國　立　交　通　大　學

電信工程學系碩士班

碩士論文

中文單詞之韻律模式研究

A Study on Prosodic Modeling for Isolated Mandarin Words

研 究 生：陳啟風

指導教授：王逸如　博士

中華民國九十六年七月

中文單詞之韻律模式研究

A Study on Prosodic Modeling for Isolated Mandarin Words

研 究 生：陳啟風　　Student： Chi-Feng Chen

指導教授：王逸如 博士　　Advisor： Dr. Yih-Ru Wang

國立交通大學

電信工程學系

碩士論文

A Thesis

Department of Communication Engineering
College of Electrical Engineering and Computer Science
National Chiao Tung University
In Partial Fulfillment of Requirements
For the Degree of
Master of Science
In Electrical Engineering

July, 2007
Hsinchu, Taiwan, Republic of China

中華民國九十六年七月

中文單詞之韻律模式研究

研究生：陳啟風　　指導教授：王逸如 博士

國立交通大學電信工程學系碩士班

中文摘要

在本文中，我們對中文單詞提出了以音節為基本單位的基頻軌跡及音節長度的韻律模型。在基頻軌跡模型中，我們考慮了聲調、音節在詞的位置以及前後音節連音的三種影響因素，並假設這些影響因素彼此獨立相加而組成音節基頻軌跡。在音節長度模型中，我們考慮了聲調、音節在詞的位置、基本音節以及前後音節連音的四種影響因素，我們同樣假設這些影響因素為彼此獨立且具加成性。我們使用一個含 107,936 個詞的單一女性語者的語料庫來評估所提方法是否有效，我們並用決策樹來分析音節長度如何受音節音素結構的影響，也用決策樹來分析音節間 pause 的長度和前後音節音素結構的關係，實驗結果顯示訓練後此兩韻律模型的影響因素都符合我們對中文韻律的知識。

A Study on Prosodic Modeling for Isolated Mandarin Words

Student： Chi-Feng Chen Advisor： Dr. Yih-Ru Wang

Department of Communication Engineering

National Chiao Tung University

Abstract

In this paper, syllable-based prosody modelings of pitch contour and syllable duration for isolated Mandarin words are proposed. In the syllable pitch contour model, three main affecting factors of tone, syllable position in word, and inter-syllable coarticulation are considered. These three affecting factors are assumed to be independent and additive. Similarly, in the syllable duration model, four affecting factors of tone, syllable position in word, base-syllable, and inter-syllable coarticulation are considered. We also assume that these affecting factors are independent and additive. A large single female-speaker speech database containing 107,936 words was used to evaluate the performance of the proposed methods. After well-training, the decision tree method was used to analyze the 411 affecting factors of base-syllable and to explore the relationship between inter-syllable pause duration and the nearby linguistic features. Experimental results showed that all these affecting factors conformed to our knowledge about Mandarin prosody.

誌謝

本篇論文的完成，首先要感謝陳信宏老師，在我就讀研究所的日子裡，陳老師以他豐富的學識不斷的教導及啟發我，使我在短短兩年來，對研究及學識有了最大進步；還有感謝我的指導教授王逸如老師，由於老師的細心態度，對研究專業或事務分析，能從我沒有考慮的問題，適時的提出及指正，讓我以後對任何事務的考慮更加周全，謝謝老師這兩年耐心的教導，使我成長。此外，我要感謝性獸學長，在我的研究的路上，給我一些新的想法和基礎的建設，讓我能不斷吸收新的資訊，使我研究順順利利的進行，還有坐在我後面的輝哥，當我程式或專業有疑惑時，一定細心的教導我，讓我受用無窮，而且輝哥你是個不錯的朋友，常常跟你聊天，讓我開心的過了這二年，還有智合學長，你是個讓人感覺很舒服的人，我的 HTK 及 LINUX 有你的指導，讓我進步不少唷！最後還有希群、阿德、BAKING，雖然你們人不常在實驗室，不過有了你們的歡笑，讓實驗室多了家的感覺；同時要感謝實驗室的同學：友駿、迷彩、大大、宏宇、小肚腩、小鄧、胤賢、小傅，跟你們在同一個實驗裡，讓我感到十分的快樂及自在。還有常常愛用斜音聊天的阿勇哥及 A 級愛好者士凡，有你們陪伴，讓我不覺得無聊。

最後，我要感謝我的雙親，在我求學的過程中，不斷給我支持和鼓勵。還有感謝我的女朋友雅萍，常常給我支持及陪伴，使我更有動力去完成本篇論文。僅將這本論文獻給所有愛我及幫助過我的人，沒有你們，也就沒有今天的我。

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究動機

隨著科技的蓬勃發展，人類越來越仰賴電腦來處理身邊各項事務，於是乎，電腦科技的發展已從原本的運算能力導向轉變為以溝通與訊息交換為主要研究目標；在這個過程中，早期的研究主要是致力於如何提供最有用，最有價值的資訊，資訊檢索系統、網路搜尋引擎、資料探勘技術應運而生，然而，資訊最終的目的是要提供給使用者，所以人與電腦間的溝通就顯得格外重要。

觀察人類最自然的溝通方式，不外乎聽與說：聽出正確的訊息(辨識)，說出要表達的話(合成)，為了讓這兩種表達方式，也能成為人機間的溝通模式，語音辨識和語音合成技術的研究與發展，扮演了舉足輕重的地位。

一套有聲電子書，要以語音的方式唸出書的內容，或是要以語音方式唸出我們所接收到的 e-mail，這樣一種可以將無限制的文句自動轉成語音的合成系統，稱為文句語音系統(Text-to-Speech system, TTS System)。本論文主要著重於由詞組語料研究韻律訊息，期能進一步了解連續語音的韻律訊息，能使合成聲音的自然流暢度更為提升。

1.2 研究方向

本論文之研究重點，在於設計各種韻律訊息產生的模型，包括音節的 pitch、duration、能量及音節間停頓長度，再分析由大型語料庫所估計得到的韻律模型參數，以更了解韻律訊息的產生機制，作為 TTS 系統合成詞的韻律訊息之用。

1.3 章節概要

本論文共分為七章：

第一章 緒論：介紹本論文之研究動機與方向。

第二章 大量詞語料庫介紹：說明如何求取語料庫中許多不同的語音特性，諸如音節位置、音高、能量、音節長度之類的參數

第三章 pitch 模型：說明如何設計 pitch 模型及預測音節的基頻軌跡。

第四章 duration 模型：說明如何設計 duration 模型及預測音節的長度。

第五章 能量模型：說明如何設計能量模型及預測音節的能量。

第六章 韻律訊息展示：說明如何設計及使用介面來表達韻律訊息資訊。

第七章 結論與未來展望。

第二章 大量詞語料庫介紹

為了訓練模型方便，我們事先求取出語料庫中許多不同的語音特性，諸如音節位置、音高、能量、音節長度之類的參數，本章即是探討如何求取這些資訊。

2.1 適用於訓練模型之大量詞語料庫條件

一個語料庫是否適用於訓練模型，癥結在於其是否擁有各種不同的合成單元。就中文而言，每個中文字對應一個音節(syllable)，音節有五種聲調(tone)的變化，中文字約有 12000 多字(character)，但如果只以發音來區分，總共大約只有 1300 種音節，如果再去除聲調的差別，則只有 411 種基本音節(base-syllable)。

一般認為，適合用於訓練模型的語料庫應具有『豐富語音』(phonetically rich)與『豐富韻律』(prosodically rich)兩個特性。所謂『豐富語音』是指具有各式各樣的音節連接方式;而『豐富韻律』則是指語料庫具有多種不同的韻律變化。

2.2 語料庫資料來源及錄製音檔

目前語料庫的文字部份來自於『文句分析器的辭典』,選擇的條件以聲調(tone)平衡為主要目的，總共有 4321 篇短文，在接下來的小節中，將介紹文字內容之萃取、統計及錄製音檔的規格。

2.2.1 文字內容之萃取及統計

首先我們由詞典抽取詞，格式如下：

公司 今年 因此 單位 他們

將其處理，標記聲碼及詞類 (part of speech, POS)，聲碼與實際發音之對照關係

請參照附錄一，『國語 411 基本音節總音表』，產生的文件格式如下，其中第一行為原始文字、第二行為音碼，第三行為詞序的編碼，第四行為詞類：

```
公 1380 201 12
司 1007 202 12
今 1270 201 16
年 2264 202 16
因 1269 201 40
此 3006 202 40
單 1125 201 14
位 4324 202 14
他 1019 201 37
們 5147 202 37
```

整個語料庫共包含 107936 個詞，共有 277218 個字，詳細的詞長及聲調(tone)分佈如下：

表 2.2.1-1:語料庫詞長分佈表

詞長	數量
二字詞	64872
三字詞	26026
四字詞	16062
五字詞	797
六字詞	124
七字詞	49
八字詞	6

表 2.2.1-2:語料庫聲調分佈表

聲調	數量
一聲	62349
二聲	69278
三聲	48904
四聲	94786
五聲	1901

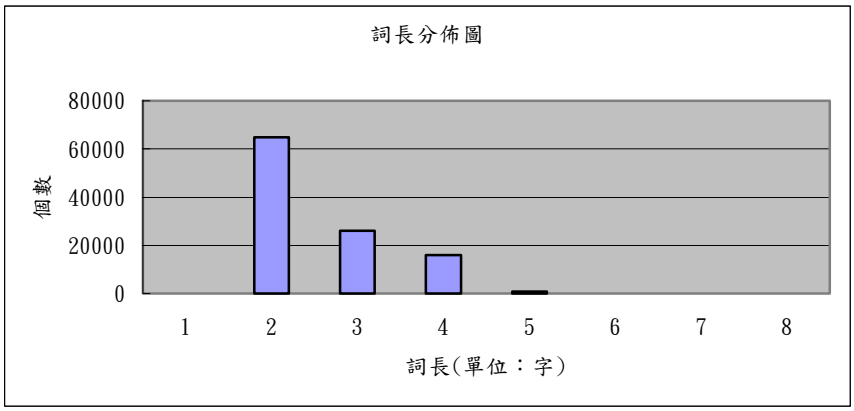

圖 2.2.1-1:語料庫詞長分佈圖

2.2.2 錄製音檔

產生如上所提的文字檔後，接下來要錄製音檔，我們將每個文字檔錄製成一個音檔，共計有 4321 個音檔，請專業的女性廣播人員以流利的方式唸出錄製語音。錄音軟硬體設備及格式詳如下表：

錄音軟體	Cool Edit Pro 直接錄成聲音檔案
麥克風	單一指向性
錄音場所	普通房間
錄音情境	依照所選出文稿唸出
取樣頻率	20 kHz
發音速度	每秒約 3.5 個音節
取樣大小	16 bits (位元)
聲道	單聲道 (mono)
檔案格式	Pcm

表 2.2.2-1 錄音軟硬體及格式表

2.3 語音參數資料庫建置

在建立了語料庫之文字資料庫與音檔後，為了之後訓練各種模型之用，我們需要更多的語音特性以供訓練各種模型之用，在此小節中，將介紹我們是如何由

原始的文字檔與音檔求取出各類語音特性參數。

2.3.1 切割資訊的求取

實際用於訓練模型的語料庫，需要更多的資訊，如標示每處音節所在位置的切割資訊，此節即是要說明如何產生此一訊息。

我們所使用的軟體為 HTK(Hidden Markov Model Toolkit)，而我們所採用的訓練模型方法在 HTK 說明手冊[2]中稱為 Isolated Word Style Training。此訓練使用參數的設定簡述如下：

關於參數的設定為：38 維的參數，包含 12 階的梅爾倒頻譜參數(Mel-frequency cepstral coefficients, MFCCs) 與能量對數值(log energy)，及其一階微分與二階微分，扣除原本的能量對數值後共 38 維;其音框大小(frame size)設為 32 ms;音框位移(frame rate)設為 10 ms。

其中有 406 個音檔由人工校正，剩餘的音檔利用過零率及能量，來修正切割資訊，使得切割資訊更加合理。

2.3.2 求取語料庫的能量資訊

在之後的訓練模型中，我們需要語料庫中更多的語音特性。在這一節中，說明能量大小的求取方法。首先，我們依照切割資訊將每段音節的波形取出，在此波形中，依照下式將每一個音框的能量求出。

$$E_x(m) = 2\log\sum_{n=1}^{N}\left|w_n \times f_x(n;m)\right|^2 \qquad (2.1)$$

m ：frame index

N ：total number of samples in a frame

w_n ：The n-th value of the Hamming window

$f_x(n;m)$ ：The magnitude of the n-th sample in the m-th frame

2.3.3 求取語料庫的音高(pitch)軌跡資訊

在語音的特性參數中，韻律訊息扮演了重要的角色，其中音高 (pitch) 的變化是我們所關注的一項議題，此節說明此資訊的求取方式。首先，我們利用 WaveSurfer[3]軟體所提供的 ESPS 演算法，求出每段音檔的音高軌跡，下圖為某個音檔的音高軌跡。

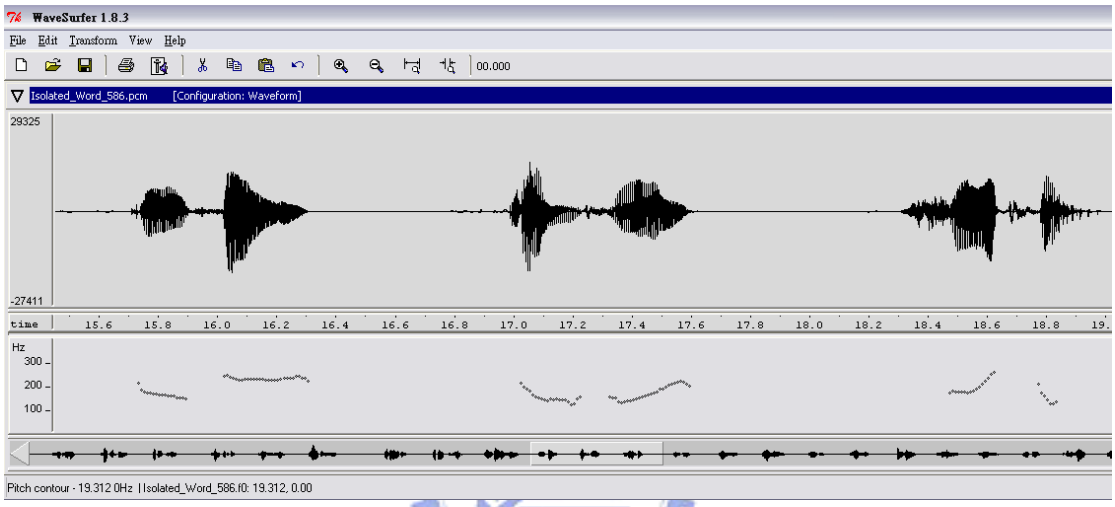

圖 2.3.3-1:利用 WaveSurfer 軟體求取音高的例子

我們將此音高軌跡藉由 WaveSurfer 軟體提供的功能存成文字檔。然而此資料是以音框為單位，每個音框有一個音高數據。為了將此資料轉變為每段音節一組數據，我們藉由前人所提出的轉換式[4]，將一段連續的音高軌跡轉化為四個正交參數表示，其數學式如下：

$$a_j = \frac{1}{N+1}\sum_{i=0}^{N} Pitch(i)\cdot \Phi_j\left(\frac{i}{N}\right) \quad \text{for } j = 0,\cdots,3 \tag{2.2}$$

其中，$Pitch(i)$為原始基頻軌跡，$0 \le i \le N$，N+1為基頻軌跡的長度。而$\Phi_j\left(\frac{i}{N}\right)$為正交化函數，定義如下：

$$\Phi_0\left(\frac{i}{N}\right) = 1 \tag{2.3}$$

$$\Phi_1\left(\frac{i}{N}\right)=\left[\frac{12\cdot N}{(N+2)}\right]^{\frac{1}{2}}\cdot\left[\left(\frac{i}{N}\right)-\frac{1}{2}\right] \tag{2.4}$$

$$\Phi_2\left(\frac{i}{N}\right)=\left[\frac{180\cdot N^3}{(N-2)(N+2)(N+3)}\right]^{\frac{1}{2}}\cdot\left[\left(\frac{i}{N}\right)^2-\left(\frac{i}{N}\right)+\frac{N-1}{6\cdot N}\right] \tag{2.5}$$

$$\Phi_3\left(\frac{i}{N}\right)=\left[\frac{2800N^5}{(N-1)(N-2)(N+2)(N+3)(N+4)}\right]^{\frac{1}{2}}\cdot\left[\left(\frac{i}{N}\right)^3-\frac{3}{2}\left(\frac{i}{N}\right)^2+\frac{6N^2-3N+2}{10N^2}\left(\frac{i}{N}\right)-\frac{(N-1)(N-2)}{20N^2}\right] \tag{2.6}$$

若知基頻軌跡參數，可以利用下式重建基頻軌跡。

$$Pitch'(i)=\sum_{j=0}^{3}p_j(k)\cdot\Phi\left(\frac{i}{N}\right),\quad \text{for } 0\le i\le N \tag{2.7}$$

藉由上述的轉換式，每段音節的音高變化皆被表示為四個正交參數，如此統一的格式就更容易被使用於訓練模型。

2.3.4 求取語料庫的音節長度(duration)資訊

在語音的特性參數中，音節長度關係到整個詞的流暢度，此節說明此資訊的求取方式。我們依照切割資訊，將每段音節的長度取得。如圖 2.3.4-1 所示：

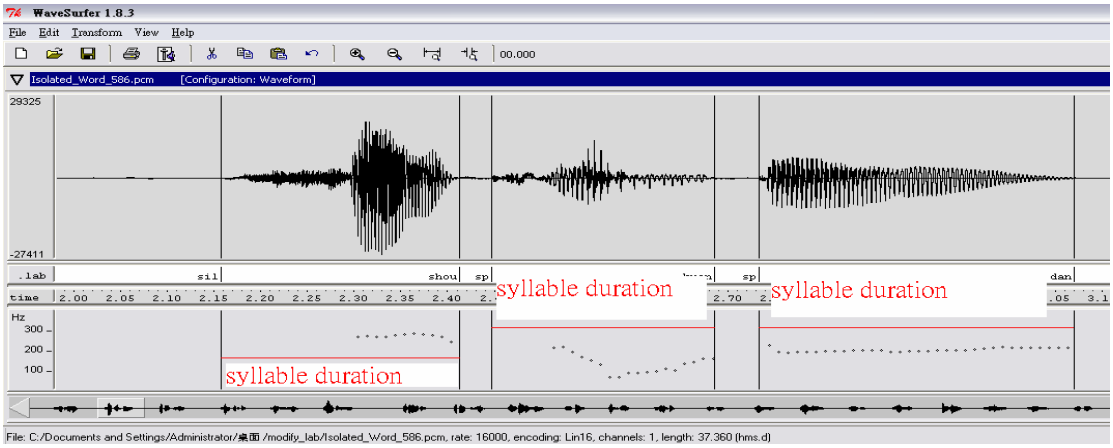

圖 2.3.4-1:由切割資訊獲得音節長度資訊

2.3.5 求取語料庫中的連音狀態(coarticulation state)資訊

所謂的連音狀態是指某個詞的兩個音節間相互影響的程度，直覺上連音狀態和前後音節互相影響有關，所以我們將連音狀態資訊也納入考慮。假設影響連音狀態的因素有二個(如下圖所示)，一為此音節的基頻軌跡與下一個音節的基頻軌跡間的長度(pitch pause)，若長度為零，則明顯看出音節間的影響最大；另一為音節與音節間最低點的能量 (energy-deep level)。

