聲學模型進一步化簡，可以得到五個子模型，如下式：

$$\begin{aligned} & P(\mathbf{Y},\mathbf{Z}|\Upsilon_s,\Lambda_p,\Lambda_l) \\ & \approx P(\mathbf{pd},\mathbf{ed},\mathbf{pj},\mathbf{dl}|\Upsilon_s,\Lambda_p,\Lambda_l) \\ & \approx \prod_{n=1}^{N-1}\Big\{ g(pd_n;\alpha_{B_n,\Lambda_{l,n}},\beta_{B_n,\Lambda_{l,n}}) N(ed_n;\mu_{ed,B_n,\Lambda_{l,n}},\sigma^2_{ed,B_n,\Lambda_{l,n}}) \\ & \quad \cdot N(pj_n;\mu_{pj,B_n,\Lambda_{l,n}},\sigma^2_{pj,B_n,\Lambda_{l,n}}) N(dl_n;\mu_{dl,B_n,\Lambda_{l,n}},\sigma^2_{dl,B_n,\Lambda_{l,n}})\Big\} \end{aligned} \tag{3.10}$$

其中 pause duration (pd_n) 為第 n 個音節跟隨的接合點停頓長度，以 Gamma distribution 來模擬；energy dip (ed_n) 為第 n 個接合點的能量下降程度，以 normal distribution 模擬；pitch jump (pj_n) 為跨越第 n 個接合點的正規化基頻差，其定義為(3.11)式，以 normal distribution 模擬；normalized duration lengthening (dl_n) 則為正規化的音節長度拉長因子，其定義為(3.12)式，同樣以 normal distribution 模擬。

$$pj_n = (\mathbf{sp}_{n+1}(1) - \boldsymbol{\beta}_{t_{n+1}}(1)) - (\mathbf{sp}_n(1) - \boldsymbol{\beta}_{t_n}(1)) \tag{3.11}$$

(3.11)式中，$sp_n(1)$ 為第 n 個音節音高軌跡的第一維度；$\beta_{t_n}(1)$ 為第 n 個音節聲調影響因素的第一維度。

$$dl_n = (sd_n - \gamma_{t_n} - \gamma_{s_n}) - (sd_{n-1} - \gamma_{t_{n-1}} - \gamma_{s_{n-1}}) \tag{3.12}$$

(3.12)式中，sd_n 為第 n 個音節長度；γ_{t_n} 為第 n 個音節長度影響因素。

在實作過程中，$P(\mathbf{pd},\mathbf{ed},\mathbf{pj},\mathbf{dl}|\Upsilon_s,\Lambda_p,\Lambda_l)$ 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 maximum likelihood gain，而 CART 演算法主要是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音節的 pd_n、ed_n、pj_n 和 dl_n 做好分類。

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：

$$\begin{aligned} &P(\mathbf{P}|\mathbf{B}) = P(\mathbf{p}|\mathbf{B})P(\mathbf{q}|\mathbf{B})P(\mathbf{r}|\mathbf{B}) \approx P(p_1)P(q_1)P(r_1) \\ &\cdot\left[\prod_{n=2}^{N} P(p_n \mid p_{n-1}, B_{n-1})P(q_n \mid q_{n-1}, B_{n-1})P(r_n \mid r_{n-1}, B_{n-1})\right] \end{aligned} \tag{3.13}$$

其中 $P(p_n \mid p_{n-1}, B_{n-1})$、$P(q_n \mid q_{n-1}, B_{n-1})$ 及 $P(r_n \mid r_{n-1}, B_{n-1})$ 分別表示三種不同韻律狀態，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。

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：

$$\begin{aligned} &P(\mathbf{X}|\Upsilon_s,\Lambda_p,\Lambda_l) \\ &\approx P(\mathbf{sp}|\Upsilon_s,\mathbf{B},\mathbf{p},\mathbf{t})P(\mathbf{sd}|\Upsilon_s,\mathbf{B},\mathbf{q},\mathbf{t},\mathbf{s})P(\mathbf{se}|\Upsilon_s,\mathbf{B},\mathbf{r},\mathbf{t},\mathbf{f}) \\ &\approx \prod_{n=1}^{N} P(sp_n|p_n, B_{n-1}^{n}, t_{n-1}^{n+1})P(sd_n \mid q_n, s_n, t_n)P(se_n \mid r_n, f_n, t_n) \end{aligned} \tag{3.14}$$

分別為音節軌跡、音節長度、音節能量，這三個子模型可以拆解成各個影響因子(affecting factor)之貢獻，以下分別介紹這三個子模型：

◆ **Syllable pitch contour model**

$$p(\mathbf{sp}|\mathbf{B},\mathbf{p},\mathbf{t}) \approx \prod_{n=1}^{N} P(sp_n | p_n, B_{n-1}^{n}, t_{n-1}^{n+1}) \tag{3.15}$$

其中

$$\mathbf{sp}_n = \begin{cases} \mathbf{sp}_n^r + \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\beta}^f_{B_{n-1},tp_{n-1}} + \boldsymbol{\beta}^b_{B_n,tp_n} + \boldsymbol{\mu}_{sp} & \text{, if } n\text{th syllable is syllable-like unit} \\ \mathbf{sp}_n^{pr\_r} + \boldsymbol{\beta}_{pr_n} + \boldsymbol{\beta}_{pr\_p_n} + \boldsymbol{\mu}_{pr\_sp} & \text{, if } n\text{th syllable is particular unit} \end{cases} \quad \text{for } 1 \le n \le N$$

為第 n 個音節之 syllable log-F0 contour，是由四維正交化係數 $[a_0 \quad a_1 \quad a_2 \quad a_3]^T$ 以向量的方式表示；$B_{n-1}^{n}=(B_{n-1},B_n)$；$t_{n-1}^{n+1}=(t_{n-1},t_n,t_{n+1})$；$\boldsymbol{\mu}_{sp}$ 及 $\boldsymbol{\mu}_{pr\_sp}$ 分別為 syllable-like unit 及 particular unit 的總體平均值(global mean)，僅第一維為非零值；$\mathbf{sp}_n^r$ 及 $\mathbf{sp}_n^{pr-r}$ 為 n 屬於 syllable-like unit 或 particular unit 的 $\mathbf{sp}_n$ 正規化（normalization）後的基頻殘餘值(residual)；$\boldsymbol{\beta}_x$ 則為某一影響因子 x 之影響型態（Affecting Pattern, AP），各種 AP 說明如下：

