國 立 交 通 大 學

電信工程學系

碩士論文

中文單詞之韻律模擬與其應用

Prosody Modeling for Isolated Mandarin Words and Its Application

研 究 生：施宏廣

指導教授：陳信宏　博士

中華民國九十七年八月

中文單詞之韻律模擬與其應用

Prosody Modeling for Isolated Mandarin Words and Its Application

研 究 生：施宏廣　　Student： Hung-Kuang Shih

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

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

August, 2008

Hsinchu, Taiwan, Republic of China

中華民國九十七年八月

國立交通大學
電信工程學系碩士班

論文口試委員會審定書

本校 電信工程學系 碩士班 施宏廣 君

所提論文(中文) 中文單詞之韻律模擬與其應用

(英文) Prosodic Modeling for Isolated Mandarin Words and Its Application

合於碩士資格水準、業經本委員會評審認可。

口試委員：王小川　陳信宏

王逸如

李琳山

指導教授：陳信宏

系主任：陸天寧 教授

中 華 民 國　97 年　8 月　25 日

中文單詞之韻律模擬與其應用

研究生：施宏廣　　指導教授：陳信宏 博士

國立交通大學電信工程學系碩士班

中文摘要

本論文探討中文單詞的音節基頻軌跡、長度和能量三種韻律參數的模式，考慮了音節的聲調、與前後音節的連音影響、音節在詞中的位置和基本音節類別等因素對三種韻律參數的影響，藉由假設這些影響因素彼此獨立且具加成性，我們設計了一個逐項最佳化的遞迴訓練方法來由實際語料估計模型參數。以一套包含 107,936 個單詞的單一女性語者的語料庫訓練韻律模型，並分析各種影響因素的物理意義和模式的誤差；實驗結果顯示此模型能有效描述此三種韻律參數的變化。

在驗證此韻律模式的有效性後，我們使用它建立了一套中文韻律學習系統，提供非中文母語的使用者學習。使用者可依需要輸入單詞，系統會自動合成該單詞之語音及顯示正確的三種韻律參數變化讓使用者模仿；並且在使用者由麥克風錄音後進行語音切割、求取基頻軌跡和能量等處理，並提供使用者相關韻律資訊回饋學習。

Prosody Modeling for Isolated Mandarin Words and Its Application

Student：Hung-Kuang Shih　　Advisor：Dr. Sin-Horng Chen

Department of Communication Engineering

National Chiao Tung University

Abstract

This thesis can be divided into two parts. In the first part, syllable-based prosodic models for syllable F0 contour, duration and energy are proposed. Four affecting factors: syllable tone, inter-syllable coarticulation, syllable position in a word and base syllable type are considered. These affecting factors are assumed to be independent and additive. A large speech database containing 107,936 isolated Mandarin words and recorded by a professional female announcer is used to train the prosodic models. The affecting factors and modeling errors are analyzed after the convergence. It shows that the proposed model is effective.

In the second part of the thesis, a Mandarin prosody learning system for non-native speakers is built as an application of the prosodic model. The user can first enter a Mandarin word, and an ideal speech and prosodic features, including syllable F0 contour, duration and energy, will be generated based on the prosodic model. The user can then record his/her own voice, and similarly, the prosodic features of the recorded voice will be extracted by the system. The user can learn and adjust the speaking style by comparing the difference between targeted and recorded voice and prosodic features.

誌謝

本論文能夠順利完成，首先要感謝陳信宏和王逸如老師；從大學時期就跟著陳老師做專題了，這幾年間老師對我的關心和指導從來沒有少過，也讓我對問題思考的深度和廣度有所成長；王老師在 meeting 時對我們要求很嚴格，私底下卻像個大學長和我們打成一片，亦師亦友，而我在老師的訓練下在做研究和報告上更加嚴謹有邏輯。