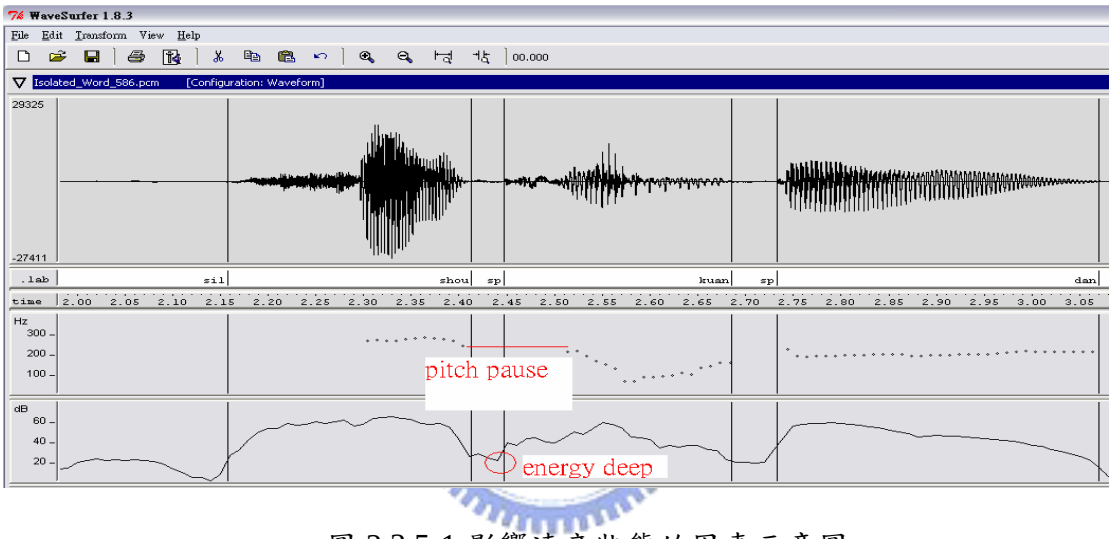

圖 2.3.5-1:影響連音狀態的因素示意圖

如何由音節間能量的最低點來作為判斷連音狀態的依據，我們由語料庫中統計出音節間能量最低點的分佈圖，如圖 2.3.5-2，很明顯可分成三群，若將 pitch pause 為零的族群去除後，剩下的可分成二群，其中 pitch pause 為零的族群分佈如圖 2.3.5-3，去除 pitch pause 為零的族群後分佈，如圖 2.3.5-4，由圖可知，能量值為 35dB 可做為連音狀態的分界點。

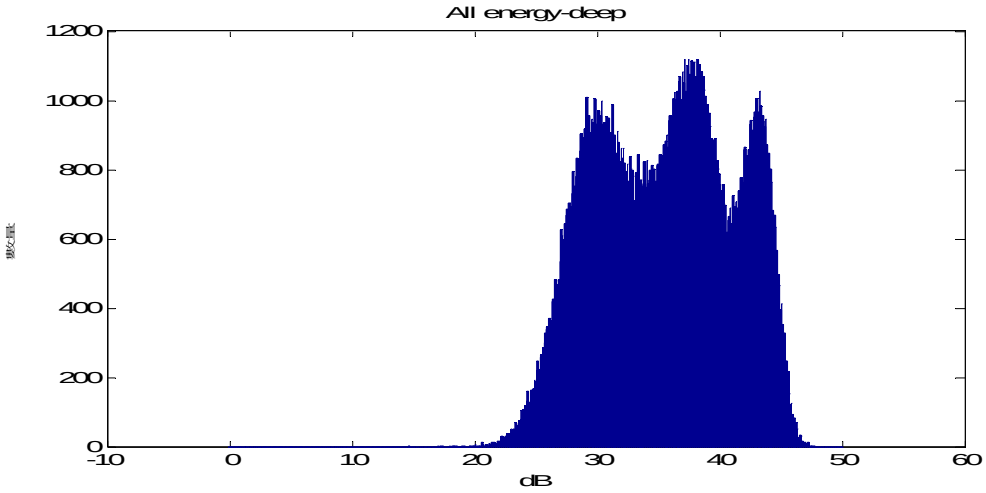

圖 2.3.5-2:語料庫上所有 energy-deep level 之分佈

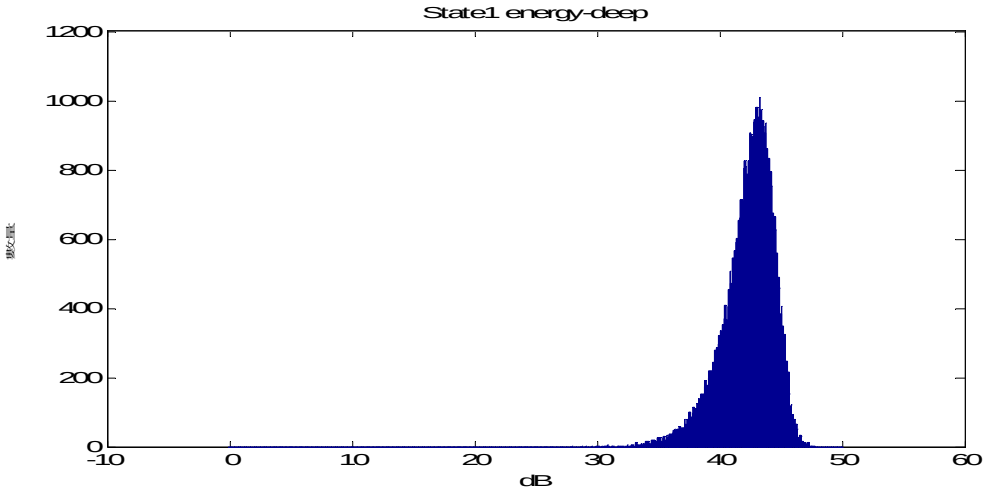

圖 2.3.5-3:將 pitch pause = 0 取出之 energy-deep level 分佈

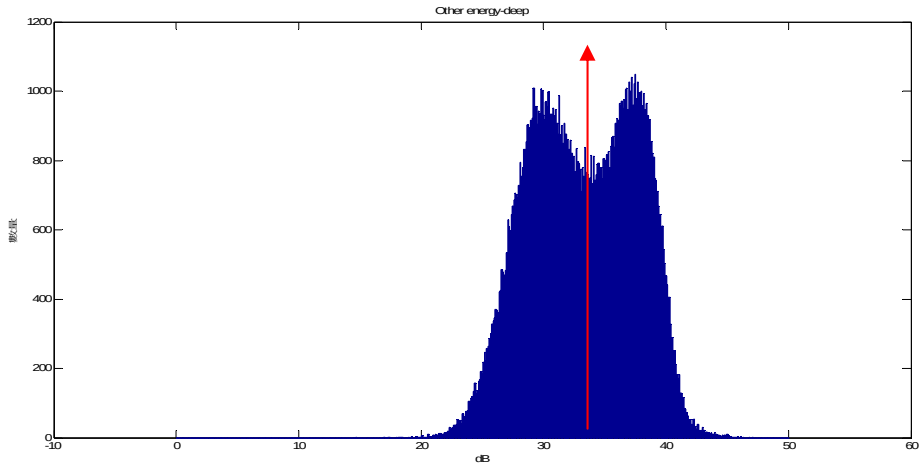

圖 2.3.5-4:pitch pause = 0 以外之 energy-deep level 分佈

為了使往後的分析較簡易，我們將上面所討論連音狀態的影響因素，做簡易的分類，目前將連音狀態分成三類，如圖 2.3.5-5：

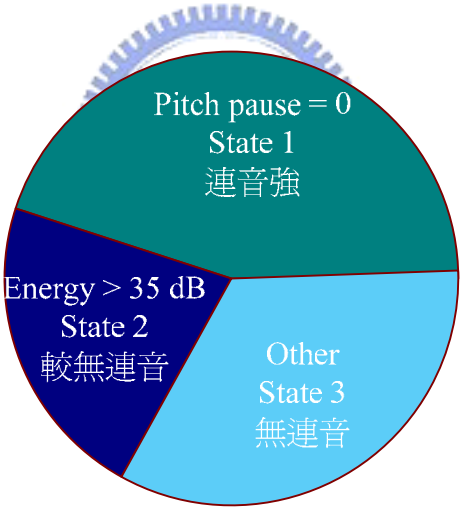

圖 2.3.5-5:連音狀態分類圖

2.3.6 音節聲母及韻母分類資訊

在我們的國語共有 411 個音節與 22 個聲母、39 個韻母，由觀察可得知某些聲母及韻母可分類成一類，依照我們分類，聲母分 6 類，韻母分 17 類，可增加分析上的便利性及運算上簡易。如下表 2.3.6-1 及表 2.3.6-2。在第四章 duration 模型及第五章能量模型皆參考下列分類，做為模型的韻母及聲母分類參考。

表 2.3.6-1:聲母分類表

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

表 2.3.6-2:韻母分類表

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

第三章 Pitch 模型

訓練 pitch 模型是為了在給予特定的資訊後，使之能預測音節基頻軌跡，能做為韻律訊息，方便未來 TTS 系統所使用，並希望更了解音節基頻軌跡被那些因素的影響程度較多。本章即是探討如何訓練 pitch 模型，及分析各種影響因素對音節基頻軌跡的影響程度。

3.1 設計 pitch 模型的方法

本論文所要設計的 pitch 模型主要有三個影響因素，分別為：聲調(tone)、音節在詞的位置(word position)、音節間的連音狀態(inter-syllable coarticulation state)。我們假設所有的影響因素可用累加的方式來表示，如下式：

$$\mathbf{sp}_n = \mathbf{sp}_n^r + \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{w_n} + \boldsymbol{\beta}^f_{c_{n-1},tp_{n-1}} + \boldsymbol{\beta}^b_{c_n,tp_n} + \boldsymbol{\mu}^p \tag{3.1}$$

其中各項向量變數定義如下，其向量為一段連續的音高軌跡轉化為四個正交參數表示，轉換方法詳述於[4]。

$\mathbf{sp}_n$：第 n 個音節的音高軌跡參數向量。

$\mathbf{sp}_n^r$：第 n 個音節的殘餘(residual)/正規化(normalized)音高軌跡參數向量。

$\boldsymbol{\beta}_{t_n}$：第 n 個音節的聲調影響音素，$t_n \in \{1,2,3,4,5\}$

$\boldsymbol{\beta}_{w_n}$：第 n 個音節的在詞的位置影響音素，$w_n \in \{(2,1)(2,2)(3,1)\ldots,(8,8)\}$

c_n：在第 n 個音節與第 n+1 個音節間的連音狀態，$c_n \in \{1,2,3\}$

tp_n：在第 n 個音節與第 n+1 個音節間的聲調配對(tone pair)。

$\boldsymbol{\beta}^b_{c_n,tp_n}$：給定的連音狀態及 tone pair 時，第 n 個音節受第 n+1 個音節的影響因素。

$\boldsymbol{\beta}^{f}_{c_{n-1},tp_{n-1}}$：給定的連音狀態及 tone pair 時，第 n 個音節受第 n-1 個音節的影響因素。

$\boldsymbol{\mu}^{p}$ ：整體音高軌跡參數的平均(global pitch mean)

在此研究，我們假設 $\mathbf{sp}^{r}_{n}$ 呈高斯分佈(Gaussian distribution)。圖 3.1-1 為音節的音高軌跡和主要的三個影響因素之關係的示意圖：

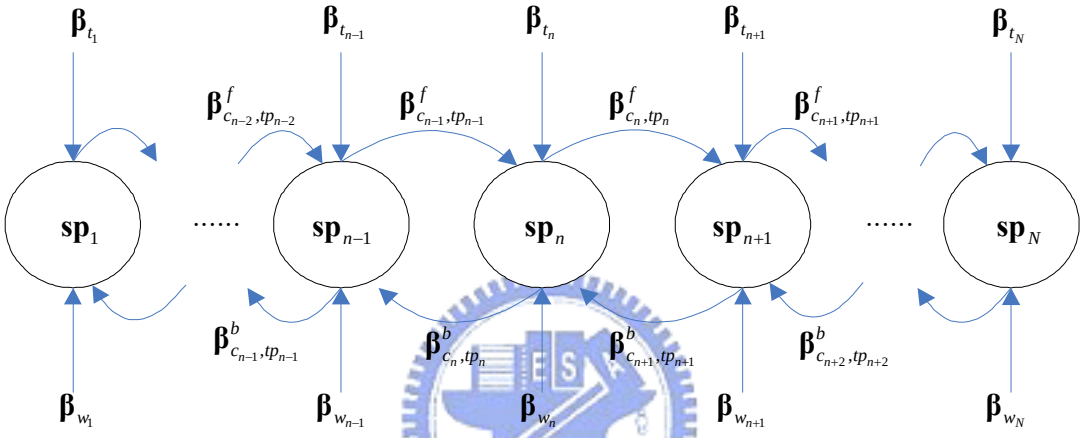

圖 3.1-1:音節音高軌跡與影響因素的示意圖

3.2 訓練 pitch 模型

3.2.1 推導各個影響因素

在給定所觀察音節的聲調、在詞的位置、與受前後音節的連音狀態及 tone pair，假設所觀察到每個音節的音高軌跡參數分佈滿足高斯分佈，其數學式如下式：

$$P(\mathbf{sp}_n \mid t_n, w_n, c_{n-1}, c_n, tp_{n-1}, tp_n) = N(\mathbf{sp}_n; \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{w_n} + \boldsymbol{\beta}^{f}_{c_{n-1},tp_{n-1}} + \boldsymbol{\beta}^{b}_{c_n,tp_n} + \boldsymbol{\mu}^{p}, \mathbf{R}^{p}) \quad (3.2)$$

利用最大概似法則(Maximum likelihood criterion)推導出各個影響因素之式子，整理如下：

$$\boldsymbol{\beta}_t = \frac{\sum_{n=1}^{N}\left(\mathbf{sp}_n - \boldsymbol{\beta}_{w_n} - \boldsymbol{\beta}^f_{c_{n-1},tp_{n-1}} - \boldsymbol{\beta}^b_{c_n,tp_n} - \boldsymbol{\mu}^p\right)\delta\left(t_n = t\right)}{\sum_{n=1}^{N}\delta\left(t_n = t\right)} \tag{3.3}$$

$$\boldsymbol{\beta}_w = \frac{\sum_{n=1}^{N}\left(\mathbf{sp}_n - \boldsymbol{\beta}_{t_n} - \boldsymbol{\beta}^f_{c_{n-1},tp_{n-1}} - \boldsymbol{\beta}^b_{c_n,tp_n} - \boldsymbol{\mu}^p\right)\delta\left(w_n = w\right)}{\sum_{n=1}^{N}\delta\left(w_n = w\right)} \tag{3.4}$$

$$\boldsymbol{\beta}^f_{c,tp} = \frac{\sum_{n=1}^{N}\left(\mathbf{sp}_n - \boldsymbol{\beta}_{t_n} - \boldsymbol{\beta}_{w_n} - \boldsymbol{\beta}^b_{c_n,tp_n} - \boldsymbol{\mu}^p\right)\delta\left(c_{n-1} = c\right)\delta\left(tp_{n-1} = tp\right)}{\sum_{n=1}^{N}\delta\left(c_{n-1} = c\right)\delta\left(tp_{n-1} = tp\right)} \tag{3.5}$$

$$\boldsymbol{\beta}^b_{c,tp} = \frac{\sum_{n=1}^{N}\left(\mathbf{sp}_n - \boldsymbol{\beta}_{t_n} - \boldsymbol{\beta}_{w_n} - \boldsymbol{\beta}^f_{c_{n-1},tp_{n-1}} - \boldsymbol{\mu}^p\right)\delta\left(c_n = c\right)\delta\left(tp_n = tp\right)}{\sum_{n=1}^{N}\delta\left(c_n = c\right)\delta\left(tp_n = tp\right)} \tag{3.6}$$

$$\mathbf{R}^p = \frac{\sum_{n=1}^{N}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}}{N} \tag{3.7}$$

其中 $\mathbf{Y} = \mathbf{sp}_n - \boldsymbol{\beta}_{t_n} - \boldsymbol{\beta}_{w_n} - \boldsymbol{\beta}^f_{c_{n-1},tp_{n-1}} - \boldsymbol{\beta}^b_{c_n,tp_n} - \boldsymbol{\mu}^p$，$\mathbf{R}^p$ 為 Covariance matrix。

3.2.2 各個影響因素的初始值設定

利用最大概似法則推導出各個影響因素，在訓練模型的角度來看，影響因素的初始值可能決定各個影響因素是否學習到該學習的地方，所以影響因素的初始值設定是個不可忽略的重要性。

在聲調影響因素的初始值設定中，我們先假設無其他的影響因素，可將(3.3)式簡略成下式：

$$\boldsymbol{\beta}_t = \frac{\sum_{n=1}^{N}\left(\mathbf{sp}_n - \boldsymbol{\mu}^p\right)\delta\left(t_n = t\right)}{\sum_{n=1}^{N}\delta\left(t_n = t\right)} \tag{3.8}$$

考慮受前音節的影響因素的初始值設定，在本篇論文是假設目前所考慮的音節聲調為 j 而前音節的聲調為 i，與前音接的連音狀態為 c，為了前音節影響因素的訓練不會學習到聲調 j 的影響因素，所以前音節影響因素扣除掉聲調 j 所造成的影響作為初始值設定，必能更加使前音節影響因素訓練更加準確，如(3.9)式，若考慮的音節聲調為 i 而後音節的聲調為 j，與後音節的連音狀態為 c，為了扣除掉聲調 i 所造成的影響作為初始值設定，如(3.10)式：

$$\boldsymbol{\beta}^f_{c,tp} = \frac{\sum_{n=1}^{N}\mathbf{sp}_n\delta\left(c_{n-1} = c\right)\delta\left(tp_{n-1} = tp\right)}{\sum_{n=1}^{N}\delta\left(c_{n-1} = c\right)\delta\left(tp_{n-1} = tp\right)} - \frac{\sum_{n=1}^{N}\mathbf{sp}_n\delta\left(c_{n-1} = c\right)\delta\left(t_n = j\right)}{\sum_{n=1}^{N}\delta\left(c_{n-1} = c\right)\delta\left(t_n = j\right)} \tag{3.9}$$

$$\boldsymbol{\beta}^b_{c,tp} = \frac{\sum_{n=1}^{N}\mathbf{sp}_n\delta\left(c_n = c\right)\delta\left(tp_n = tp\right)}{\sum_{n=1}^{N}\delta\left(c_n = c\right)\delta\left(tp_n = tp\right)} - \frac{\sum_{n=1}^{N}\mathbf{sp}_n\delta\left(c_n = c\right)\delta\left(t_n = i\right)}{\sum_{n=1}^{N}\delta\left(c_n = c\right)\delta\left(t_n = i\right)} \tag{3.10}$$

For c = 1~3 and tp=(i,j)

因為目前已經有了聲調、前音節(forward syllable)及後音節(backward syllable)的影響因素初始值，最後音節在詞的位置影響因素(word position)初始值設定可直接用(3.4)式。

3.2.3 訓練流程

訓練模型時，初始化後利用(3.3)式到(3.7)式，依序將聲調、音節在詞的位置、受前後音節等影響因素及 covariance matrix 的參數值更新，然後使用更新後的參數值，算出整個語料庫的最大概似法則分數，一直重覆更新參數值及分數，直到分數的變化量小於10^{-7}。訓練流程圖如圖 3.2.3-1：