➢ Syllable-like unit 的 APs

$\boldsymbol{\beta}_{t_n}$ 為目前音節聲調 $t_n \in$ {lexical tone 1~5} 的 APs；$\boldsymbol{\beta}_{p_n}$ 為目前音節韻律狀態 $p_n \in$ {prosody state 1~16} 的 APs，其中韻律狀態的影響只限制對目前音節的 LogF0 level，故將 $\boldsymbol{\beta}_{p_n}$ 的四維正交係數，僅第一維設為非零值；$\boldsymbol{\beta}^f_{B_{n-1},tp_{n-1}}$ 及 $\boldsymbol{\beta}^b_{B_n,tp_n}$ 分別是第 n-1 個和第 n+1 個音節所貢獻的前後連音效應(forward and backward coarticulations) APs，其中 tp_{n-1} 是 tone pair $t_{n-1}^{n}=(t'_{n-1},t_n)$，$t'_{n-1} \in$ {lexical tone 1~5 for (*n-1*)*th* syllable is syllable-like unit；otherwise is particular unit}；tp_n 是 tone pair $t_n^{n+1}=(t_n,t'_{n+1})$，$t'_{n+1} \in$ {lexical tone 1~5 for (*n+1*)*th* syllable is syllable-like unit; otherwise is particular unit}。此外，針對 utterance boundary 再另外給定兩個特殊的 break type B_b 和 B_e 來表示所有 utterance 的起頭和結束

位置(i.e. $B_0=B_b$ 和 $B_N=B_e$)，以及兩個代表開頭和結尾連音效應的特殊 APs $\boldsymbol{\beta}^{f}_{B_b,t_1}=\boldsymbol{\beta}^{f}_{B_0,tp_0}$ 和 $\boldsymbol{\beta}^{b}_{B_e,t_N}=\boldsymbol{\beta}^{b}_{B_N,tp_N}$ 。

- Particular unit 的 APs

$\boldsymbol{\beta}_{pr_n}$ 為各種 particular unit $pr_n \in$ {particular unit type 1~26} 的 APs；$\boldsymbol{\beta}_{pr\_p_n}$ 為 particular unit 的目前音節韻律狀態 $pr\_p_n \in$ {prosody state 1~4} 的 APs。

◆ **Syllable duration model**

$$P(\mathbf{sd}|\Upsilon_s,\mathbf{B},\mathbf{q},\mathbf{t},\mathbf{s}) \approx \prod_{n=1}^{N} P(sd_n|q_n,s_n,t_n) \qquad (3.16)$$

其中

$$sd_n = \begin{cases} sd_n^r + \gamma_{t_n} + \gamma_{s_n} + \gamma_{q_n} + \mu_{sd} & \text{, if } n\text{th syllable is syllable-like unit} \\ sd_n^{pr-r} + \gamma_{pr_n} + \gamma_{pr\_q_n} + \mu_{pr\_sd} & \text{, if } n\text{th syllable is particular unit} \end{cases} \quad \text{for } 1 \le n \le N$$

為第 n 個音節之音節長度，其中 μ_{sd} 及 μ_{pr_sd} 分別為 syllable-like unit 及 particular unit 的總體平均值(global mean)；sd_n^r 及 sd_n^{pr-r} 為 n 屬於 syllable-like unit 或 particular unit 的 sd_n 正規化後的殘餘值；γ_x 則為某一影響因子 x 之影響型態（Affecting Pattern, AP），各種 AP 說明如下：

- Syllable-like unit 的 APs

γ_{t_n} 為目前的音節聲調 $t_n \in$ {lexical tone 1~5} 對 sd_n 的 APs；γ_{s_n} 為基本音節類型 $s_n \in$ {base syllable type 1~82} 對 sd_n 的 APs；γ_{q_n} 為目前韻律狀態 $q_n \in$ {prosody state 1~16} 對 sd_n 的 APs

- Particular unit 的 APs

γ_{pr_n} 為各種 particular unit $pr_n \in$ {particular unit type 1~26} 的 APs；$\gamma_{pr\_q_n}$ 為 particular unit 的目前音節韻律狀態 $pr\_q_n \in$ {prosody state 1~4} 的 APs。

◆ **Syllable energy model**

$$P(\mathbf{se}|\Upsilon_s,\mathbf{B},\mathbf{r},\mathbf{t},\mathbf{f}) \approx \prod_{n=1}^{N} P(se_n|r_n,f_n,t_n) \tag{3.17}$$

其中

$$se_n = \begin{cases} se_n^r + \alpha_{t_n} + \alpha_{f_n} + \alpha_{r_n} + \mu_{se} & \text{, if } n\text{th syllable is syllable-like unit} \\ se_n^{pr\_r} + \alpha_{pr_n} + \alpha_{pr\_r_n} + \mu_{pr\_se} & \text{, if } n\text{th syllable is particular unit} \end{cases} \quad \text{for } 1 \le n \le N$$

為第 n 個音節之音節能量，其中 μ_{se} 及 μ_{pr_se} 分別為 syllable-like unit 及 particular unit 的總體平均值(global mean)；se_n^r 及 $se_n^{pr\_r}$ 為 n 屬於 syllable-like unit 或 particular unit 的 se_n 正規化後的殘餘值；α_x 則為某一影響因子 x 之影響型態（Affecting Pattern, AP），各種 AP 說明如下：

- Syllable-like unit 的 APs

α_{t_n} 目前音節聲調 $t_n \in$ {lexical tone 1~5} 對 se_n 的 APs；α_{f_n} 為聲母類型 $f_n \in$ {final type 1~40} 對 se_n 的 APs；α_{r_n} 為目前韻律狀態 $r_n \in$ {prosody state 1~16} 對 se_n 的 APs

- Particular unit 的 APs

α_{pr_n} 為各種 particular unit $pr_n \in$ {particular unit type 1~26} 的 APs；$\alpha_{pr\_r_n}$ 為 particular unit 的目前音節韻律狀態 $pr\_r_n \in$ {prosody state 1~4} 的 APs。

第四章 語音辨認系統架構

在本章節中，我們利用劉銘傑【19】論文中使用階層式語音韻律模型於中文語音辨認系統，本研究則用在自發性語音上，希望藉由韻律資訊來改善辨識率。在 4.1 節中會說明如何將韻律模型以 two-stage 的方式加入語音辨認中，首先利用 HMM-based 語音辨認器來做 first-stage 語音辨認，將產生之 word lattice 做音節切割後，加入韻律資訊做重新評分的動作，得到新的辨認結果。4.2 節則是說明本研究使用鑑別式模型組合(Discriminative Model Combination)來解決重新評分的過程中，多個模型之權重問題。