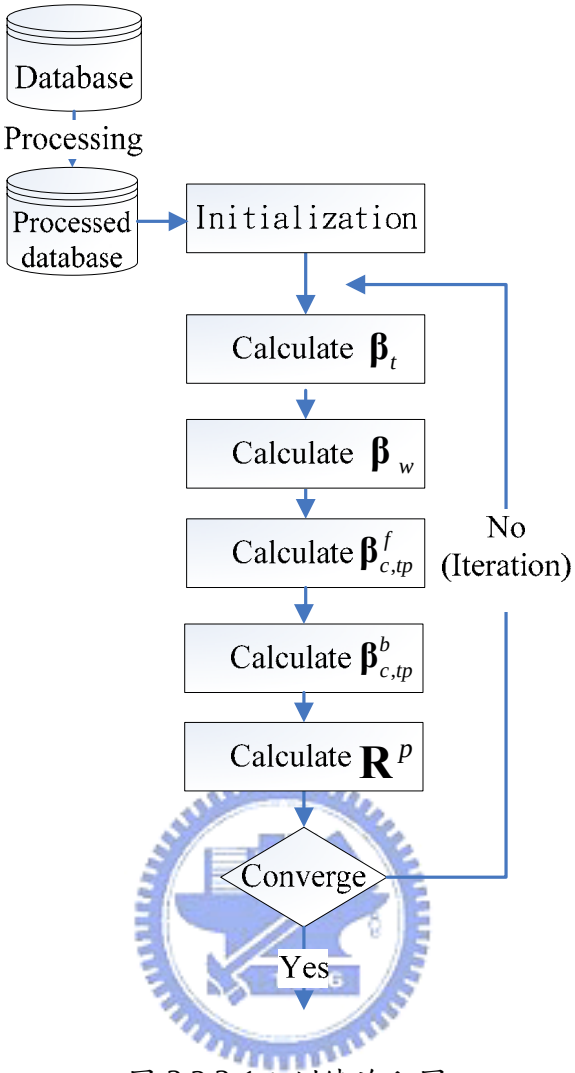

圖 3.2.3-1：訓練流程圖

3.3 訓練結果與分析

訓練模型後，使用 pitch 模型去預測基頻軌跡，我們進一步討論各個影響因素對於聲調的影響及貢獻。

3.3.1 利用 pitch 模型預測基頻軌跡

Pitch 模型預測基頻軌跡結果如下圖所示，其中黑色線(實線)為原始音節的基頻軌跡，紅色線(點線)為 pitch 模型所預測該音節的基頻軌跡，由圖可知大致的基頻軌跡走勢相似。

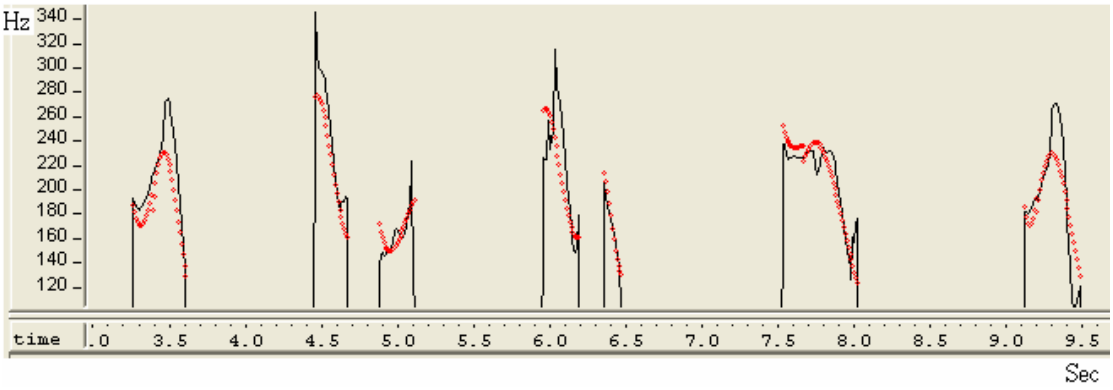

圖 3.3.1-1：Pitch 模型預測基頻軌跡

3.3.2 聲調影響因素(tone affecting factor)

聲調影響因素對音節的聲調由下圖所示，由圖觀察得，聲調影響因素的基頻軌跡較相似於國語聲調，所以音節聲調主要受聲調影響因素的影響。

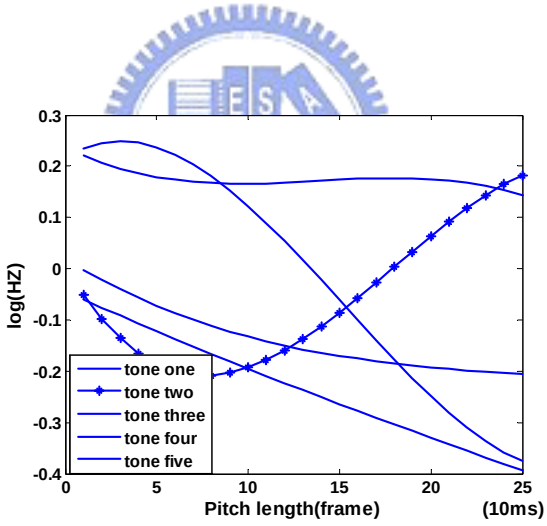

圖 3.3.2-1：聲調影響因素基頻軌跡

3.3.3 音節在詞的位置影響因素(word position affecting factor)

音節在詞的位置影響因素對音節的基頻軌跡影響如下列圖所示，分別有二字詞至五字詞位置影響因素對音節的基頻軌跡影響及將詞首與詞末的基頻軌跡整合比較，由圖 3.3.3-1 至圖 3.3.3-4 可觀察得詞的位置會直接影響基頻軌跡的平均基頻高度，愈接近詞首愈高，愈接近詞末愈低，且詞首的後半部音節軌跡會有上

仰趨勢，是為了接下個音節基頻軌跡而準備，也可由圖 3.3.3-5 所知，詞的字數愈多，詞首的平均基頻高度愈高且上仰趨勢愈大。

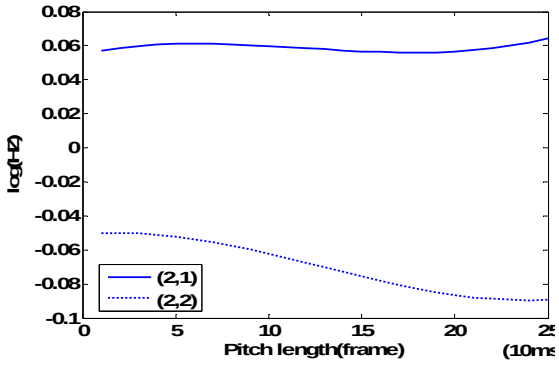

圖 3.3.3-1：二字詞影響因素基頻軌跡

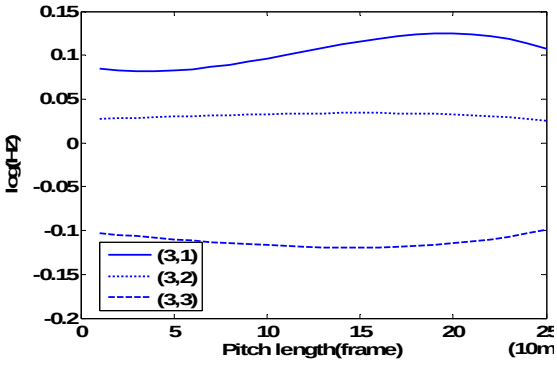

圖 3.3.3-2：三字詞影響因素基頻軌跡

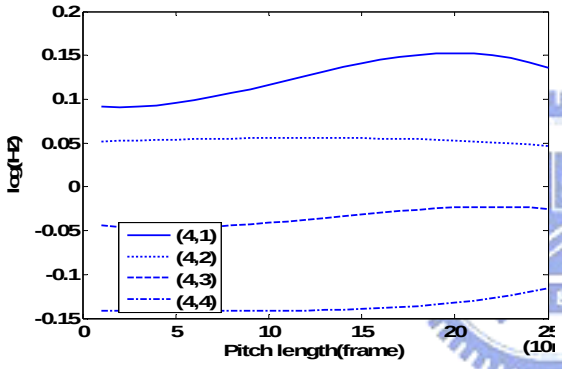

圖 3.3.3-3：四字詞影響因素基頻軌跡

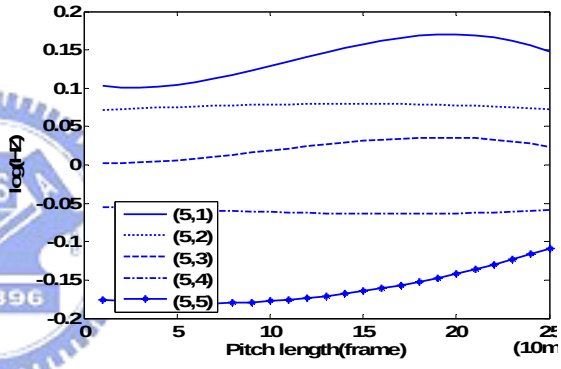

圖 3.3.3-4：五字詞影響因素基頻軌跡

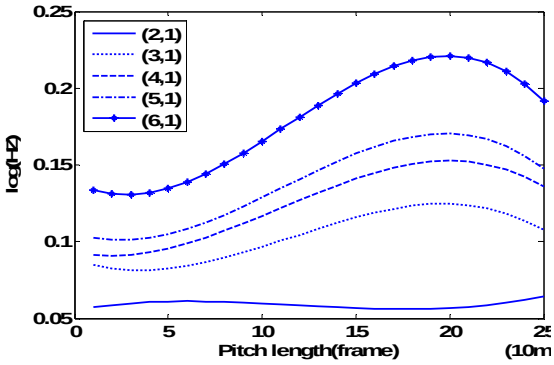

圖 3.3.3-5：整合詞首基頻軌跡

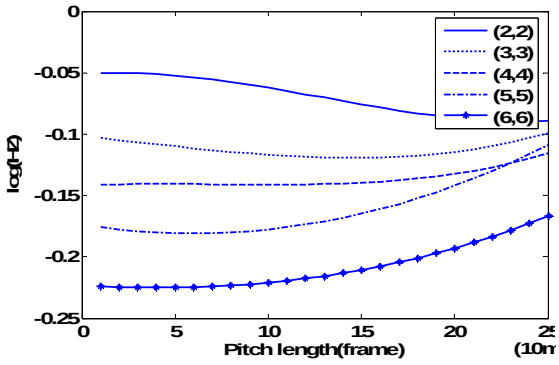

圖 3.3.3-6：整合詞末基頻軌跡

3.3.4 受前面音節影響因素(forward affecting factor)

為方便觀察受前音節影響的基頻軌跡，單純假設圖 3.3.4-1 的基頻軌跡音節為二字詞字末音節，僅受前音節的影響，可以看出五個聲調受前音節各個聲調在各個連音狀態的影響程度。由圖 3.3.4-1 觀察得 state1 基頻軌跡前半部受前音節影響改變量較大，且 state2 與 state3 較不受影響，與軌跡相似。

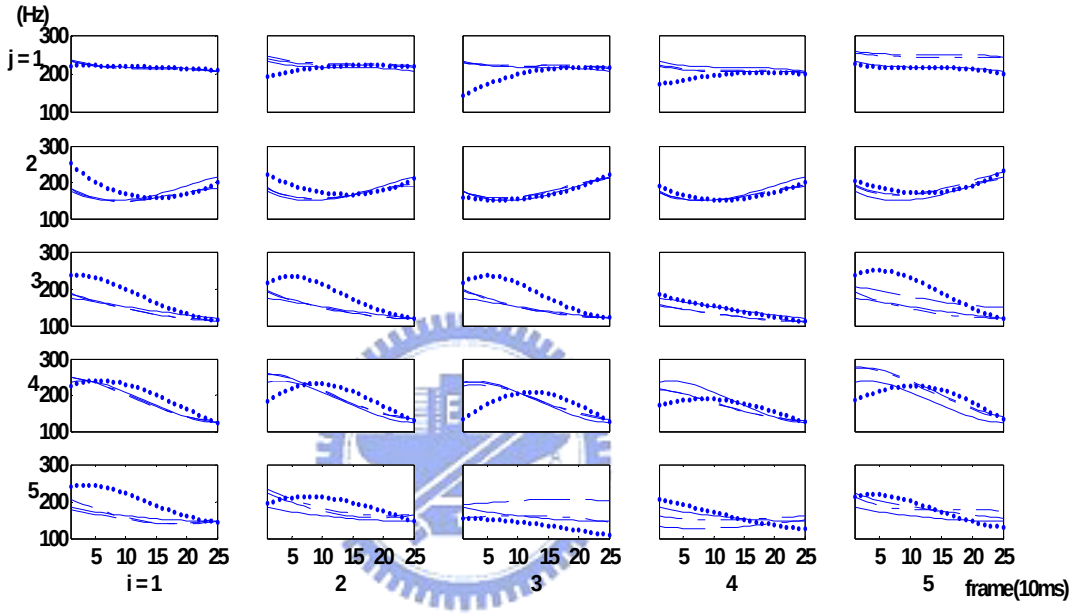

圖 3.3.4-1：聲調受前面音節影響基頻軌跡變化圖，其中實線為不加前後影響因素、點線(…)為 state1、點虛線(--・)為 state2、虛線(--)為 state3，其 j 為音節的聲調受前音節聲調 i 的影響

3.3.5 受後面音節影響因素(backward affecting factor)

為方便觀察受前音節影響的基頻軌跡，單純假設圖 3.3.5-1 的基頻軌跡音節為二字詞字首音節，僅受後音節的影響。可以看出五個聲調受後音節各個聲調在各個連音狀態的影響程度。由圖 3.3.5-1 觀察得 state1 基頻軌跡後半部由後面音節影響改變較大，且比較圖 3.3.4-1 可明確了解受前面音節的影響大於受後面音節的影響。

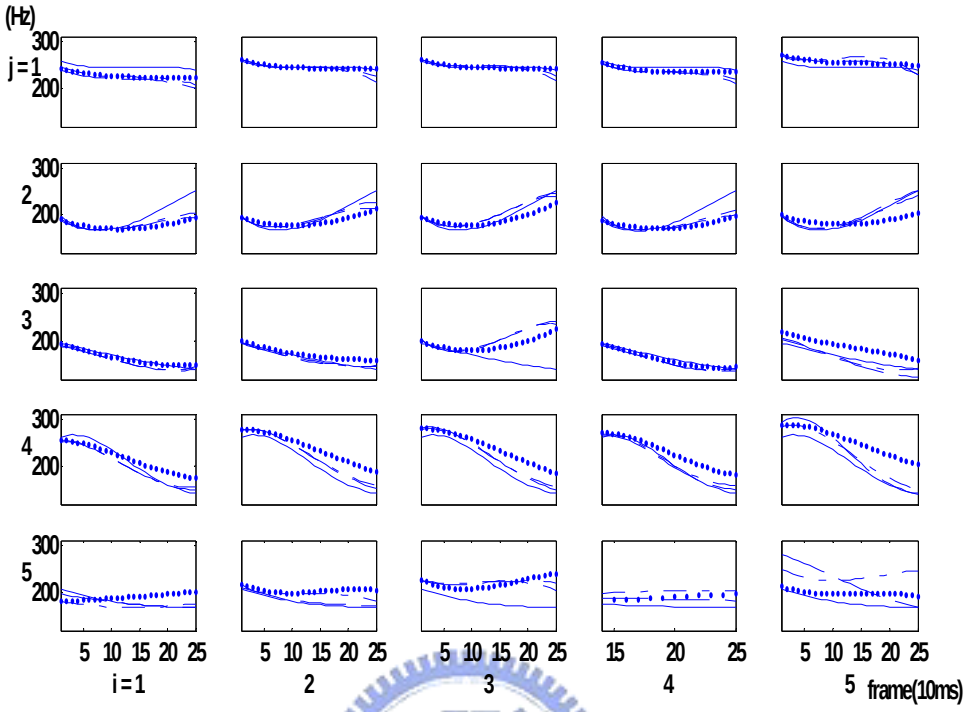

圖 3.3.5-1：聲調受後音節影響基頻軌跡變化圖，其中實線為不加前後影響因素、點線(…)為 state1、點虛線(--・)為 state2、虛線(--)為 state3，其 j 為音節的聲調受後音節聲調 i 的影響

3.3.6 二字詞基頻軌跡預測

圖 3.3.6-1 為各個 tone pair 的基頻軌跡預測結果，假設圖的二個音節基頻軌跡相連，可觀察得各個 tone pair 的預測結果。其中 state1 明顯觀察得易受影響，也可證實三聲接三聲分變成二聲接三聲的語言特性。

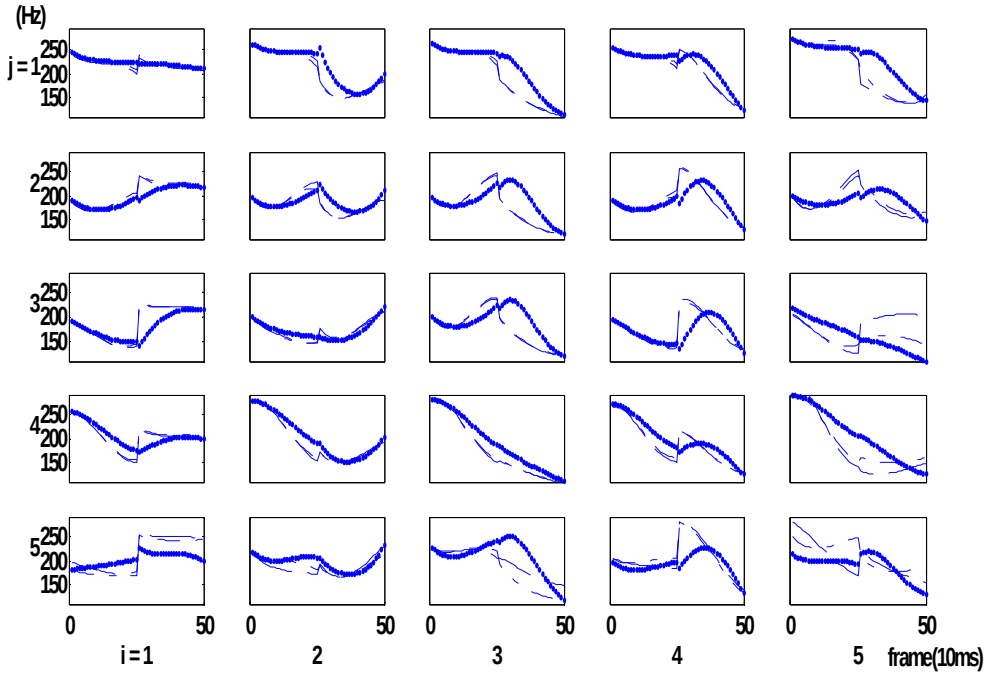

圖 3.3.6-1：二字詞基頻軌跡預測，其中點線(…)為 state1、點虛線(--・)為 state2、虛線(--)為 state3，其 j 為第一音節，i 為第二個音節

第四章 Duration 模型

訓練 duration 模型是為了在給予特定的資訊後，使之能預測音節長度，能做為韻律訊息，方便未來 TTS 系統使用，可以了解音節長度被那些影響因素的影響程度較多。本章即是探討如何訓練 duration 模型，及分析各種影響因素對音節長度的影響程度。

4.1 設計 duration 模型的方法

本論文所要設計的 duration 模型主要有四個影響因素，分別為：聲調、音節在詞的位置、基本音節種類(base syllable)、音節間的連音狀態。我們假設所有的影響因素可用累加的方式來表示，如下式：

$$sd_n = sd_n^r + \gamma_{t_n} + \gamma_{w_n} + \gamma_{sy_n} + \gamma^f_{c_{n-1}, fi\_in_{n-1}} + \gamma^b_{c_n, fi\_in_n} + \mu^d \tag{4.1}$$

其中各項變數定義如下。

sd_n：第 n 個音節的長度。

sd_n^r：第 n 個音節的長度殘餘 (residual)。

γ_{t_n}：第 n 個音節的聲調影響音素，$t_n \in \{1,2,3,4,5\}$

γ_{w_n}：第 n 個音節在詞的位置調影響音素，$w_n \in \{(2,1)(2,2)(3,1)\ldots,(8,8)\}$

γ_{sy_n}：第 n 個音節的音節影響因素，$sy_n \in \{1,2\ldots411\}$

c_n：在第 n 個音節與第 n+1 個音節的連音狀態，$c_n \in \{1,2,3\}$

fi_in_n：第 n 個音節的韻母與第 n+1 個音節的聲母組合。

$\gamma^b_{c_n, fi\_in_n}$： 第 n 個音節受第 n+1 個音節影響的影響因素。

$\gamma^{f}_{c_{n-1},fi\_in_{n-1}}$：第 n 個音節受第 n-1 個音節影響的影響因素。

μ^{d}：整體長度平均。

在此研究，我們假設 sd_n^r 呈高斯分佈。圖 4.1-1 為每個音節長度和主要的四個影響因素之關係的示意圖：

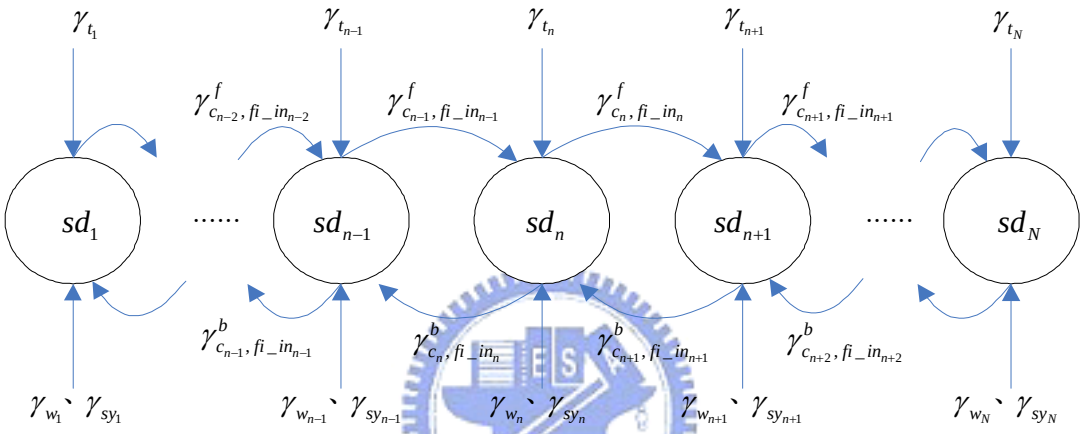

圖 4.1-1 音節長度與影響因素的示意圖

4.2 訓練 duration 模型

4.2.1 推導各個影響因素

給定觀察音節的聲調、在詞的位置、基本音節與受後前音節的連音狀態及音節間的韻母及聲母組合，則觀察到每個音節的長度分佈滿足下式：

$$P(sd_n \mid t_n, w_n, sy_n, c_{n-1}, c_n, fi\_in_{n-1}, fi\_in_n) = N(sd_n; \gamma_{t_n} + \gamma_{w_n} + \gamma_{sy_n} + \gamma^{f}_{c_{n-1},fi\_in_{n-1}} + \gamma^{b}_{c_n,fi\_in_n} + \mu^{d}, R^{d}) \quad (4.2)$$

利用最大概似法則推導出各個影響因素之式子，整理如下：

$$\gamma_t = \frac{\sum_{n=1}^{N}\left(sd_n - \gamma_{w_n} - \gamma_{sy_n} - \gamma^f_{c_{n-1},fi\_in_{n-1}} - \gamma^b_{c_n,fi\_in_n} - \mu^d\right)\delta\left(t_n = t\right)}{\sum_{n=1}^{N}\delta\left(t_n = t\right)} \tag{4.3}$$

$$\gamma_w = \frac{\sum_{n=1}^{N}\left(sd_n - \gamma_{t_n} - \gamma_{sy_n} - \gamma^f_{c_{n-1},fi\_in_{n-1}} - \gamma^b_{c_n,fi\_in_n} - \mu^d\right)\delta\left(w_n = w\right)}{\sum_{n=1}^{N}\delta\left(w_n = w\right)} \tag{4.4}$$

$$\gamma_{sy} = \frac{\sum_{n=1}^{N}\left(sd_n - \gamma_{t_n} - \gamma_{w_n} - \gamma^f_{c_{n-1},fi\_in_{n-1}} - \gamma^b_{c_n,fi\_in_n} - \mu^d\right)\delta\left(sy_n = sy\right)}{\sum_{n=1}^{N}\delta\left(sy_n = sy\right)} \tag{4.5}$$

$$\gamma^f_{c,fi\_in} = \frac{\sum_{n=1}^{N}\left(sd_n - \gamma_{t_n} - \gamma_{w_n} - \gamma_{sy_n} - \gamma^b_{c_n,fi\_in_n} - \mu^d\right)\delta\left(c_{n-1} = c\right)\delta\left(fi\_in_{n-1} = fi\_in\right)}{\sum_{n=1}^{N}\delta\left(c_{n-1} = c\right)\delta\left(fi\_in_{n-1} = fi\_in\right)} \tag{4.6}$$

$$\gamma^b_{c,fi\_in} = \frac{\sum_{n=1}^{N}\left(sd_n - \gamma_{t_n} - \gamma_{w_n} - \gamma_{sy_n} - \gamma^f_{c_{n-1},fi\_in_{n-1}} - \mu^d\right)\delta\left(c_n = c\right)\delta\left(fi\_in_n = fi\_in\right)}{\sum_{n=1}^{N}\delta\left(c_n = c\right)\delta\left(fi\_in_n = fi\_in\right)} \tag{4.7}$$

$$R^d = \frac{\sum_{n=1}^{N}\left(sd_n - \gamma_{t_n} - \gamma_{w_n} - \gamma_{sy_n} - \gamma^f_{c_{n-1},fi\_in_{n-1}} - \gamma^b_{c_n,fi\_in_n} - \mu^d\right)^2}{N} \tag{4.8}$$

其中，R^d 為 covariance matrix。本實驗中將所有影響因素的初始值全部預設為 0，由訓練流程更新參數值。

4.2.2 訓練流程

訓練模型時，初始化後依序將聲調、詞的位置、基本音節、受前後音節等影響因素及 covariance matrix 的參數值更新，然後使用更新後的參數值，算出整個語料庫的最大概似法則分數，一直重覆更新參數值及分數，直到分數的變化量小於10^{-7}。訓練流程圖如圖 4.2.2-1：

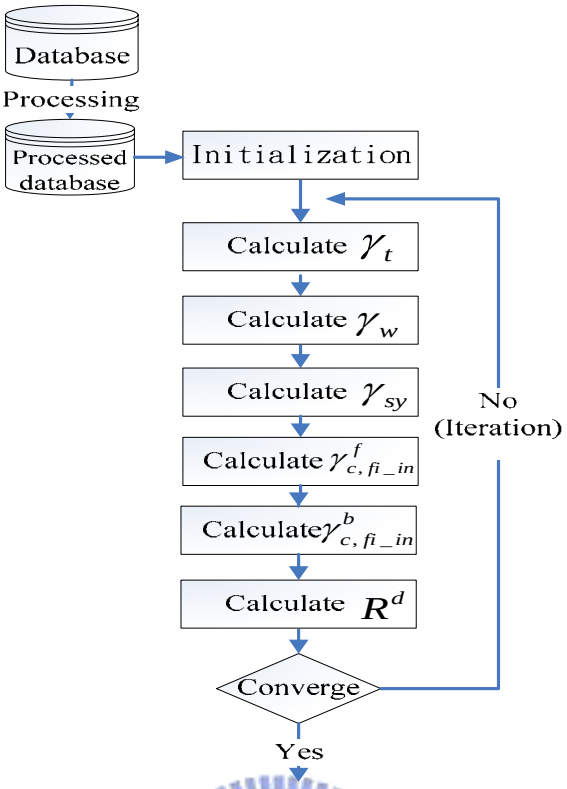

圖 4.2.2-1：訓練流程圖

4.3 決策樹分析音節長度

由語料庫的統計，平均音節長度為 283 ms，及詞裡面的音節間長度(pause)的個數為 167,635 個。下一個小節，利用決策樹，將語料庫的音節長度，音節間長度及連音狀態做適當的分類，分類的結果做為產生 pause 訊息的依據。

4.3.1 決策樹分裂依據

圖 4.3.1-1 為決策樹分裂條件示意圖，分裂條件需滿足分裂後的 likelihood 分數總和(L_1+L_2)比分裂前的 likelihood 分數(L)多 α 分，如(4.9)式，而且分裂後的群組成員至少有 β 個，如式，(4.10)式，當滿足分裂的問題，則該問題應從決策樹的問題集(question set)裡移除，當不滿足分裂條件任一條，或者問題集沒有任何問題時，則停止分裂。Question set 陳列如表 4.3.1-1。

$$L_1 + L_2 - L > \alpha \tag{4.9}$$

$$N_1 > \beta, N_2 > \beta \tag{4.10}$$

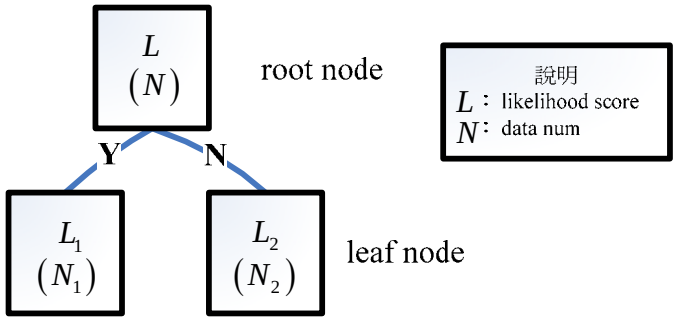

圖 4.3.1-1：決策樹分裂條件示意圖

表 4.3.1-1：決策樹問題集

問題項目	問題	問題項目	問題
Q1	NULL initial	Q6	Initial in {j,z,zh}
Q2	Initial in {b,d,g}	Q7	Initial in {p,t,k}
Q3	Initial in {f,s,sh,x,h}	Q8	Single vowel
Q4	Initial in {m,n,l,r}	Q9	Compound vowel
Q5	Initial in {c,ch,q}	Q10	Nasal ending

4.3.2 以決策樹分類語料庫連音狀態

語料庫的連音狀態分類如圖 2.3.5-5 所示，應用於問題集中的 initial 和 final 為此連音狀態相鄰的 initial 及 final。連音狀態分類做為決策樹分類的目標，由語料庫音節間的連音狀態設為定值，state 1 設值為 1，state 2 設值為 2，state 3 設值為 3。將這些 pause 對應的值當成決策樹的分類目標，所以在圖 4.3.2-1 呈現連音狀態的分類結果，其中 mean = 1.047 意指分佈較集中於 state 1，mean=1.1173 也較集中於 state 1，mean=2.1341 是指分佈較集中於 state 2。

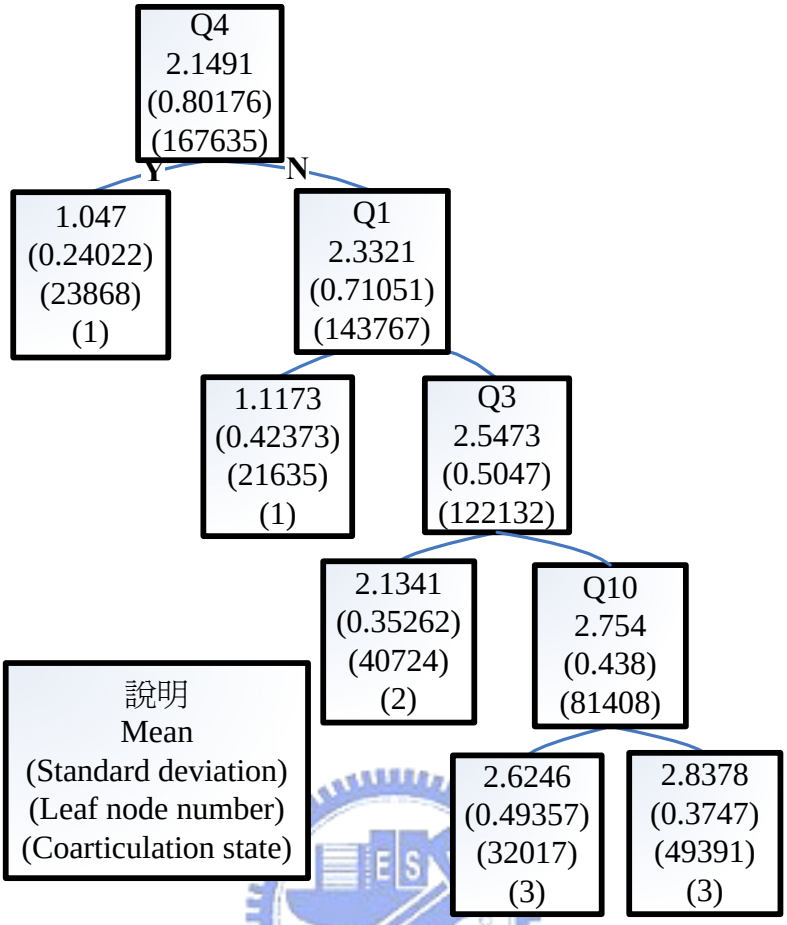

圖 4.3.2-1：以決策樹分析連音狀態的分類結果

由圖 4.3.2-1 可觀察得 NULL initial(Q1)和 initial 為{m,n,l,r}(Q4)這類歸為 state 1，而 Initial 為{f,s,sh,x,h}(Q3)這類歸為 state 2，除了上述二種狀況外以鼻音結束的音節(Q10)，雖 mean 為 2.6246 剛好處於 2(state 2)與 3(state 3)之間，因為在 4.3.3 小節中了解到可以利用音節的 initial 來當成判斷 pause 長度的依據，所以目前先假設除了 Q1、Q3、Q4 外，其餘設為 state3。

4.3.3 以決策樹分析語料庫 pause 長度

應用於問題集中的 initial 和 final 為 pause 相鄰的 initial 及 final。Pause 長度做為決策樹分類的目標，由語料庫中 pause 的長度視為目標值，其單位為毫秒(ms)。 其中圖 4.3.3-1 為以決策樹分析語料庫 pause 長度的分類結果。

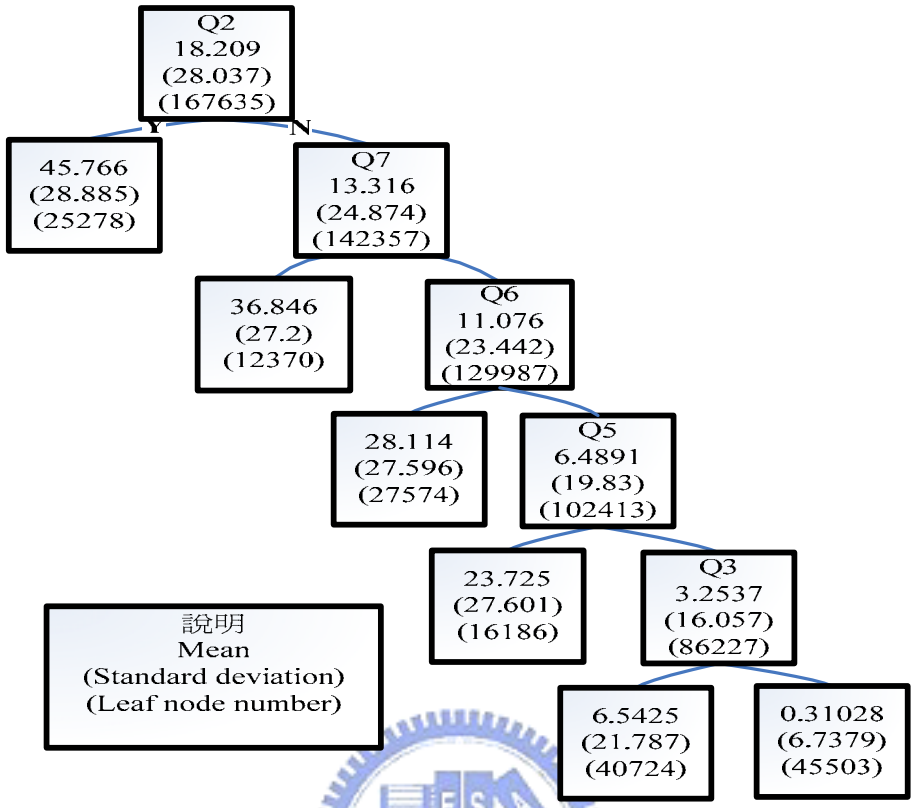

圖 4.3.3-1：以決策樹分析語料庫 pause 長度的分類結果

圖 4.3.3-1 可觀察得 Pause 長度分類明顯與 pause 相鄰的 initial 有關，與 final 相關性不大。我們將 Pause 長度與 initial 關係整理如下：

- 爆破音_不送氣 > 爆破音_送氣 > 塞擦音_不送氣 > 塞擦音_送氣 >
- 摩擦音_清音 > m, n, l, r 及空聲母

4.3.4 整合 pause 長度

整合 4.3.2 及 4.3.3 章節，由圖 4.3.2-1 得知音節的 initial 為 INULL 或{m,n,l,r}這類，被歸為 state 1，且由圖 4.3.3-1 得知此分類長度接近於 0，我們假設此分類 pause 長度為 0。而 state 2 及 state 3 均可由 initial 類別決定，藉著圖 4.3.3-1 查詢該 initial 類別的 mean 而決定 pause 長度。選取 pause 長度整理如表 4.3.4-1：

表 4.3.4-1：pause 長度選取表

類別	聲母	Pause 長度
1	ㄇ、ㄋ、ㄌ、ㄖ、空聲母 (鼻音_濁音)	0 ms
2	ㄏ、ㄒ、ㄕ、ㄈ、ㄙ (摩擦音_清音)	7 ms
3	ㄅ、ㄉ、ㄍ (爆破音_不送氣)	46 ms
4	ㄐ、ㄓ、ㄗ (塞擦音_不送氣)	28 ms
5	ㄆ、ㄊ、ㄎ (爆破音_送氣)	37 ms
6	ㄑ、ㄔ、ㄘ (塞擦音_送氣)	24 ms

4.4 訓練結果與分析

訓練模型後，使用 duration 模型去預測音節長度，我們更進一步去討論各個影響因素對於長度的影響及貢獻。

4.4.1 利用 duration 模型預測音節長度

圖 4.4.1-1 為語料庫音檔部份音節長度與 duration 模型預測的長度做比較，其中 training data 為音節實際長度，而 duration model 為預測長度。