4.1 加入韻律訊息於 two-stage 語音辨認系統

下圖 4.1 為 two-stage 語音辨認系統流程圖，以下將對此系統第二個 stage 做詳細介紹。

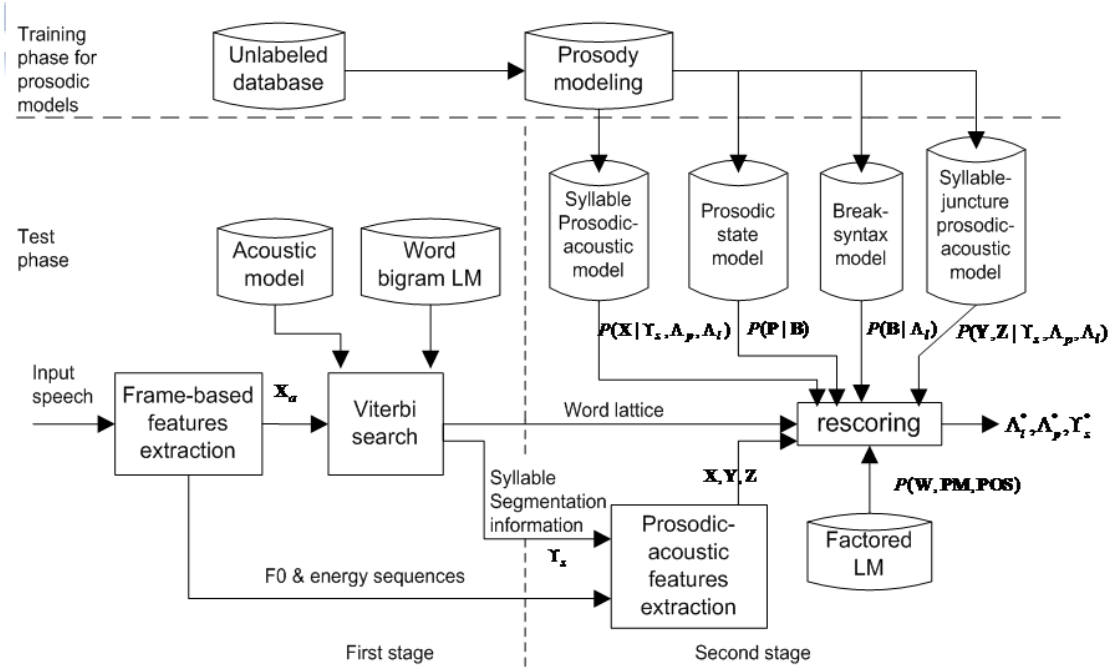


圖 4.1：以 two-stage 方式之韻律輔助中文語音辨認系統流程圖

4.1.1 Joint Syntax Model 之架構與建立

由 3.3 小節可以發現到，韻律模型的建立需要給定語言參數資訊，其中包含詞性(POS)以及標點符號(PM)這兩種語言參數資訊，本研究所使用的 joint syntax model 包含一個 trigram LM、一個 factored POS model 與一個 factored PM model，在這裡我們是使用 FLM approach【20】建構 factored POS model 和 factored PM model，並使用 SRILM toolkit 及利用 Witten-Bell smoothing 的方式來訓練模型，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。

當使用多種語言資訊做預估時難免會面臨到資料量不足的問題，因此會採取退化(back-off)的架構，factored POS model 與 factored PM model 的退化路徑如下：

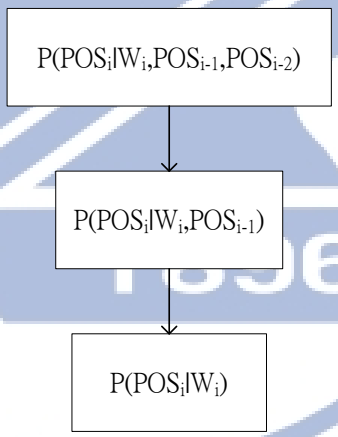


圖 4.2：factored POS model 的 back-off 路徑

如上圖 4.2 所示 factored POS model 的 back-off 路徑，在最上層的情況，期望以前兩個 POS 與目前的詞來預估，若此機率的組合沒有出現，則丟棄最遠的 POS，若仍是沒有出現，就退化到最下層的狀態，此時就一定有機率。

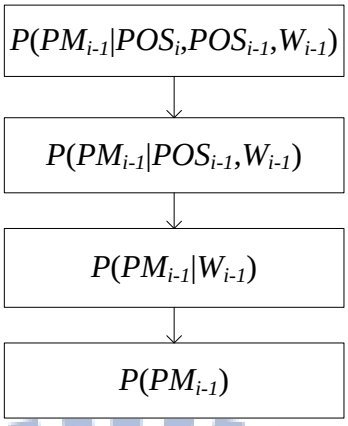


圖 4.3：factored PM model 的 back-off 路徑

factored PM model 亦是如此，如上圖 4.3 所示，我們使用了前一個詞、POS 以及現在的 POS 來預估中間的 PM 機率為何，若是沒此機率，則丟掉現在的 POS，若再沒有機率，接著丟掉前一個 POS，最後退化到最下層的狀態，此時就一定有機率。

另外，在 3.2.2 節中我們依其特性加入了兩類詞「particle」與「marker」，而以數量上來看，MCDC 的資料量太少，必須經過調適後才能更符合我們要的模型，所以考慮像調適 Word tri-gram LM 的方法，使用 model interpolation 來調適 Factored POS model 與 Factored PM model。

在訓練 factored POS model 時分成 48 類詞姓，另外加上 particle 與 marker 兩類，共 50 類詞姓；factored PM model 則依據標點符號的性質分成：逗號(COM)、頓號(DOT)、無標點符號(NONE)、句點或驚嘆號等具有句子結束意義的標點符號(OTH)，共 4 類。

兩種 factored model 訓練完成後，其 perplexity 效能評估於表 4.1 及表 4.2，觀察 perplexity 的下降，可以看出每加入一個 factor 都有助於預估 POS 或 PM。

表 4.1：factored POS model 的 perplexity

factored POS model with different factors	perplexity
$P(POS_i\|W_i)$	2.24832
$P(POS_i\|W_i,POS_{i-1})$	2.06332
$P(POS_i\|W_i,POS_{i-1},POS_{i-2})$	1.92078

表 4.2：factored PM model 的 perplexity

factored PM model with different factors	perplexity
$P(PM_{i-1}\|W_{i-1})$	2.48816
$P(PM_{i-1}\|W_{i-1},POS_{i-1})$	2.44734
$P(PM_{i-1}\|W_{i-1},POS_{i-1},POS_{i})$	2.28617

※ *PPL*= 10^(-logprob / (words - OOVs + sentences))

4.1.2 參數正規化

針對圖 4.1 中 second stage，如需要加入音節能量、音節長度或音節基頻等韻律聲學參數時，為了消除不同語者先天上發音的差異，我們必須先對其作正規化的動作。首先我們將 first stage 所產生的最佳辨認結果(top 1)對測試音檔特徵參數做強迫對齊(force alignment)，接著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，完成後就可執行正規化。

首先，音節基頻正規化我們以 frame-based 的對數基頻做正規化，其數學式如下：

$$\hat{f}_i^s = \left(\frac{f_i^s - \mu_s}{\sigma_s} \right) \cdot \sigma_{average} + \mu_{global} \tag{4.1}$$