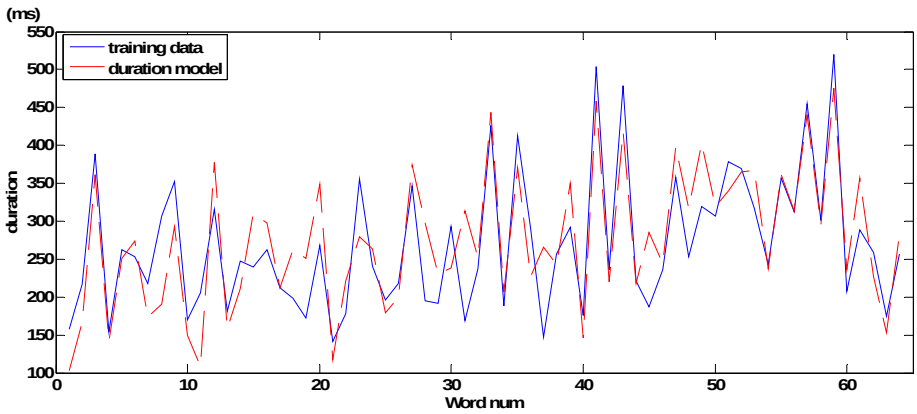

圖 4.4.1-1：音節實際長度與模型預測長度之比較圖(1)

4.4.2 聲調影響因素(tone affecting factor)

聲調對長度影響程度如圖 4.4.2-1，由圖形可知五聲音節最短，二聲音節最長。

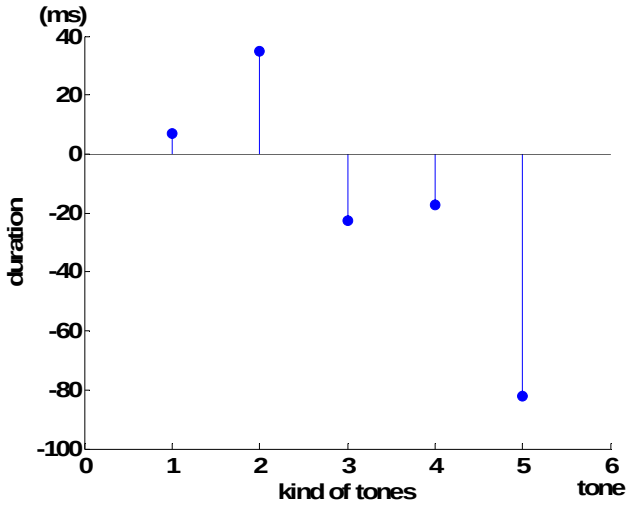

圖 4.4.2-1：聲調影響因素

4.4.3 詞的位置影響因素(word position affecting factor)

詞的位置對長度影響程度如圖 4.4.3-1，由圖可知在詞的最後的音節較長，圖的五字詞、六字詞、七字詞可觀察得詞愈長愈容易產生愈短的音節。

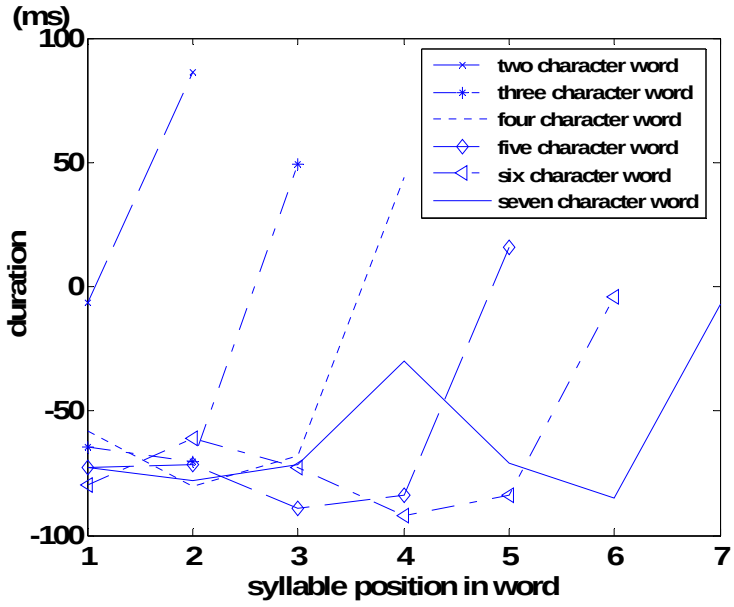

圖 4.4.3-1：詞的位置影響因素

4.4.4 音節影響因素(syllable affecting factor)

使用決策樹觀察音節影響因素對長度影響程度，如圖 4.4.4-1，其中問題集的 initial 和 final 為該音節的 initial 及 final。由決策樹可觀察出 initial 為 INULL (Q1)，{b、d、g}(Q2)或 final 為單母音(Q8)這類的音節長度較短，而鼻音結尾及{f,s,sh,x,h}(Q3)、{c,ch,q}(Q5)這類摩擦音音節長度會較長，這符合我們對音節長度的一般認知。

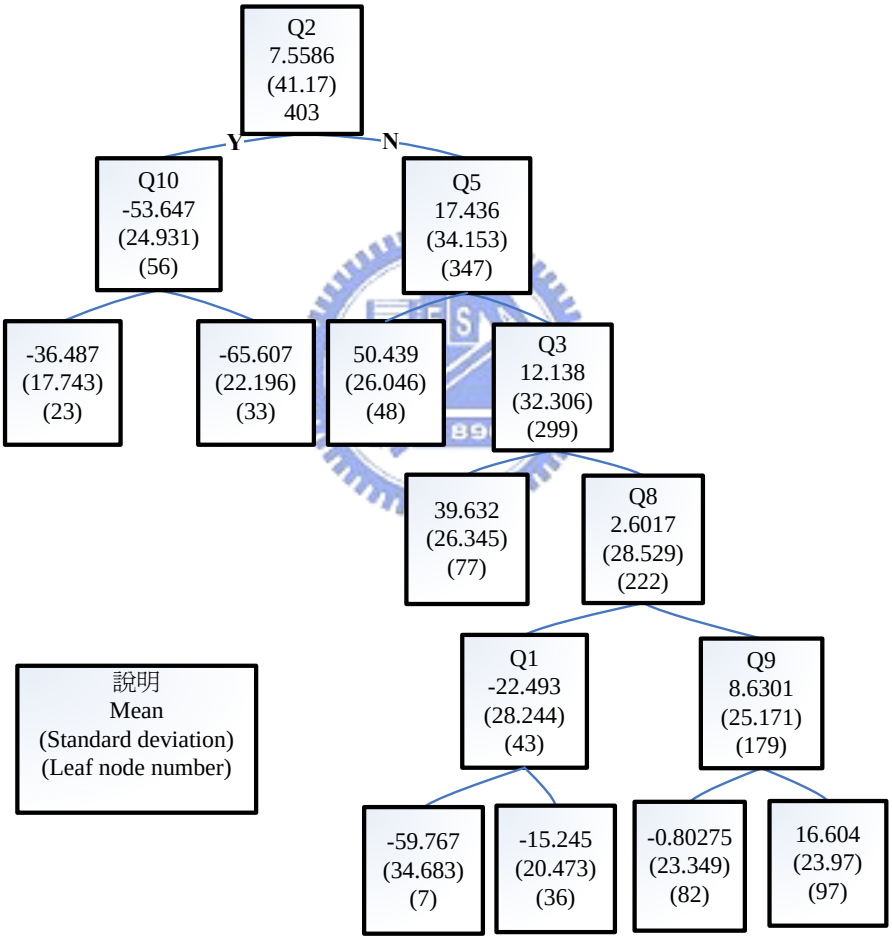

圖 4.4.4-1：以決策樹分析音節影響因素分類結果

第五章 能量模型

訓練能量模型是為了在給予特定的資訊後，使之能預測音節能量軌跡，做為韻律訊息，方便未來 TTS 系統所使用，可以了解音節能量軌跡被那些影響因素的影響程度較多。本章即是探討如何訓練能量模型，及分析各種影響因素對音節能量軌跡的影響程度。

5.1 設計能量模型的方法

本論文所要設計的能量模型主要有四個影響因素，分別為：聲調(tone)、音節在詞的位置、基本音節種類、音節間的連音狀態，假設音節能量可由 8 個能量狀態所組成， 每一個能量狀態內之聲音訊號利用 13 維 MFCC 參數表示，其中第 1 維參數為能量狀態上的平均能量，如何決定音節各個能量狀態長度的問題，在後面的章節討論。假設所有的影響因素可用累加的方式來表示，如下式：

$$\mathbf{se}_{n,s} = \mathbf{se}^{r}_{n,s} + \boldsymbol{\alpha}_{t_{n,s}} + \boldsymbol{\alpha}_{w_{n,s}} + \boldsymbol{\alpha}_{sy_{n,s}} + \boldsymbol{\alpha}^{f}_{c_{n-1},fi\_in_{n-1},s} + \boldsymbol{\alpha}^{b}_{c_n,fi\_in_n,s} + \boldsymbol{\mu}^{e}_{s} \tag{5.1}$$

其中各項變數定義為

$\mathbf{se}_{n,s}$：第 n 個音節的第 s 個能量狀態的能量值。

$\mathbf{se}^{r}_{n,s}$：第 n 個音節的第 s 個能量狀態殘餘(residual)。

$\boldsymbol{\alpha}_{t_{n,s}}$：第 n 個音節的第 s 個能量狀態的聲調影響音素，$t_n \in \{1,2,3,4,5\}$

$\boldsymbol{\alpha}_{w_{n,s}}$：第 n 個音節的第 s 個能量狀態音節在詞的位置影響音素，$w_n \in \{(2,1),(2,2),(3,1)\ldots,(8,8)\}$

$\boldsymbol{\alpha}_{sy_{n,s}}$：第 n 個音節的第 s 個能量狀態的音節影響因素，$sy_n \in \{1,2\ldots411\}$

c_n：在第 n 個音節與第 n+1 個音節的連音狀態，$c_n \in \{1,2,3\}$

fi_in_n：第 n 個音節的韻母與第 n+1 個音節的聲母組合。

$\boldsymbol{\alpha}^{b}_{c_n, fi\_in_n, s}$：第 n 個音節的第 s 個能量狀態受第 n+1 個音節影響的影響因素。

$\boldsymbol{\alpha}^{f}_{c_{n-1}, fi\_in_{n-1}, s}$：第 n 個音節的第 s 個能量狀態受第 n-1 個音節影響的影響因素。

$\boldsymbol{\mu}^{e}_{s}$：整體第 s 個能量狀態的平均能量。

在此研究，我們假設 $\mathbf{se}^{r}_{n,s}$ 呈高斯分佈。圖 5.1-1 為音節能量和主要的四個影響因素之關係的示意圖：

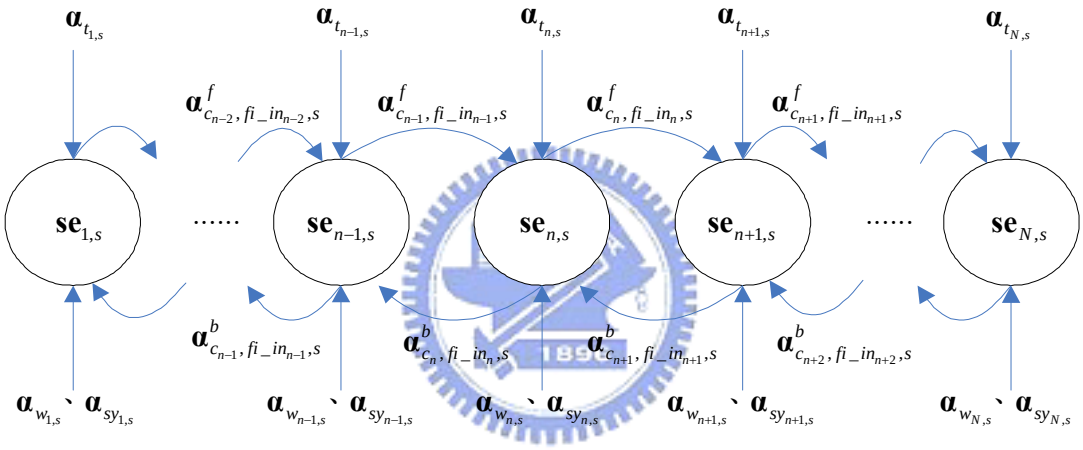

圖 5.1-1：音節能量與影響因素的示意圖

5.2 訓練能量模型

5.2.1 推導各個影響因素

在給定所觀察音節的聲調、在詞的位置、基本音節與受前後音節的連音狀態及音節間的韻母及聲母組合，則觀察到每個音節的某個能量狀態的能量分佈滿足下式：

$$P(\mathbf{se}_{n,s} \mid t_{n,s}, w_{n,s}, sy_{n,s}, c_{n-1}, c_n, fi\_in_{n-1}, fi\_in_n) = N(\mathbf{se}_{n,s}; \boldsymbol{\alpha}_{t_{n,s}} + \boldsymbol{\alpha}_{w_{n,s}} + \boldsymbol{\alpha}_{sy_{n,s}} + \boldsymbol{\alpha}^{f}_{c_{n-1}, fi\_in_{n-1}, s} + \boldsymbol{\alpha}^{b}_{c_n, fi\_in_n, s} + \boldsymbol{\mu}^{e}_{s}, \mathbf{R}^{e}) \quad (5.2)$$

利用最大概似法則推導出各個影響因素之式子，整理如下：

$$\boldsymbol{\alpha}_{t_s}=\frac{\sum_{n=1}^{N}\left(\mathbf{se}_{n,s}-\boldsymbol{\alpha}_{w_{n,s}}-\boldsymbol{\alpha}_{sy_{n,s}}-\boldsymbol{\alpha}^{f}_{c_{n-1},fi\_in_{n-1},s}-\boldsymbol{\alpha}^{b}_{c_n,fi\_in_n,s}-\boldsymbol{\mu}^{e}_{s}\right)\delta\left(t_{n.s}=t\right)}{\sum_{n=1}^{N}\delta\left(t_{n.s}=t\right)} \tag{5.3}$$

$$\boldsymbol{\alpha}_{w_s}=\frac{\sum_{n=1}^{N}\left(\mathbf{se}_{n,s}-\boldsymbol{\alpha}_{t_{n,s}}-\boldsymbol{\alpha}_{sy_{n,s}}-\boldsymbol{\alpha}^{f}_{c_{n-1},fi\_in_{n-1},s}-\boldsymbol{\alpha}^{b}_{c_n,fi\_in_n,s}-\boldsymbol{\mu}^{e}_{s}\right)\delta\left(w_{n,s}=w\right)}{\sum_{n=1}^{N}\delta\left(w_{n,s}=w\right)} \tag{5.4}$$

$$\boldsymbol{\alpha}_{sy_s}=\frac{\sum_{n=1}^{N}\left(\mathbf{se}_{n,s}-\boldsymbol{\alpha}_{t_{n,s}}-\boldsymbol{\alpha}_{w_{n,s}}-\boldsymbol{\alpha}^{f}_{c_{n-1},fi\_in_{n-1},s}-\boldsymbol{\alpha}^{b}_{c_n fi\_in_n,s}-\boldsymbol{\mu}^{e}_{s}\right)\delta\left(sy_{n,s}=sy\right)}{\sum_{n=1}^{N}\delta\left(sy_{n,s}=sy\right)} \tag{5.5}$$

$$\boldsymbol{\alpha}^{f}_{c,fi\_in,s}=\frac{\sum_{n=1}^{N}\left(\mathbf{se}_{n,s}-\boldsymbol{\alpha}_{t_{n,s}}-\boldsymbol{\alpha}_{w_{n,s}}-\boldsymbol{\alpha}_{sy_{n,s}}-\boldsymbol{\alpha}^{b}_{c_n,fi\_in_n,s}-\boldsymbol{\mu}^{e}_{s}\right)\delta\left(fi\_in_{n-1}=fi\_in\right)\delta\left(c_{n-1}=c\right)}{\sum_{n=1}^{N}\delta\left(fi\_in_{n-1}=fi\_in\right)\delta\left(c_{n-1}=c\right)} \tag{5.6}$$

$$\boldsymbol{\alpha}^{b}_{fi\_in,s}=\frac{\sum_{n=1}^{N}\left(\mathbf{se}_{n,s}-\boldsymbol{\alpha}_{t_{n,s}}-\boldsymbol{\alpha}_{w_{n,s}}-\boldsymbol{\alpha}^{f}_{c_{n-1},fi\_in_{n-1},s}-\boldsymbol{\alpha}_{sy_{n,s}}-\boldsymbol{\mu}^{e}_{s}\right)\delta\left(fi\_in_{n}=fi\_in\right)\delta\left(c_{n}=c\right)}{\sum_{n=1}^{N}\delta\left(fi\_in_{n}=fi\_in\right)\delta\left(c_{n}=c\right)} \tag{5.7}$$

$$\mathbf{R}^{e}=\frac{\sum_{n=1}^{N}\mathbf{X}\mathbf{X}^{\mathrm{T}}}{N} \tag{5.8}$$

其中 $\mathbf{X}=\mathbf{se}_{n,s}-\boldsymbol{\alpha}_{t_{n,s}}-\boldsymbol{\alpha}_{w_{n,s}}-\boldsymbol{\alpha}_{sy_{n,s}}-\boldsymbol{\alpha}^{f}_{c_{n-1},fi\_in_{n-1},s}-\boldsymbol{\alpha}^{b}_{c_n,fi\_in_n,s}-\boldsymbol{\mu}^{e}_{s}$，$\mathbf{R}^{e}$ 為 covariance matrix。

5.2.2 訓練流程

各別取出分割後能量狀態內的 13 維 MFCC 參數之平均值，依序將基本音節、音節在詞的位置、聲調、受前後音節等影響因素及 covariance matrix 的參數

值更新，然後使用更新後的參數值，算出整個語料庫的最大概似法則分數，使用更新後的影響因素參數值去更新能量狀態切割位置，一直重覆更新影響因素參數值、分數及能量狀態分割位置，直到分數的變化量小於10^{-7}。訓練流程圖如圖5.2.2-1，在後面章節說明如何初始化分割狀態及如何由更新後的影響因素參數值去更新能量狀態切割位置。

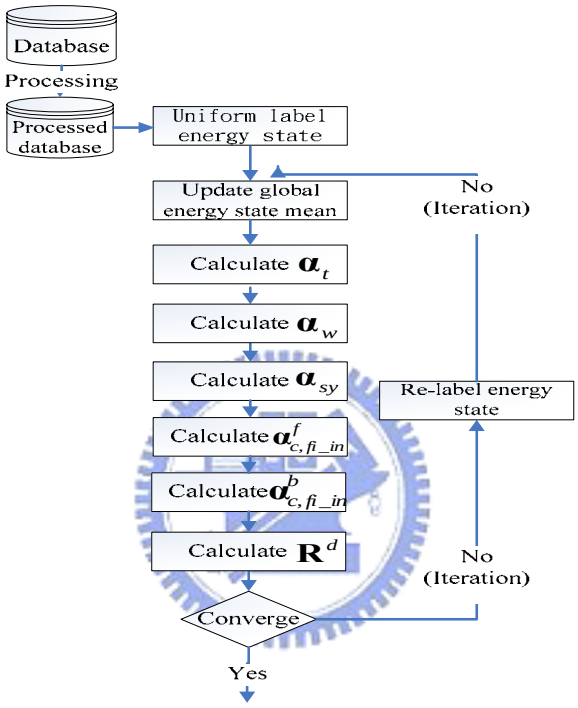

圖 5.2.2-1：訓練流程圖

5.2.3 初始設定及更新狀態分割

我們將音節上的八個能量狀態均勻分割視為初始設定。在本實驗中將所有影響因素的初始值全部預設為 0，直接由訓練流程去更新參數值。利用式子(5.3)~(5.8)更新影響因素參數值，假設音節上的某個能量狀態內抽取 13 維 MFCC 參數，均滿足高斯分佈，利用音節上的聲調、音節在詞的位置、基本音節、受前後音節的連音狀態及韻母、聲母組合等資訊，計算出該音節上各個能量狀態的 mean 和 covariance，式子如(5.9)、(5.10)。

在更新能量狀態分割時，由音節上的 frame 抽取出 13 維 MFCC 參數，利用各別能量狀態的 mean 和 covariance 計算出在各別的分數，再使用 Viterbi algorithm 決定每個 frame 最有可能屬於那一個能量狀態。由於語料庫有極短或發音不全等音節，所以我們限制除了第一個及第八個能量狀態，其餘中間能量狀態有機率允許跳過狀態，但能量狀態不可往回，如圖 5.2.3-1。

$$\mathbf{state\_\mu_s} = \boldsymbol{\alpha}_{t_{n,s}} + \boldsymbol{\alpha}_{w_{n,s}} + \boldsymbol{\alpha}_{sy_{n,s}} + \boldsymbol{\alpha}^{f}_{c_{n-1},fi\_in_{n-1},s} + \boldsymbol{\alpha}^{b}_{c_n,fi\_in_n,s} + \boldsymbol{\mu}^{e}_{s} \tag{5.9}$$

$$\mathbf{state\_R_s} = \frac{\sum_{n=1}^{N} \mathbf{X}\mathbf{X}^{\mathrm{T}}\delta(s_n = s)}{\sum_{n=1}^{N} \delta(s_n = s)} \tag{5.10}$$

其中 $\mathbf{X} = \mathbf{se}_{n,s} - \boldsymbol{\alpha}_{t_{n,s}} - \boldsymbol{\alpha}_{w_{n,s}} - \boldsymbol{\alpha}_{sy_{n,s}} - \boldsymbol{\alpha}^{f}_{c_{n-1},fi\_in_{n-1},s} - \boldsymbol{\alpha}^{b}_{c_n,fi\_in_n,s} - \boldsymbol{\mu}^{e}_{s}$

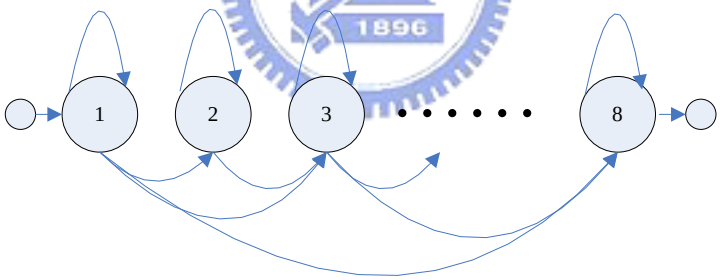

圖 5.2.3-1：狀態走勢

5.3 訓練結果與分析

訓練模型後，使用能量模型去預測音節能量，我們進一步去討論各個影響因素對於能量的影響及貢獻。

5.3.1 利用能量模型預測音節能量

整體能量軌跡預測如圖 5.3.1-1 所示，其中黑色線(實線)為原始音節的能量軌跡，紅色線(點線)為由能量模型所預測音節的能量軌跡。圖 5.3.1-2 為三字詞的能量預測圖，其中藍色分別為音節狀態切割位置。

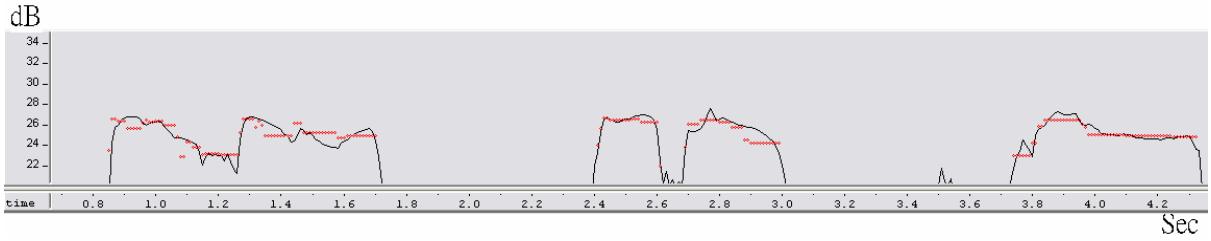

圖 5.3.1-1：整體能量預測圖

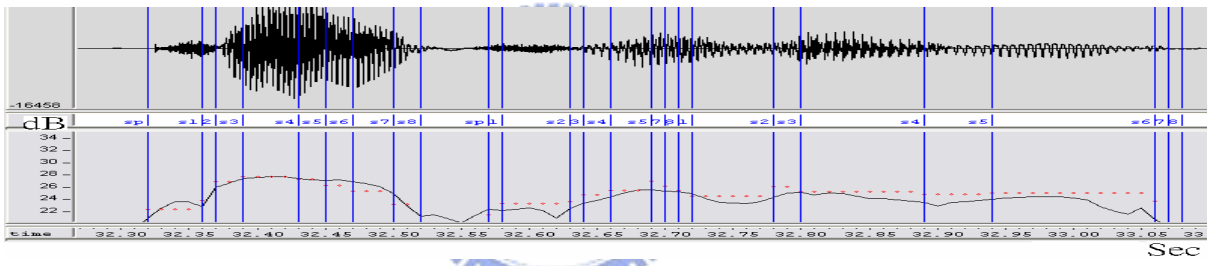

圖 5.3.1-2：三字詞的能量預測圖

5.3.2 聲調影響因素(tone affecting factor)

聲調對音節能量影響程度如圖 5.3.2-1，由圖形可觀察得能量範圍小，可能對音節能量影響較小，由觀察得一聲及二聲能量較集中於後面的能量狀態。

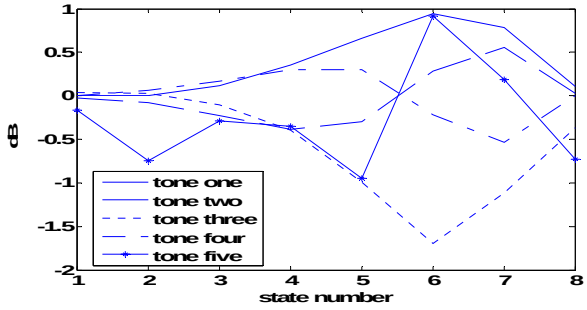

圖 5.3.2-1：聲調影響因素對能量影響

5.3.3 詞的位置影響因素(word position affecting factor)

詞的位置之影響因素對音節能量影響程度如圖 5.3.3-1~5.3.3-4，由觀察得知完整不間斷韻律詞音節隨著在詞的位置愈往後，能量愈有下降的現象。詞中位置的音節有個有趣的現象，音節前幾個能量狀態為聲母部份，使之能量往下，後幾個能量狀態為韻母部份，使提高能量值可接下個音節。

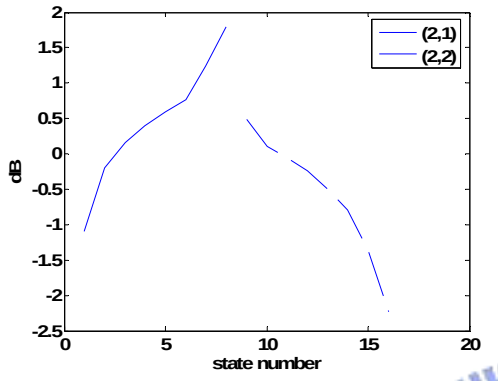

圖 5.3.3-1：二字詞

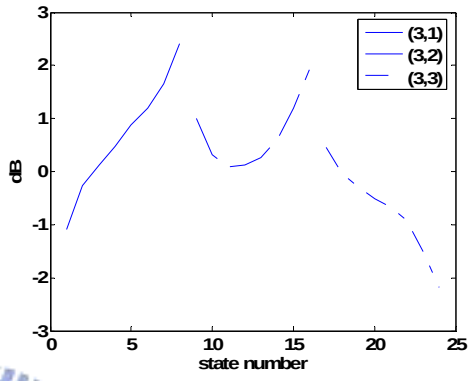

圖 5.3.3-2：三字詞

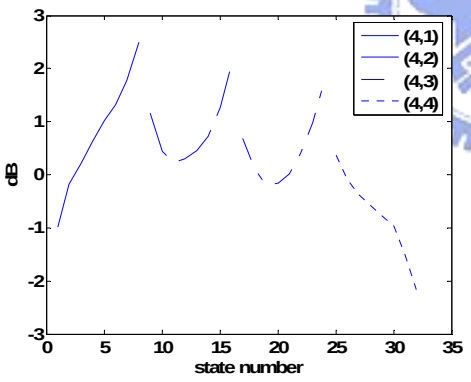

圖 5.3.3-3：四字詞

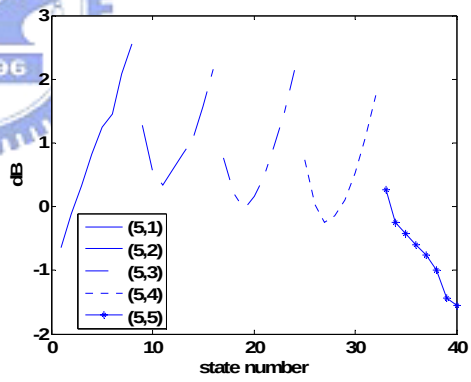

圖 5.3.3-4：五字詞

5.3.4 受前音節影響因素(forward affecting factor)

受前音節之影響因素對音節能量影響程度如圖 5.3.4-1~5.3.4-6，圖上各個連音狀態的數量可做為判斷影響因素能量軌跡的可信度。

觀察圖 5.3.4-1 聲母類別 1 為ㄇ、ㄋ、ㄌ、ㄖ及空聲母，若連音較嚴重(state 1)則能量延續，其他兩種連音狀態明顯能量有下降現象，在其他聲母類別圖形中

state 1 數量嚴重不足，參考價值不多。觀察圖 5.3.4-2 聲母類別 2 為ㄏ、ㄒ、ㄕ、ㄈ、ㄙ，此類別在表 4.3.5-1：pause 長度選取表，得知此類別 pause 較短，所以與前面音節能量相連續，且此類別聲母能量偏弱，所以對受前音節影響因素的前幾個狀態能量往上。由圖 5.3.4-3、圖 5.3.4-5 聲母類別 3、5 為爆破音，在表 4.3.5-1 得知此類別 pause 較長，所以與前面音節能量不連續，所以對受前音節影響因素的前幾個狀態能量往下。沒有被討論的類別影響不大。

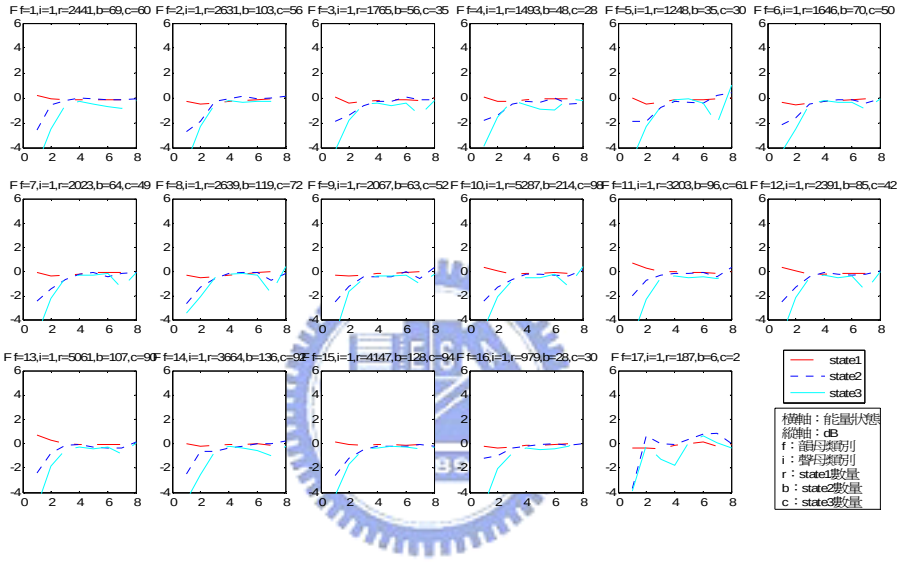

圖 5.3.4-1：聲母類別 1 受前音節韻母各種類別

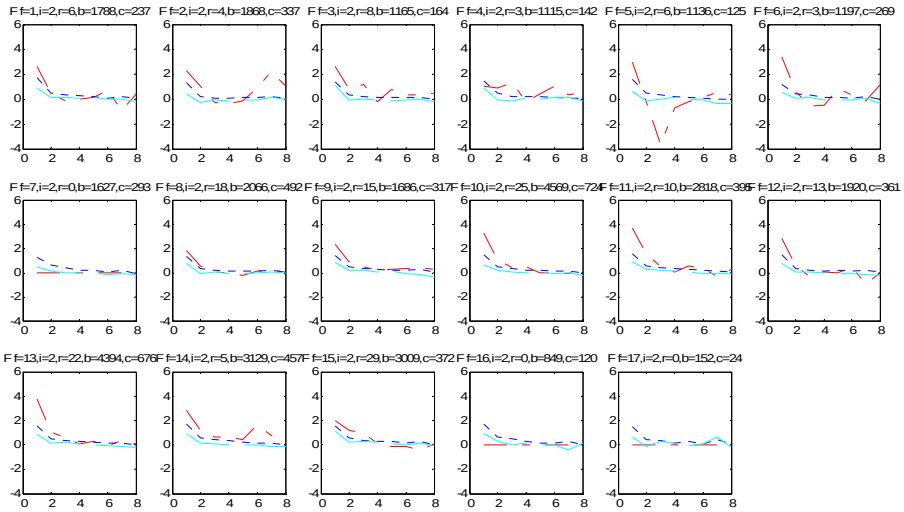

圖 5.3.4-2：聲母類別 2 受前音節韻母各種類別

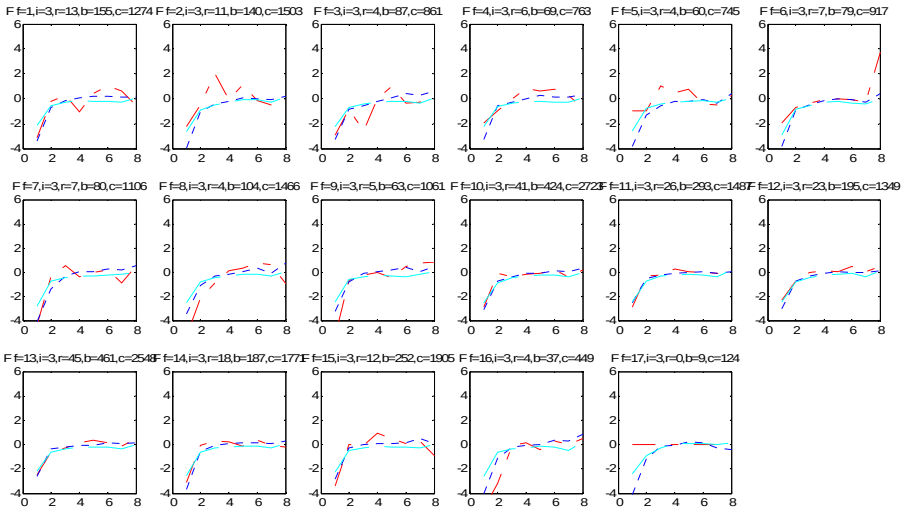

圖 5.3.4-3：聲母類別 3 受前音節韻母各種類別

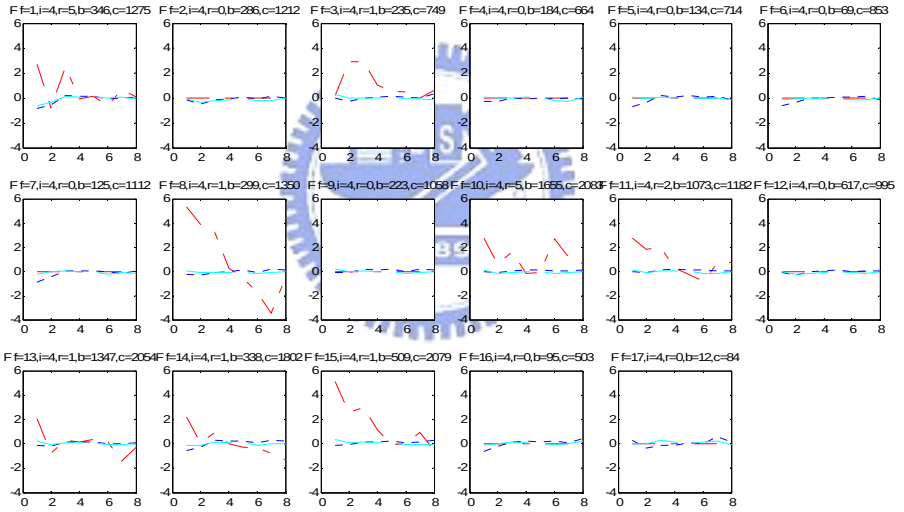

圖 5.3.4-4：聲母類別 4 受前音節韻母各種類別

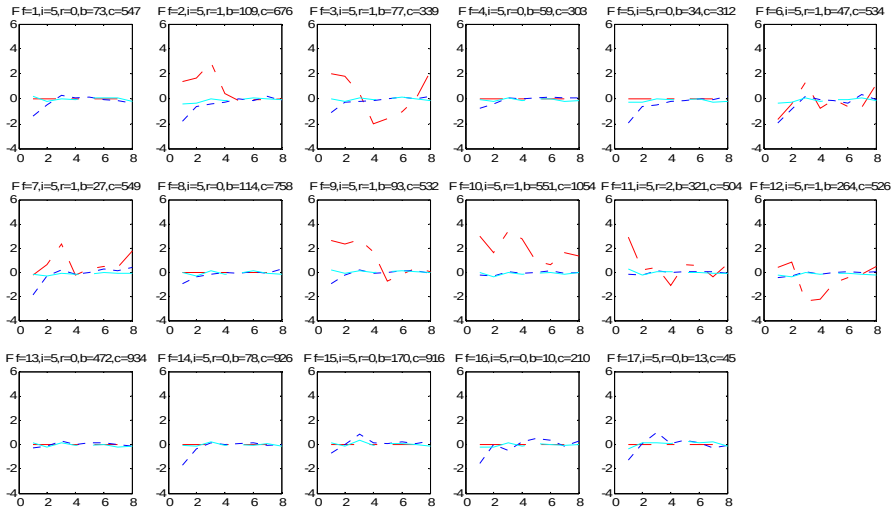

圖 5.3.4-5：聲母類別 5 受前音節韻母各種類別

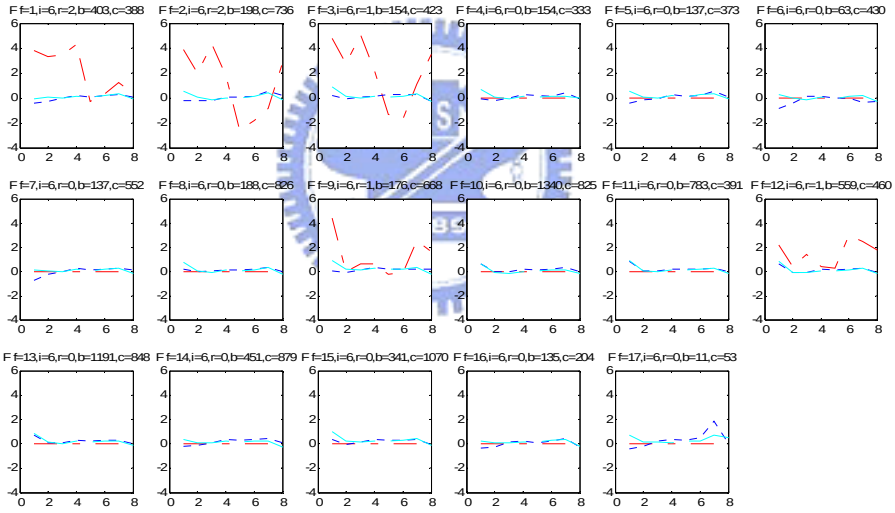

圖 5.3.4-6：聲母類別 6 受前音節韻母各種類別

5.3.5 受後音節影響因素(forward affecting factor)