其中 f_i^s 為第 s 個音檔的第 i 個音框的原始基頻對數值；$\hat{f}_i^s$ 為正規化後的基頻對數值；μ_s 及 σ_s 則分別為第 s 個音檔基頻數值之平均值及標準差，數學式如下：

$$\mu_s = \frac{\sum_{i=1}^{I(s)} f_i^s}{I(s)} \tag{4.2}$$

$$\sigma_s = \sqrt{\frac{\sum_{i=1}^{I(s)} (f_i^s - \mu_s)^2}{I(s)}} \tag{4.3}$$

在(4.2)式及(4.3)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。

而音節能量、音節長度的正規化方法也與基頻正規化類似，只是以 syllable-based 做正規化，對音節能量而言，(4.1)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，(4.1)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。

4.1.3 The Second Stage 之實作

在第二個 stage 開始之後，需要加入許多韻律資訊來幫助辨認，為了瞭解每個模型對於辨識系統的影響力，本研究根據【19】將針對 second stage 再細分成三個階段，逐次加入模型資訊並觀察實驗結果，其詳細流程圖如下：

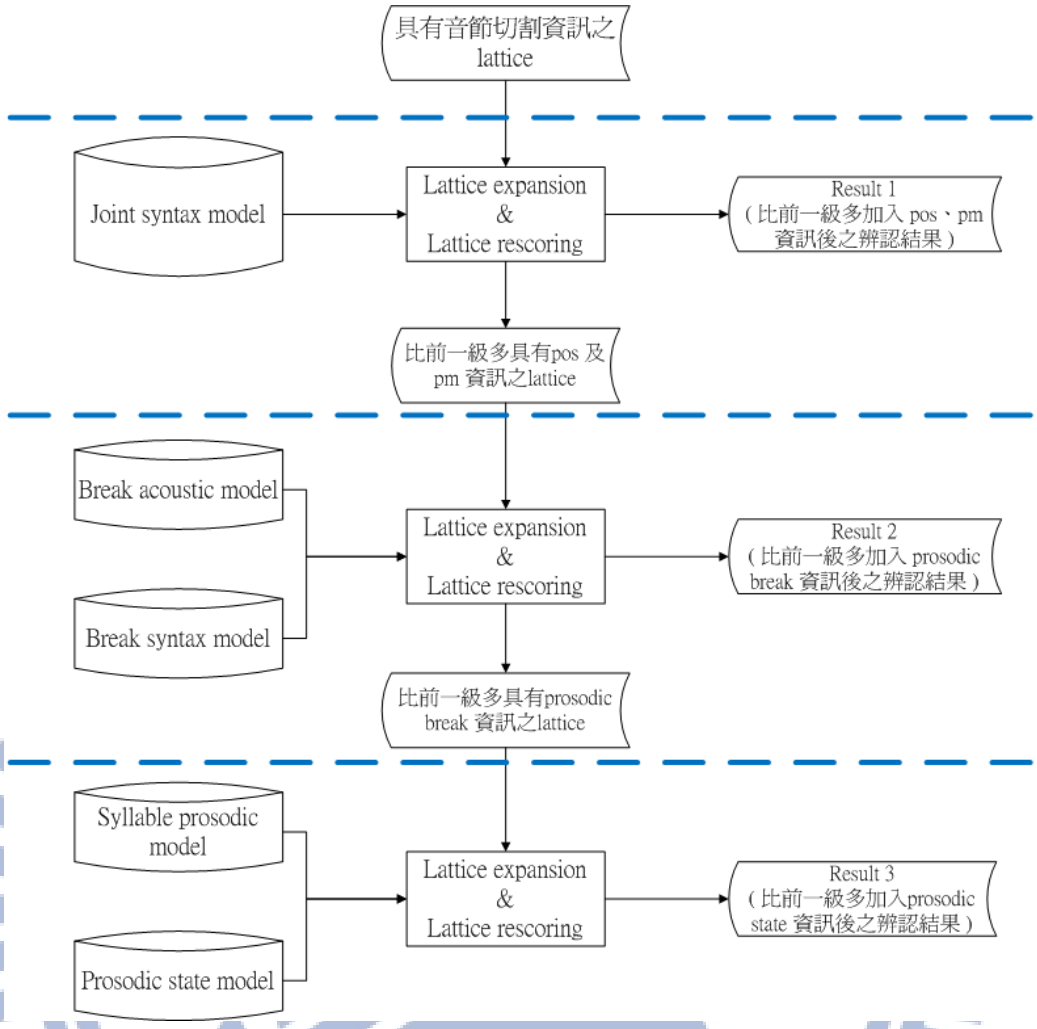


圖 4.4：辨認器第二級三階段實作流程圖

第一階段：加入多種語言資訊

如圖 4.4 所示，辨認器第二級第一階段是引入多種語言參數資訊(POS 及 PM)，在這裡需要加入的模型為 joint syntax model，此時在 word lattice 上每個 node 所帶有的 word 資訊會根據 factored POS model 與 factored PM model 找出相對應的詞性與標點符號做展開動作。

在 lattice 展開之後，我們將每個模型的分數乘上其權重後加起來，所以每一個 arc 上會累積 AM、LM、POS 及 PM 的分數，再做重新評分來得到一條最佳路徑作為辨識答案，並解碼出相關的語言資訊。

(1). node expansion：

在 word lattice 上每個 node 所帶有的 word 資訊會根據 factored POS model 找出其最佳對應的詞性(POS)，搭配 4 種標點符號(PM)，將各個 node 展開至 1*4 倍。

(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(1)及 PM 數目(4)，並對上一個 arc 中所帶有的 word 資訊，找出相對應的 POS 數目(1)及 PM 數目(4)，再將各個 arc 展開至 1*4*1*4 倍。

第二階段：加入韻律邊界停頓資訊

在第二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，根據數學式(3.9)、(3.10)算出韻律停頓相關分數，但因為韻律基本單元為音節，所以針對 intra-word syllable 的部分可由程式內部處理，不需要將 lattice 做展開，而 inter-word syllable 的部分則是將 arc 上 word 的最後一個音節作展開，至於 word 中第一個音節資訊利用 backward Viterbi search，利用資訊傳遞的特性存入下一個 word 中的第一個音節。

在此階段重新評分的方法與第一階段相同，只是加入了韻律停頓模型之分數於每個 arc 上，除了會計算一條最佳路徑作為辨識答案外，也會解碼出韻律邊界停頓資訊。

(1). node expansion：

針對原始 lattice 中各個 node，觀察其來源 arc 中所帶有的 word 資訊中最後一個音節資訊，現假設來源 arc 中最後一個音節的資訊共有 *M* 倍。

(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。

第三階段：加入音節韻律狀態資訊

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，根據數學式(3.13)、(3.14)算出韻律狀態相關分數，而這一階段不同於【19】