受後音節之影響因素對音節能量影響程度如圖 5.3.5-1~5.3.5-17。除了聲母類別為 1 時 state 1 較有可信度，其餘較無可信度。觀察得受後音節的影響比受前音節的影響較小，僅在音節後面幾個狀態能量往下趨勢。

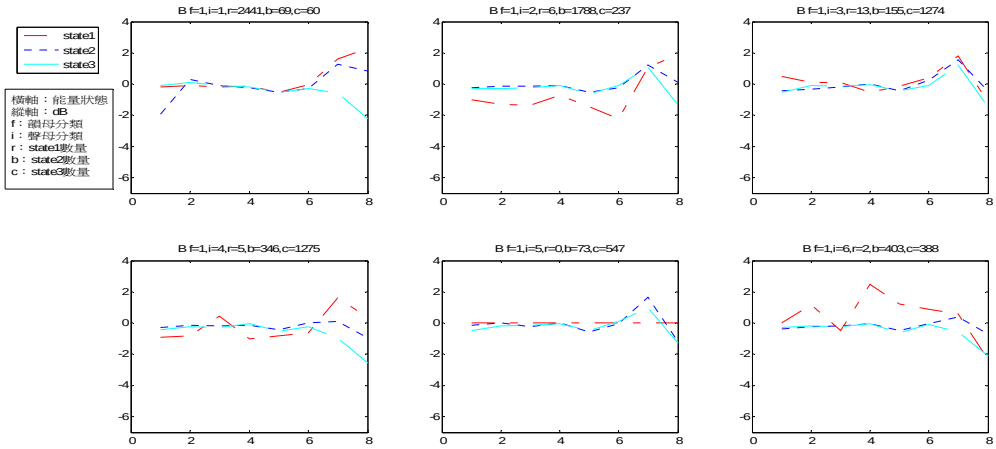

圖 5.3.5-1：韻母類別 1 受後音節聲母各種類別

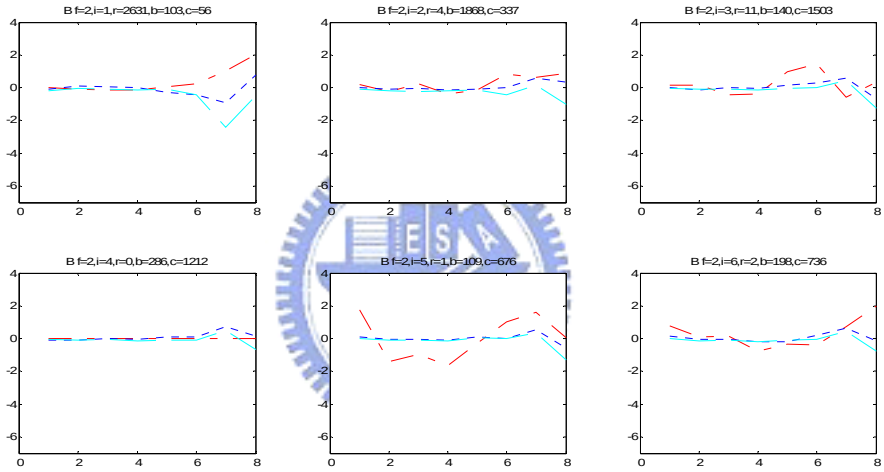

圖 5.3.5-2：韻母類別 2 受後音節聲母各種類別

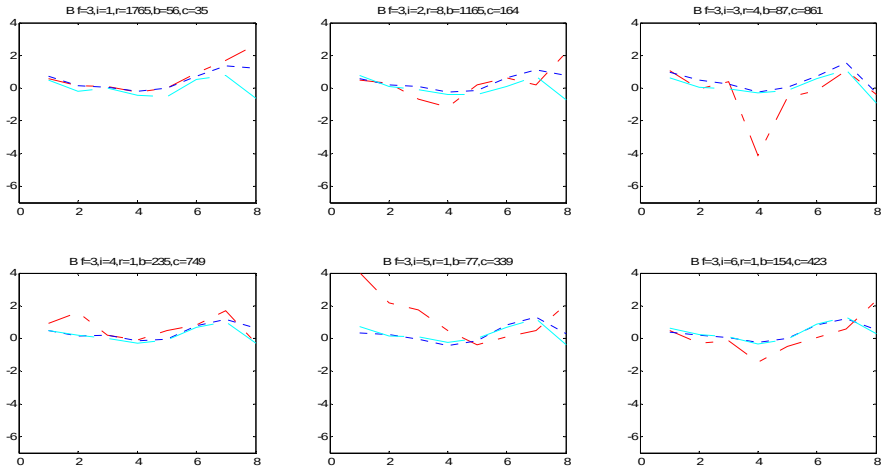

圖 5.3.5-3：韻母類別 3 受後音節聲母各種類別

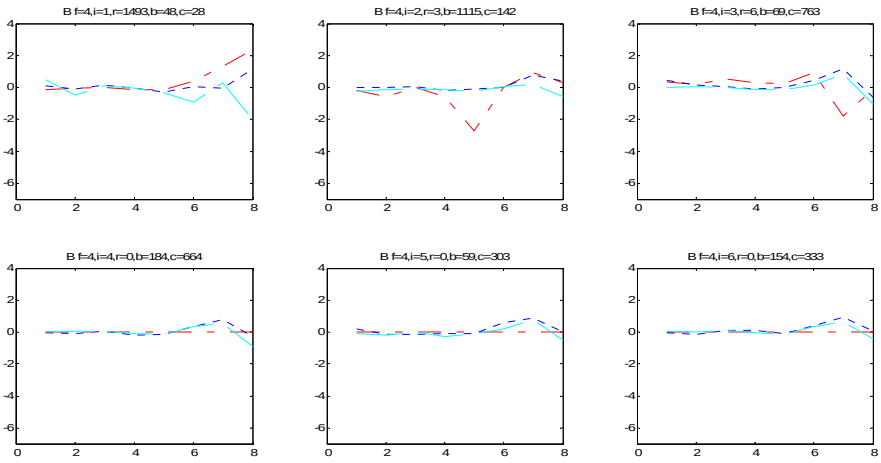

圖 5.3.5-4：韻母類別 4 受後音節聲母各種類別

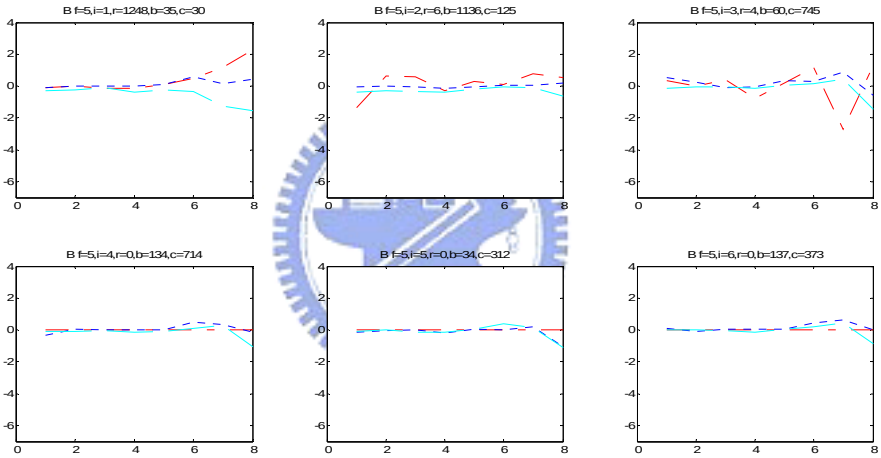

圖 5.3.5-5：韻母類別 5 受後音節聲母各種類別

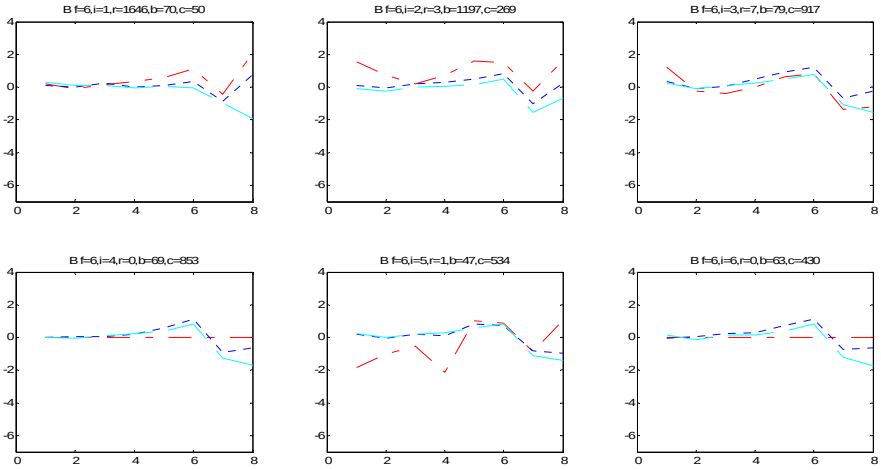

圖 5.3.5-6：韻母類別 6 受後音節聲母各種類別

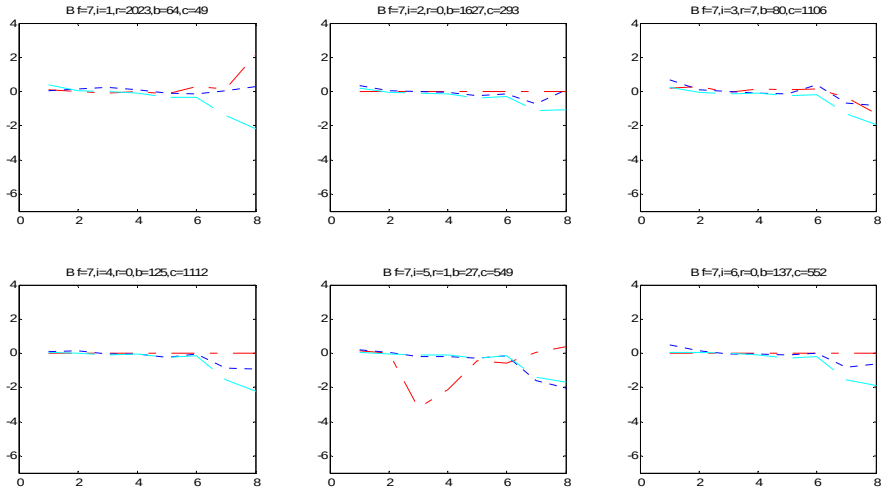

圖 5.3.5-7：韻母類別 7 受後音節聲母各種類別

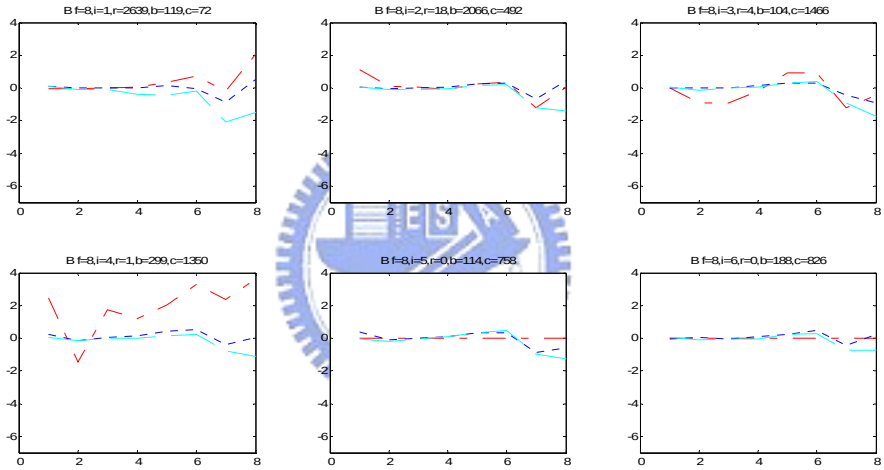

圖 5.3.5-8：韻母類別 8 受後音節聲母各種類別

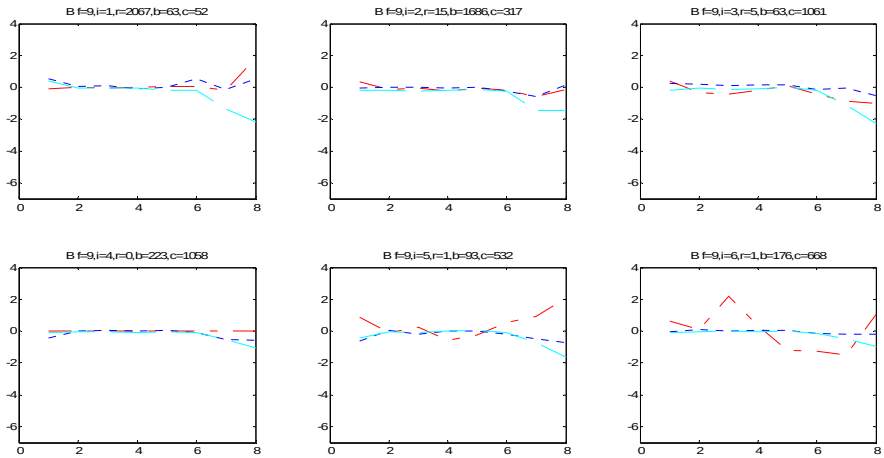

圖 5.3.5-9：韻母類別 9 受後面節聲母各種類別

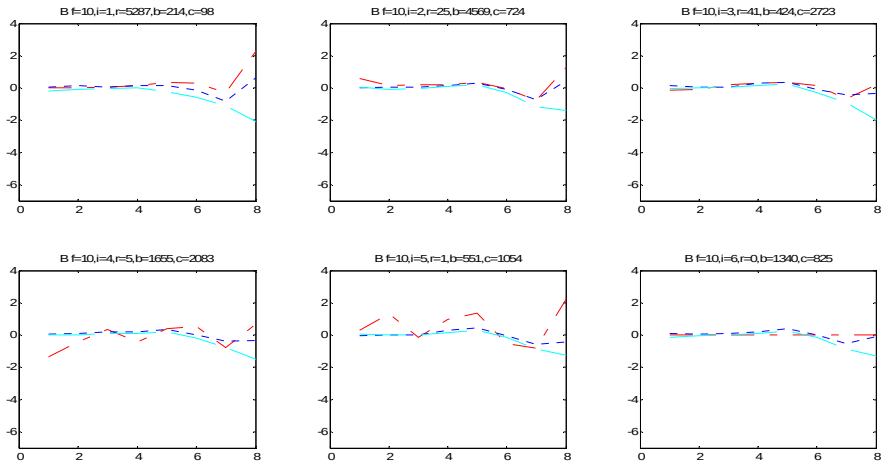

圖 5.3.5-10：韻母類別 10 受後音節聲母各種類別

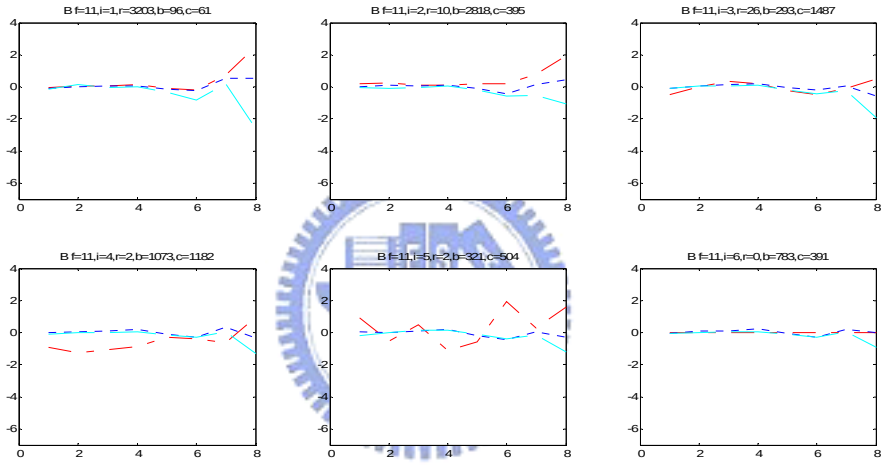

圖 5.3.5-11：韻母類別 11 受後音節聲母各種類別

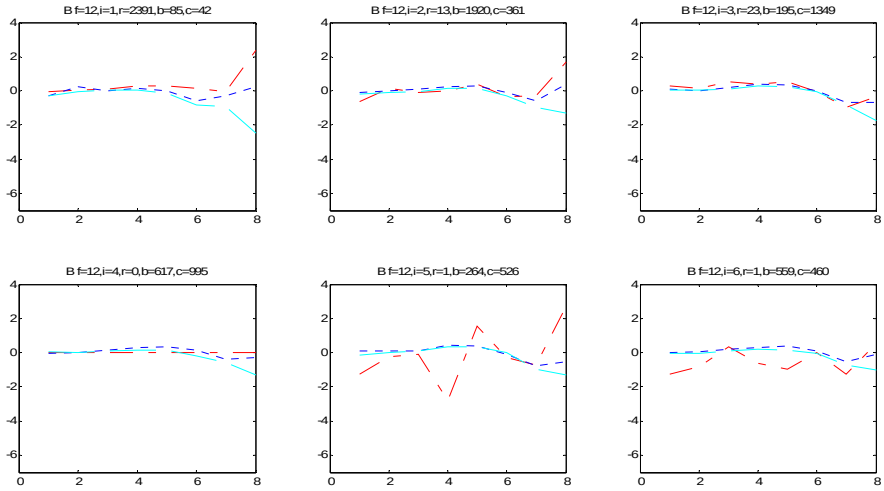

圖 5.3.5-12：韻母類別 12 受後音節聲母各種類別

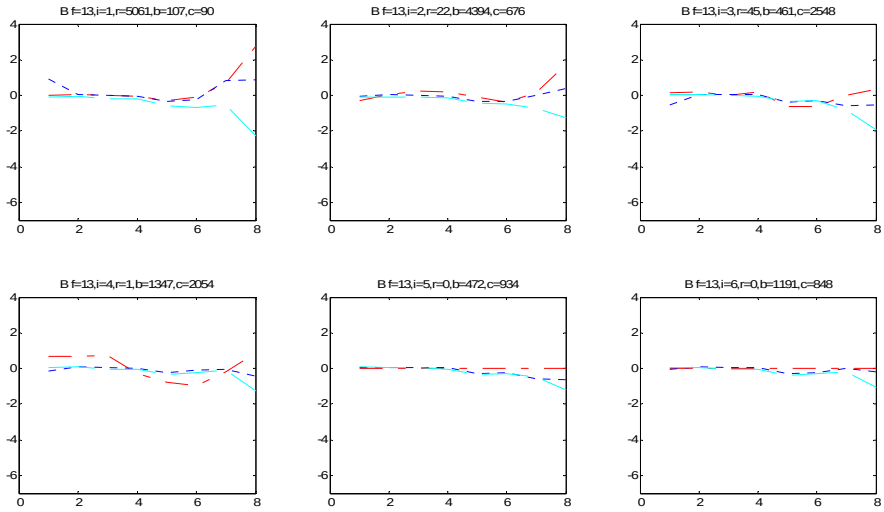

圖 5.3.5-13：韻母類別 13 受後音節聲母各種類別

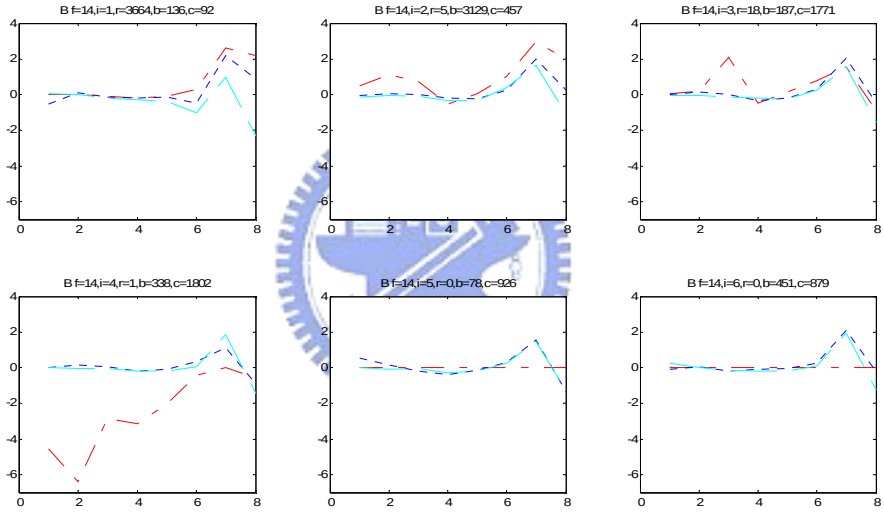

圖 5.3.5-14：韻母類別 14 受後音節聲母各種類別

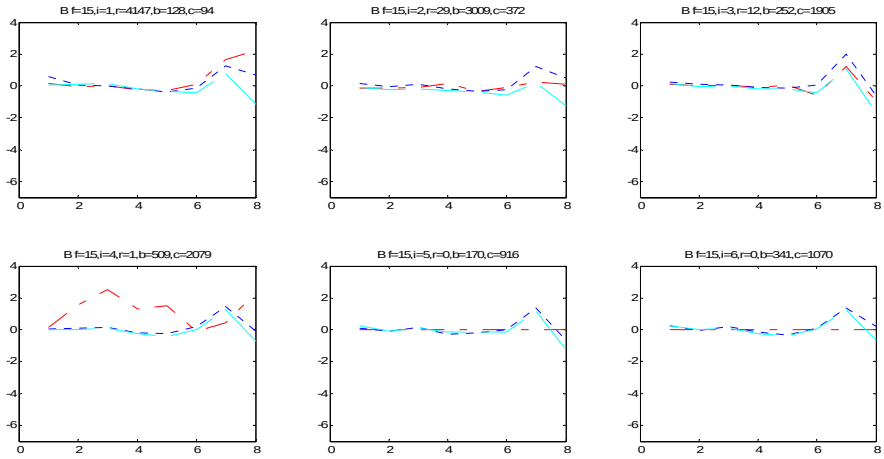

圖 5.3.5-15：韻母類別 15 受後音節聲母各種類別

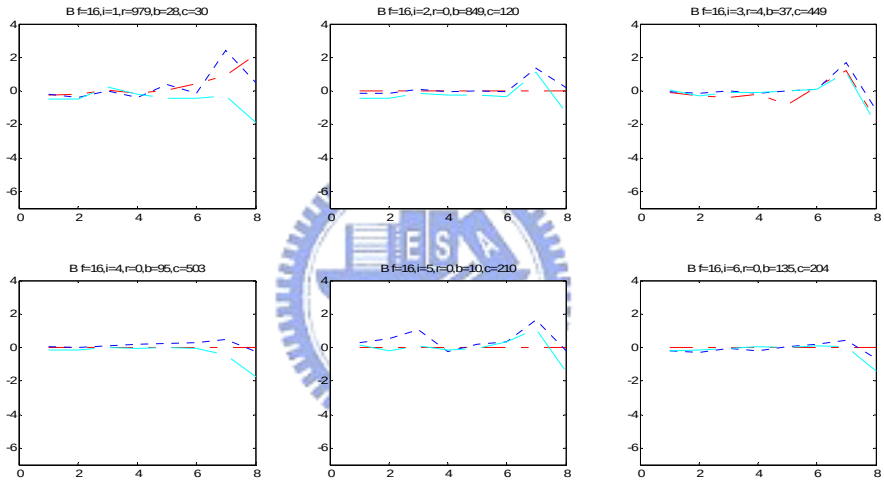

圖 5.3.5-16：韻母類別 16 受後音節聲母各種類別

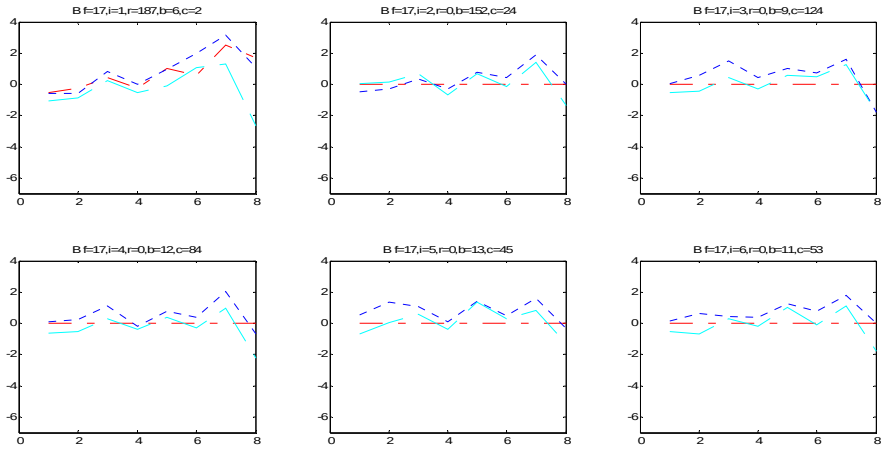

圖 5.3.5-17：韻母類別 17 受後音節聲母各種類別

5.3.6 音節影響因素(syllable affecting factor)

音節對音節能量影響程度圖如附錄二，圖 5.3.6-1 縱向分別為ㄓ、ㄔ、ㄕ、ㄖ、ㄘ的聲母，與橫向分別為ㄚ、ㄜ、ㄞ、ㄡ、ㄢ的韻母組合的音節能量軌跡，可發現相同聲母或韻母有相似的能量分佈。

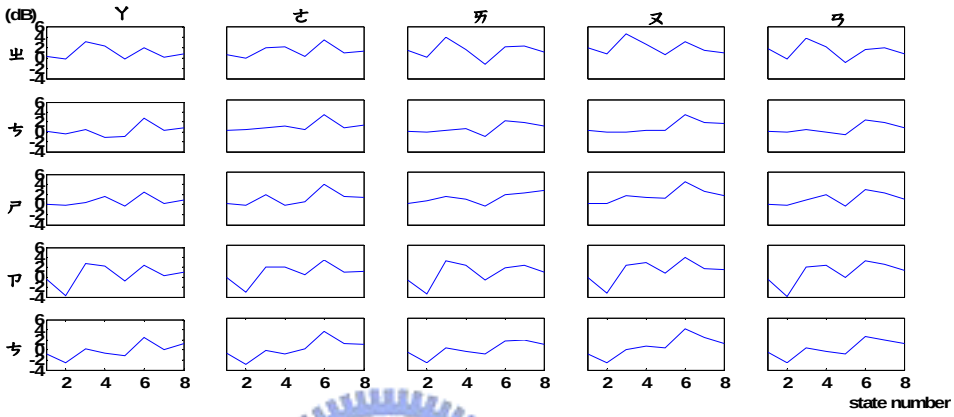

圖 5.3.6-1： 聲母及韻母組合能量軌跡比較圖

第六章 韻律系統展示

在前面章節討論第三章 pitch 模型、第四章 duration 模型及第五章能量模型，本章即是探討如何由這些資訊變成有效的韻律訊息，且如何展示韻律系統。

6.1 韻律系統整體架構

圖 6.1-1 為韻律展示系統整體架構，其中詞典(Lexicon)搜尋器含有 1 字詞到 6 字詞，當輸入 n 字詞時，則詞典搜尋器提供 n 字詞的詞典方便搜尋，當沒有搜尋到輸入的詞，則由 1 字詞的詞典搜尋，可能有破音字或變調影響搜尋結果。

由詞典搜尋器輸出的參數有音節碼、音節聲調、音節位置等，可當成韻律訊息產生器的輸入，圖 6.1-2 為韻律訊息產生器方塊圖：

圖 6.1-1：韻律展示系統整體架構

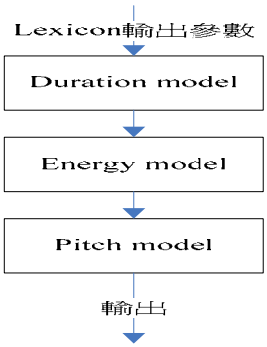

圖 6.1-2：韻律訊息產生器方塊圖

6.1.1 Duration 模型

詞典搜尋器輸出的參數當成韻律訊息產生器的輸入，根據這些參數我們可由表 4.3.5-1 去選取 pause 長度及式子(4.1)預測每個音節長度。

6.1.2 能量模型

我們假設每個音節的各個能量狀態長度有一定的比例，統計方法如式子(6.1)

$$Ratio_{sy,s} = \frac{\sum_{n=1}^{N} Dur_{n,num}\,\delta(sy = sy_n)\,\delta(s = s_{n,num})}{\sum_{n=1}^{N}\sum_{num=1}^{8} Dur_{n,num}\,\delta(sy = sy_n)} \tag{6.1}$$

$sy = \{1, 2 \ldots 411\}$　　$s = \{1, 2 \ldots 8\}$

其中：

$Ratio_{sy,s}$ ： 第 *sy* 個音節碼的第 s 能量狀態之長度比例。

$Dur_{n,num}$ ： 觀察得第 *n* 個音節的第 *num* 能量狀態長度。

sy_n ：觀察得第 *n* 個音節為第 *sy* 個音節碼。

$s_{n,num}$ ：第 *n* 個音節的第 *num* 個能量狀態。

由統計狀態長度比例，可由 duration 模型預測音節長度來決定能量狀態長度，這個假設只為了方便展示韻律系統，並無根據。獲得音節每個能量狀態長度後，可利用第五章的能量模型討論，使用式子(5.1)預測出每個能量狀態上的能量值。

6.1.3 Pitch 模型

為獲得韻律訊息，最後音節 pitch 訊息，則利用第三章的 pitch 模型的式子(3.1)預測每個音節的 pitch。如何決定 pitch 在音節的啟始位置，我們可利

用第五章所討論的能量模型，由能量狀態切割位置觀察得，若音節有摩擦音時，第一個狀態至第三個狀態為摩擦音部位，第四個狀態至第八個狀態為元音部位，所以 pitch 的啟始位置可假設由第四個狀態開始。若音節無摩擦音時，可假設 pitch 的啟始位置由第一個狀態開始。這些假設只為了方便展示韻律系統，並無意義。

6.2 展示韻律系統

利用 6.1 章節觀察到每個音節的韻律訊息以繪圖方式表達，介面操作步驟如圖 6.2-1~6.2-5，其中圖 6.2-3 為分析輸入詞的分析結果，分析的結果有詞典搜尋器輸出的音節碼、音節聲調、音節在詞的位置，有韻律訊息產生器輸出的 pitch、duration、能量等韻律訊息，其中音節長度預測結果直接輸出於介面上。

圖 6.2-1：呼叫輸入介面

圖 6.2-2：輸入詞

圖 6.2-3：相關資訊展示

圖 6.2-4：繪圖鍵

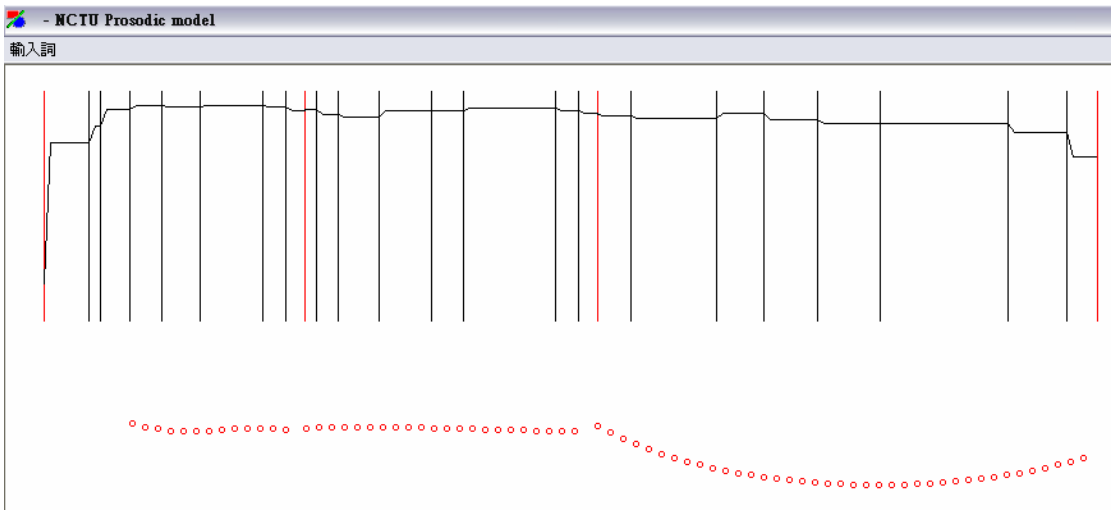

圖 6.2-5：繪圖結果

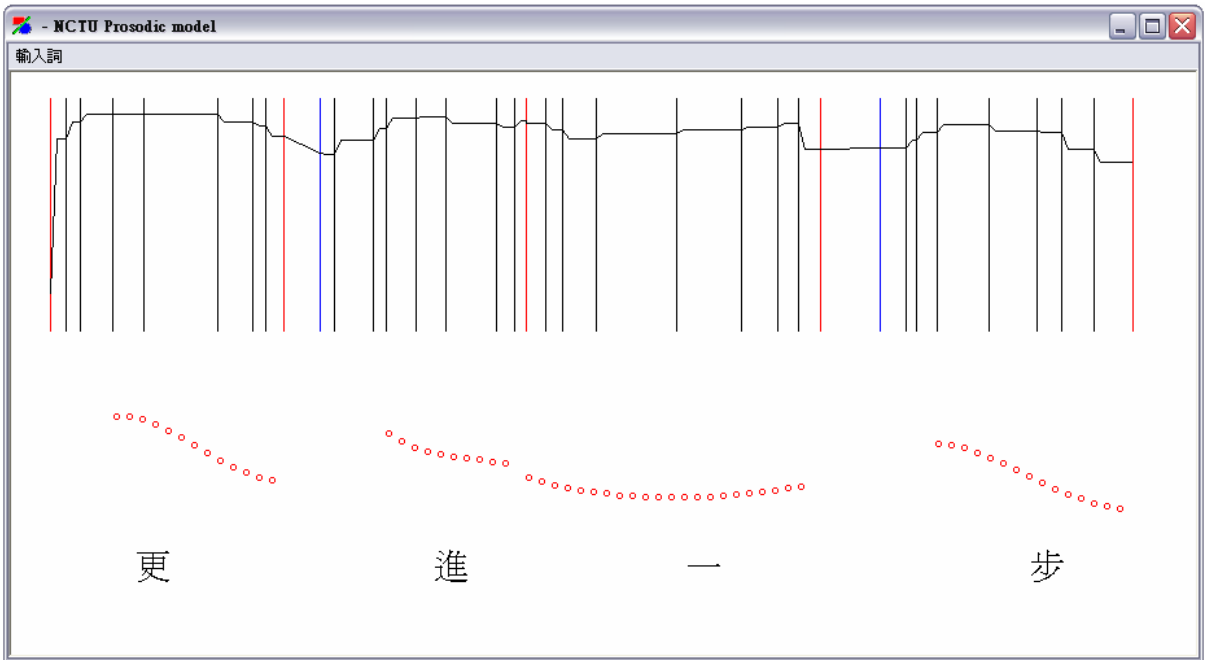

圖 6.2-6：繪圖結果

其中圖的紅色直線為音節長度邊界，黑色直線為音節的各個狀態邊界，黑色橫線為音節能量軌跡，藍色線為音節間長度邊界，最後紅色圓點為音節基頻軌跡。

第七章 結論與未來展望

本論文提出使用各種影響因素相加來預測各種韻律訊息，及討論各種影響因素對韻律訊息的分析，由實驗結果證實各種影響因素相加便可預測各種韻律訊息。相信未來應用於語音合成，可明顯自然流暢許多。

然而，因為語料庫音節切割目前是利用自己撰寫程式自動切割，若未來有更好的切割方法，便可將各種模型訓練的更好，再者，能量模型所考慮的因素尚不周全，未來可針對能量模型更深入探討，使能量模型更精緻。最後，因為本論文所提出的方法，不因語言不同而有所改變，所以未來可朝向建立一套整合國、台、客語的韻律產生器及韻律分析邁進。

參考文獻

[1] C.Y. Chiang, Y.R. Wang, and S.-H. Chen, “On the Inter-syllable Coarticulation Effect of Pitch Modeling for Mandarin Speech, “Proc. of Interspeech 2005, Lisbon, Portugal, pp. 3269-3272

[2] The HTK Book(for HTK Version 3.2.1)

[3] Wavesufer Homepage：http://www.speech.kth.se/wavesurfer/

[4] S.H.Chen,and S.H.Hwang,and Y.R.Wang, “An RNN-based prosodic information synthesizer for Mandarin text-to-speech, “IEEE Trans. Speech and Audio Processing,vol.6,no.3,pp.226-239,May 1998.

[5] S.H. Chen, and Y.R. Wang, “Vector quantization of pitch information in Mandarin speech, “IEEE Trans. Communication, vol.COM-38, pp.1317-1320,1990.

[6] S.H. Hwang, S.H. Chen, and Y.R. Wang, “A Mandarin Text-to-Speech system, “ in Proc. ICSLP-96, pp.1421-1424, Oct.1996.

[7] L.S.Lee,C.Y.Tseng,and M. Ouh-Young, ”The synthesis rules in Chinese text-to-speech system, ”IEEE Trans. Acoust, Speech, Signal Processing,vol.37,n0.9,p1309-1319,Sep. 1989

[8] L.S.Lee,C.Y.Tseng,and C.J.Hesih, ”Improved tone concatenation rules in a formant-based Chinese text-to-speech system, ”IEEE Trans.Speech and Audio Processing,Vol.1,No.3,pp.287-294,July 1993.

[9] S.H.Chen,S.G.Chang,and S.M.Lee, ”A statistical model based fundamental frequency synthesizer for Mandarin speech, ”J.Acoust. Soc. Am.,92(1),pp.114-120,July 1992

[10] 黃紹華，“中文文句翻語音系統中韻律訊息產生器之研究“，國立交通大學博士論文，民國八十五年六月。