的是，我們多給了特殊音節 4 個韻律狀態；而音節韻律模型方面，我們針對特殊音節的部分多加了 3 個分數(特殊音節之音節基頻、音節長度及音節能量分數)，依據(3.15)、(3.16)及(3.17)式的特殊音節部分對程式做修改，判斷是否為特殊音節來給予不同分數，也就是每個音節的三種音節韻律模型分數只會來自於一般音節或是特殊音節其中之一。在此 intra-word 也是由程式內部處理，而 inter-word 的部分則採用 forward viterbi search，利用資訊傳遞的特性將 word 中前一個音節狀態資訊存入上一個 word 中最後一個音節資訊。

第三階段重新評分的方式與前兩階段相同，計算出一條最佳路徑作為辨識結果，並同時解碼出詞中每個音節的三種韻律狀態資訊。

(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 資訊中第一個音節的聲調，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。

(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。

4.2 鑑別式模型組合

在辨認階段 second stage 中我們加入了韻律相關資訊做重新評分，而總共用到的模型多達 18 種之多，包含：

- Acoustic Model
- Language Model
- Factored POS Model
- Factored PM Model
- prosodic break 相關 5 個
  (pause duration、pitch jump、energy dip、duration lengthening、break syntax model)
- prosodic state 相關 3 個
  (pitch prosodic state、duration prosodic state、energy prosodic state)
- syllable prosodic model for syllable-like unit 相關 3 個
  (syllable pitch、syllable duration、syllable energy)
- syllable prosodic model for particular unit 相關 3 個
  (syllable pitch、syllable duration、syllable energy)

因此如何找到一組權重使這 18 個模型結合後能有最好的辨認率便是非常重要之課題，而本研究也是根據【19】論文中，使用鑑別式模型組合(Discriminative Model Combination, DMC) 【21】的方法來決定權重。

DMC 的方法是先定義一個 decision error rate 的鑑別式函數(discriminant function)如(4.4)式，目標是找到一組權重使此函數的 decision error rate 最佳化。

$$\begin{aligned} & g(x_1^T, w_1^S, w'^{S'}_1) \\ & = \log P(w_1^S \mid x_1^T) - \log P(w'^{S'}_1 \mid x_1^T) \\ & = \log[P(w_1^S)P(x_1^T \mid w_1^S)] - \log[P(w'^{S'}_1)P(x_1^T \mid w'^{S'}_1)] \end{aligned} \tag{4.4}$$

(4.4)式中 $w_1^S = (w_1, \ldots w_S)$ 代表詞串，$x_1^S = (x_1, \ldots x_T)$ 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，當 $P(w'^{S'}_1 \mid x_1^T)$ 分數越接近 $P(w_1^S \mid x_1^T)$ 越好，代表其 likelihood 最好，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：

$$P_{\{\Lambda\}}^{\Pi}(x_1^T \mid w_1^S) = \exp\{\log C(\Lambda) + \sum_{j=1}^{M} \lambda_j \log P_j(x_1^T \mid w_1^S)\} \tag{4.5}$$

(4.5)式中 $\Lambda = (\lambda_1, ..., \lambda_M)^{tr}$ 代表 M 個模型 P_j 分數組合時的權重；$C(\Lambda)$ 代表正規化因子。

$P(w_1^S \mid x_1^T)$ 可拆成 M 個不同模型，其鑑別式函數將改寫成下式：

$$g(x_1^T, w_1^S, w'^{S'}_1) = \sum_{j=1}^{M} \lambda_j (\log P_j(w_1^S \mid x_1^T) - \log P_j(w'^{S'}_1 \mid x_1^T)) \tag{4.6}$$

最後將定義一個 smooth misclassification function $\ell(x_n, k_{n0}, \Lambda)$，並搭配 Generalized Probabilistic Descent (GPD) algorithm【21】來得到多個模型的權重值 Λ，以下對所用到的符號做定義：

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：訓練語料的 smoothed empirical error rate $L(\Lambda)$ 為：

$$L(\Lambda) = \frac{1}{N} \sum_{n=1}^{N} \ell(x_n, k_{n0}, \Lambda) \tag{4.7}$$

$$\ell(x_n, k_{n0}, \Lambda)^{-1} = 1 + A \cdot \left(\frac{1}{K} \sum_{r=1}^{K} e^{\left\{ -\eta LD(k_{nr}, k_{n0}) \log \frac{P_{\{\Lambda\}}^{\Pi}(k_{n0} \mid x_n)}{P_{\{\Lambda\}}^{\Pi}(k_{nr} \mid x_n)} \right\}} \right)^{-\frac{B}{\eta}}, \mathrm{A} > 0, \mathrm{B} > 0, \eta > 0 \tag{4.8}$$

然後利用下面的遞迴架構即可求出權重值 λ_j：

For j=1,...,M

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

$$\lambda_j^{(I+1)} = \lambda_j^{(I)} + \varepsilon \sum_{n=1}^{N} \ell(x_n, k_{n0}, \Lambda^{(I)})\left(1 - \ell(x_n, k_{n0}, \Lambda^{(I)})\right) \cdot$$

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

$$\Lambda^{(I+1)} = (\lambda_1^{(I+1)}, \ldots, \lambda_M^{(I+1)})^{tr} \qquad (4.9)$$

在 (4.9) 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。

在實作方面，首先選取訓練語料的一部分作為 develop set，利用 HTK 產生 top-N 辨認結果詞串，對辨認答案作 force-alignment 得到正確詞串，此時詞串上會有 AM 及 LM 的分數，然後加入 factored LM 得到 POS 和 PM 分數，再分別依據 3.3.2 節之四個子模型將詞中音節資訊加入韻律停頓及韻律狀態分數，有了所有分數之後，利用上述演算法得到 18 個模型權重。下面列出三個階段之模型權重：

- **第一階段(+POS & PM)**

#AM: 10.134861
#LM: 111.274605
#POS: 1.021641
#PM: 22.432690

- **第二階段(+Prosodic break)**

#AM: 10.223469
#LM: 125.614311
#POS: 14.166499
#PM: 20.945004

------break weight------

#pd: 10.731343

#pj: 44.041885

#ed: 13.118249

#dl: 8.927853

#break syntax: 14.193851

- **第三階段(+Prosodic state)**

#AM: 10.712511

#LM: 159.646785

#POS: 14.074561

#PM: 20.771852

------prosodic state weight------

#pitch prosodic state: 86.124034

#duration prosodic state: 81.948402

#energy prosodic state: 17.294501

------syllable-like prosodic model weight------	------particular prosodic model weight------
#syllable pitch model: 42.010248	# syllable pitch model: 35.779157
#syllable duration model: 21.982106	# syllable duration model: 25.804143
#syllable energy model: 1.486401	# syllable energy model: 1.718144

(做第三階段時，前一階段已經將 break 的分數記錄在 lattice 當中，所以不需要另外給 break 的權重)

第五章 實驗結果與分析

本章列出加入韻律資訊於自發性語音辨認系統後重新評分的結果，並對此系統的辨認結果做分析討論。

5.1 加入韻律信息之辨認率

在第二階段辨識中我們分別加入不同韻律訊息來觀察其幫助，分別是詞性(POS)及標點符號(PM)資訊、韻律停頓(prosodic break)資訊和韻律狀態(prosodic state)資訊，以下數據列出詞辨識率、字元辨識率及音節辨識率，表格第一列為 Baseline 系統，代表第一階段中所產生之辨識率；第二列為加入 joint syntax model 後之辨識率；第三列為加入 break syntax model 及 break acoustic model 進行辨認考量，第四列為加入 syllable prosodic model 及 prosodic state model 進行辨識考量：

表 5.1：詞(word)辨認率

	Accuracy(%)
Baseline	53.86
+PM & POS	56.74
+Prosodic break	57.39
+Prosodic break +Prosodic state	58.29

表 5.2：字元(character)辨認率

	Accuracy(%)
Baseline	60.34
+PM & POS	63.26
+Prosodic break	64.73
+Prosodic break +Prosodic state	64.94

表 5.3：音節(syllable)辨認率

	Accuracy(%)
Baseline	65.83
+PM & POS	68.03
+Prosodic break	68.61
+Prosodic break +Prosodic state	68.89

下表 5.4 與表 5.5 則是三階段辨識中所解碼出的詞性(POS)及標點符號(PM)之辨識率，其計算辨認率採用 F-measure 的方式，以下將對此計算方法進行說明。

表 5.4：詞性(POS)辨認率

	Precision	**Recall**	**F-measure**
+PM & POS	85.16	53.91	66.02
+Prosodic break	86.03	54.10	66.43
+Prosodic break +Prosodic state	85.14	54.22	66.25

表 5.5：標點符號(PM)辨認率

	Precision	**Recall**	**F-measure**
+PM & POS	72.08	43.88	54.55
+Prosodic break	74.60	44.83	56.00
+Prosodic break +Prosodic state	72.39	44.16	54.86

5.1.1 詞性(POS)辨認率算法

在此計算 POS 辨認率不是直接拿辨認結果和標準答案做計算，因為當詞辨認證確實，其辨認出來的 POS 才有意義，所以我們使用 F-measure 的計算方式，首先統計出在詞辨認正確條件之下 POS 辨認正確的數量 ***H***；以及在詞辨認正確的條件之下 POS 總數 ***N***；和 POS 答案中的總數量 ***R***，接下來分別計算 POS 的 Recall (***H*** / ***R***)、詞辨認正確的條件之下 POS 的 Precision (***H*** / ***N***)，最後利用 Precision 及 Recall 就可以算出 F-measure，公式如下：

$$F-measure=\frac{2\times Precision\times Recall}{Precision+Recall} \quad (5.1)$$

5.1.2 標點符號(PM)辨認率算法

在此計算 PM 辨認率不是直接拿辨認結果和標準答案做計算，因為當詞辨認證確實，其辨認出來的 PM 才有意義，所以我們使用 F-measure 的計算方式，首先統計出在詞辨認正確條件之下 PM 辨認正確的數量 ***H***；以及在詞辨認正確的條件之下 PM 總數 ***N***；和 PM 答案中的總數量 ***R***，接下來分別計算 PM 的 Recall (***H*** / ***R***)、詞辨認正確的條件之下 PM 的 Precision (***H*** / ***N***)，最後有了 Precision 及 Recall 就可以利用(5.1)式算出 F-measure。

5.2 辨認結果分析與比較

從 5.1 節的辨識率中可以看出，每加入一種韻律資訊時，辨識效能都有小幅的上升，但是經過觀察發現，辨認結果常會發生一連串辨認正確、又一連串辨認錯誤的情況，所以本節首先針對這個特性分析其原因，最後再列出韻律資訊加入而改善辨認的部分。

5.2.1 錯誤分析

MCDC 自發性語音由於說話方式較不固定，常常一句話語速突然變很快、講話含糊或是能量忽大忽小，這些都是造成辨認不佳的主要原因，刪除型錯誤比朗讀式語音多上許多，下圖 5.1 為一個辨識率較低的例子，該語者在 A 及 B 處語速變快，造成許多詞辨認不出來，另外，該語句之能量偏低(43.47 dB，global energy 為 51.80 dB)，講話聽不清楚。

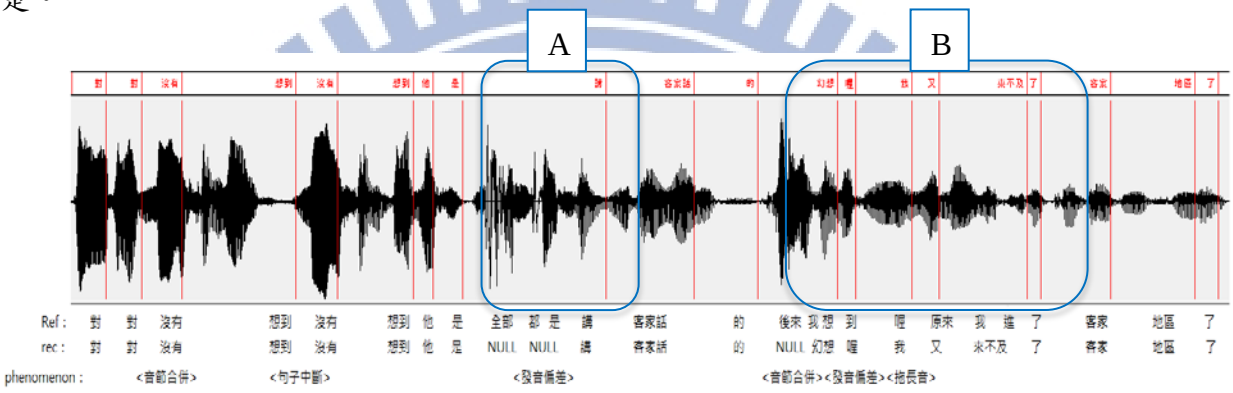


圖 5.1：辨識率較低的例子

另外，下表 5.6 則列出各語者的平均音節長度和詞辨識率，可以發現平均音節長度較短的語者，其詞辨識率也較低，這可能是影響辨識率的原因之一。

表 5.6：各語者之平均音節長度與詞辨識率

語者(句數)	平均音節長度(sec)	詞辨識率(Acc)	語者(句數)	平均音節長度(sec)	詞辨識率(Acc)
1L (7)	0.183	57.63	9L (14)	0.175	65.08
1R (12)	0.171	60.72	9R (4)	0.176	66.55
2L (10)	0.175	61.15	10L (8)	0.147	47.39
2R (7)	0.156	36.16	10R (8)	0.167	48.71
3L (12)	0.168	49.94	25L (10)	0.159	60.88
3R (8)	0.168	49.37	25R (9)	0.160	64.33
5L (12)	0.172	66.92	26L (6)	0.164	55.76
5R (11)	0.176	63.89	26R (12)	0.157	39.65

5.2.2 辨認結果之改善

下面表 5.7、表 5.8 與表 5.9 列出經由加入韻律資訊後改善詞邊界判斷錯誤的搶詞問題、改善因為語速快造成音節合併現象的辨認以及聲調修正的情況，表格中第一欄為正確答案(標出其各種特殊現象)，第二欄為第一階段 baseline 系統的辨認結果，第三欄則是加入所有韻律資訊後的辨認結果，並且將解碼出的韻律邊界停頓標示在第四欄。

表 5.7：搶詞狀況的改善

mcdc-03-19_L0080588			
就	NULL	就	B1
搶 <音節合併 1>	酒廠	搶	B1
的 <音節合併 2>	的	的	B1
很	NULL	NULL	NULL
辛苦	辛苦	辛苦	B2-1,B2-1
mcdc-03-20_R0108676			
就是 <拖長音>	就是說	就是	B0,B1
所以 <音節合併>	NULL	NULL	NULL
我	NULL	我	B0
就	就	就	B1
覺得	覺得	覺得	B0,B1
有點	有點	有點	B1,B2-1
壓力	壓力	壓力	B1,B2-1
大	NULL	大	B2-1
吧	打法	吧	B1

呼吸聲

表 5.8：音節合併辨認的改善

mcdc-01-10_R0172836			
我們	我們	我們	B1,B1
沒	沒	沒	B1
話	話	話	B2-1
講	講	講	B2-1
只要 <音節合併>	叫	只要	B0,B1
他 <發音偏差>	他	他	B0

在 基本 標準	在 基本 標準	在 基本 標準	B1 B1,B2-1 B0,B1
mcdc-05-17_R0150960			
壹週刊 就 把 <音節合併 1> 人家<音節合併 2> 刊 出來 嘛	壹週刊 NULL NULL 大量 看 出來 嗎	壹週刊 就 NULL 人家 看 出來 嗎	B2-1,B1,B1 B1 NULL B0,B1 B1 B0,B1 B1

表 5.9：聲調的修正

mcdc-05-05_L0100477			
還 有 就是 本身 自己 的 房子 貸款	還 有 就是 本 省 之 一 NULL 房子 貸款	還 有 就是 本身 是 以 NULL 房子 貸款	B2-1 B2-2 B2-1,B2-1 B2-1,B1 B1,B1 NULL B1,B2-2 B1,B2-1
mcdc-10-08_L0089277			
那 <音節合併 1> 我們<音節合併 2> 眷村 就是 <音節合併> 哪 邊 蓋好<音節合併 1> 了 <音節合併 2>	那 我們 捐 髓 就是 哪 邊 蓋好 NULL	那 我們 眷村 就是 哪 邊 蓋好 NULL	B1 B0,B3 B1,B1 B1,B2-2 B1 B1 B0,B2-1 NULL

句子中斷

第六章 結論與未來展望

6.1 結論

中文自發性語音辨認效能不佳有幾個原因，我們針對其特性做修正及改進，首先，語言模型因為語料不足及與朗讀式特性不同，本研究使用 MAP 調適法可以降低其複雜度，並且使第一階段 baseline 辨識效能有不錯的改善；第二階段使用經由調適後的 factored 語言模型及針對自發性語音所設計出來的韻律模型幫助辨認，藉由更多語言資訊和韻律停頓、韻律狀態資訊的加入，改善搶詞、音節合併等問題，並且解碼出各種語言及韻律資訊。

實驗結果顯示將階層式語音辨認系統應用在自發性語音中，使詞(word)、字(character)及音節(syllable)的辨識率分別改進了 4.43%、4.60%及 3.06%，對於韻律訊息的加入整體辨識率改善幅度有限，仔細探討其原因，發現這些改善比較在正常語流上，而對於不正常語流的辨識率仍是很低，這些特殊現象會有許多取代型錯誤(substitution)及刪除型錯誤(insertion)發生。另外，在聽過測試語料中辨識率較低的句子，發現該語者講話過於含糊且速度快，這些說話方式及語速的特性或許是辨識效能增加不顯著的主要原因。

6.2 未來展望

自發性語音普遍語速較快，造成許多音節合併及發音變異現象，若能將常發生音節合併和發音變異現象的聲音建立在聲學模型當中，也許可以改善這個問題；語言模型上，若能加入更多自發性文字語料，並針對不流暢現象加入 filler，更能模擬出適合的語言模型；而在韻律模型上，如果能把語速的考慮進去，偵測出不流暢的地方，並將 word lattice 切斷分不同模型做評分，整個辨認系統效能可以更好。而加入韻律信息於辨認系統中使得 word lattice 過於龐大、辨識時間過長，未來必須盡量縮小辨認時的搜尋路徑，或是將自發性語音應用在 WFST(Weighted Finite State Transducers)上。

參考文獻

【1】 M.Y. Tsai, F.C. Chou, and L.S. Lee, “Pronunciation modeling with reduced confusion for Mandarin Chinese using a three-stage framework,” *IEEE Transactions on Speech and Audio Processing*, Vol. 15, No. 2, pp. 661-675, Feb. 2006.

【2】 Y. Liu, and P. Fung, “State-Dependent Phonetic Tied Mixtures with Pronunciation Modeling for Spontaneous Speech Recognition,” IEEE Transactions on Speech and Audio Processing, Vol. 12, No. 4, pp. 351-364, July 2004.

【3】 V. Rangarajan, S, Narayanan, “Analysis of disfluent repetitions in spontaneous speech recognition,” 14th European Signal Processing Conference, Sep, 2006.

【4】 A. Zgank, M.S. Maucec, “Modelling of Filled Pauses and Onomatopoeias for Spontaneous Speech Recognition,” Advances in Speech Recognition, pp.164, Sep. 2010.

【5】 K. Ohta, M. Tsuchiya, S. Nakagawa, “Construction of Spoken Language Model Including Fillers Using Filler Prediction Model,” in proceeding of INTERSPEECH, 2007

【6】 T. Ng, M. Ostendorf, M.Y. Hwang, M. Siu, I. Bulyko, X. Lei, “Web-Data Augmented Language Models for Mandarin Conversational Speech Recognition,” in Proceedings of ICASSP, 2005, pp. 589-592.

【7】 許誌宏，“中文自發性語音辨識系統”，國立交通大學碩士論文，民國九十九年八月。

【8】 Stolcke, E. Shriberg, D. Hakkani-Tur , and G. Tur, “Modeling the prosody of hidden events for improved word recognition,” in Proc. of Eurospeech, 1999, pp. 311-314.

【9】 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.

【10】 K. Chen and 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.

【11】 李柏蒼，“自發性國語語音辨識”，國立交通大學碩士論文，民國九十六年八月。

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

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

【14】 吳聲鋒，“使用於中文自發性語音辨認之聲學模式及韻律模式”， 國立交通大學碩士論文，民國一百零三年八月。

【15】 蘇仲銘，“基於加權有限狀態轉換器國語語音辨認系統之設計”，國立交通大學碩士論文，民國一百零二年十一月。

【16】 Michiel Bacchiani, Brian Roark, “Unsupervised Language Model Adaptation,” in Proceedings of the IEEE ICASSP

【17】 Z. Sheng, J.-H. Tao, and D.-L. Jiang,“Chinese prosodic phrasing with extended features,” in Proceedings of the IEEE ICASSP 2003, Vol. 1, pp.492-495, 2008

【18】 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

【19】 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。

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

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

附錄一：詞類分類表

2 類詞類 (Level-0)			8 類詞類(Level-1)			23 類詞類(Level-2)			46 類詞類(Level-3)		
編號	中文詞類	代號	編號	中文詞類	代號	編號	中文詞類	代號	編號	中文詞類	代號
1	實詞	S	1	非謂形容詞	A	1	非謂形容詞	A	1	非謂形容詞	A
2	功能詞	F	2	連接詞	C	2	連接詞	C	2	對等連接詞	Caa
2	功能詞	F	2	連接詞	C	2	連接詞	C	3	連接詞，如：等等	Cab
2	功能詞	F	2	連接詞	C	2	連接詞	C	4	連接詞，如：的話	Cba
2	功能詞	F	2	連接詞	C	2	連接詞	C	5	關聯連接詞	Cbb
2	實詞	S	3	副詞	D	5	副詞	D	6	數量副詞	Da
2	實詞	S	3	副詞	D	3	動詞前程度副詞	Dfa	7	動詞前程度副詞	Dfa
2	實詞	S	3	副詞	D	4	動詞後程度副詞	Dfb	8	動詞後程度副詞	Dfb
2	實詞	S	3	副詞	D	5	副詞	D	9	時態標記	Di
2	實詞	S	3	副詞	D	5	副詞	D	10	句副詞	Dk
2	實詞	S	3	副詞	D	5	副詞	D	11	副詞	D
1	實詞	S	4	體詞	N	6	普通名詞	N	12	普通名詞	Na
1	實詞	S	4	體詞	N	6	普通名詞	N	13	專有名詞	Nb
1	實詞	S	4	體詞	N	6	普通名詞	N	14	地方詞	Nc
1	實詞	S	4	體詞	N	9	後置詞,位置詞	Ng	15	位置詞	Ncd
1	實詞	S	4	體詞	N	7	時間詞	Nd	16	時間詞	Nd
1	實詞	S	4	體詞	N	8	定詞,量詞	Ne	17	數詞定詞	Neu
1	實詞	S	4	體詞	N	8	定詞,量詞	Ne	18	特指定詞	Nes
1	實詞	S	4	體詞	N	8	定詞,量詞	Ne	19	指代定詞	Nep
1	實詞	S	4	體詞	N	8	定詞,量詞	Ne	20	數量定詞	Neqa
1	實詞	S	4	體詞	N	8	定詞,量詞	Ne	21	後置數量定詞	Neqb
1	實詞	S	4	體詞	N	8	定詞,量詞	Ne	22	量詞	Nf
1	實詞	S	4	體詞	N	9	後置詞,位置詞	Ng	23	後置詞	Ng
1	實詞	S	4	體詞	N	6	普通名詞	N	47		Nv
1	實詞	S	4	體詞	N	10	代名詞	Nh	24	代名詞	Nh
2	功能詞	F	5	感嘆、語助詞	T	12	感嘆、語助詞	T	25	感嘆詞	I
2	功能詞	F	6	介詞	P	11	介詞	P	26	介詞	P
2	功能詞	F	5	感嘆、語助詞	T	12	感嘆、語助詞	T	27	語助詞	T
1	實詞	S	7	動詞	V	13	不及物動詞	VA	28	動作不及物動詞	VA
1	實詞	S	7	動詞	V	13	不及物動詞	VA	29	動作使動動詞	VAC
1	實詞	S	7	動詞	V	14	及物動詞	VC	30	動作類及物動詞	VB

1	實詞	S	7	動詞	V	14	及物動詞	VC	31	動作及物動詞	VC
1	實詞	S	7	動詞	V	14	及物動詞	VC	32	動作接地方賓語動詞	VCL
1	實詞	S	7	動詞	V	14	及物動詞	VC	33	雙賓動詞	VD
1	實詞	S	7	動詞	V	14	及物動詞	VC	34	動作句賓動詞	VE
1	實詞	S	7	動詞	V	14	及物動詞	VC	35	動作謂賓動詞	VF
1	實詞	S	7	動詞	V	13	不及物動詞	VA	36	分類動詞	VG
1	實詞	S	7	動詞	V	15	狀態不及物動詞	VH	37	狀態不及物動詞	VH
1	實詞	S	7	動詞	V	15	狀態不及物動詞	VH	38	狀態使動動詞	VHC
1	實詞	S	7	動詞	V	15	狀態不及物動詞	VH	39	狀態類及物動詞	VI
1	實詞	S	7	動詞	V	14	及物動詞	VC	40	狀態及物動詞	VJ
1	實詞	S	7	動詞	V	14	及物動詞	VC	41	狀態句賓動詞	VK
1	實詞	S	7	動詞	V	14	及物動詞	VC	42	狀態謂賓動詞	VL
1	實詞	S	7	動詞	V	16	有	V_2	43	有	V_2
2	功能詞	F	8	的	DE	17	的，之，得	DE	44	的，之，得	DE
2	功能詞	F	9	是	SHI	18	是	SHI	45	是	SHI
1	實詞	S	11	外文	FW	20	外文標記	FW	46	外文標記	FW
1	實詞	S	10	定量複合詞	DM	19	定量複合詞	DM	58	定量複合詞	DM
4	Paralinguistic	ParaL	13	Paralinguistic	ParaL	24	Paralinguistic	ParaL	59	Paralinguistic	ParaL
5	Particle	Particle	14	Particle	Particle	25	Particle	Particle	60	Particle	Particle
5	Particle	Particle	14	Particle	Particle	25	Particle	Particle	61	Marker	Marker
6	無法辨認的語音	UnRec Spch	15	無法辨認的語音	UnRec Spch	26	無法辨認的語音	UnRec Spch	62	無法辨認的語音	UnRecSpch
6	無法辨認的語音	UnRec Spch	15	無法辨認的語音	UnRec Spch	26	無法辨認的語音	UnRec Spch	63	不確定音	uncertain
4	Paralinguistic	ParaL	13	Paralinguistic	ParaL	24	Paralinguistic	ParaL	64	noise	noise