接著要感謝什麼都會的性獸，任何問題找你都可以順利解決，讓我的研究可以很快地上手；楊智合學長脾氣好又有耐心，在我不太會寫程式的年代是你帶我一步步學習；阿德是痞子開心果，因為有你實驗室常常充滿歡笑；希群是個文藝好青年，總覺得你多才多藝深藏不露；巴金有豐富的社會經驗，和你聊天總讓我有新的啟發；輝哥白天當老師晚上還努力地來實驗室做研究，也祝你今年順利畢業！阿宅吃素也可以呷卡肥，口試前二個月你讓我看見宅男的驕傲；科科達上大夜班，和你討論程式讓我受益良多；志豪惦惦吃三碗公，愛情事業兩得意；翔耀是個性情中人，朋友有事一定義氣相挺的好漢子。還有普烏、QQ、小宋和小帥哥，和你們在一起的生活很開心，相信在你們的努力下實驗室會愈來愈好；也祝福新進實驗室的學弟妹在接下來的兩年也收穫滿載！

最後要感謝我的爸媽和三位姊姊，你們總是默默地支持著我，沒有你們就沒有今天的我；還有一路陪伴我的念真，若不是妳，這條研究之路將無法走得如此堅定踏實！

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究動機

隨著科技的發展，電腦在我們的生活中扮演著舉足輕重的角色，因此如何增進人與機器之間溝通的便利性，已成為重要的課題。目前最常見的人機介面為鍵盤輸入和螢幕輸出，但受限於體積等限制，對於手持行動裝置來說仍嫌不便。語音是人與人之間最直接的溝通，不但效率高且能傳達出情緒等更高階的訊息，被視為未來取代鍵盤和螢幕的最佳介面之一。在語音科技中，語音辨識和語音合成是相當重要的二個領域，前者使人的語言能夠被電腦理解；而後者則正好相反，讓人能夠自然的理解電腦傳達的訊息。如果這兩種技術發展純熟，有朝一日將能取代鍵盤和螢幕。

在語音合成技術中，語音的韻律模擬是相當重要的一環，直接影響了合成語音品質的好壞。為了使合成的語音自然悅耳，我們需事先模擬自然語音的韻律，提供給文字轉語音(Text-to-Speech, TTS)系統做為合成的標準。傳統的韻律模擬方法有規則法[1,2]和類神經網路法[3]，前者以語言學的知識歸納出若干規則，但由於規則繁複而難以盡善盡美；後者模擬人腦神經學習和記憶的方式將語言參數和韻律參數的規則連結，雖然效果良好，但二種參數間的關係卻像個黑盒子般難以分析。因此，在本論文中我們以統計法訓練韻律模型，嚐試將韻律表現拆解成若干個影響因素，期望能在韻律的良好模擬外，更能進一步分析韻律訊息的產生的機制和意義。

1.2 研究方向

本論文研究的重點在於設計一套韻律模型，包括音節基頻軌跡、音節長度、音節能量、音節間的停頓和連音狀態等，再由一套包含大量中文單詞的語料庫訓練韻律模型，並分析各種韻律影響因素的意義。在研究的最後，我們建立一套中文韻律學習系統，做為此韻律模型的應用。

1.3 章節概要

本論文共分為六章：

第一章　緒論：介紹本論文之研究動機與方向。

第二章　中文單詞語料庫介紹：說明如何建立語料庫，以及如何求取各種韻律參數，如音節基頻軌跡、長度、能量等。

第三章　音節基頻軌跡模型：說明如何設計音節基頻軌跡模型，並分析各種影響因素和模擬結果。

第四章　音節長度和能量模型：說明如何設計音節長度、能量、音節間停頓和連音狀態等模型，並分析各種影響因素和模擬結果。

第五章　中文韻律學習系統：運用本論文的韻律模型建立一套應用系統，供非母語的使用者學習中文韻律。

第六章　結論與未來展望。

第二章 中文單詞語料庫介紹

在進行語音韻律模擬之前，我們需要準備一套中文單詞語料庫，而語料庫品質的好壞將直接影響韻律模擬的效能。在本章中我們探討語料庫適用於韻律模擬與否的標準，並且說明本論文語料庫的來源與錄製方法，最後則討論韻律參數(含音節基頻軌跡、長度和能量等)的求取方法。

2.1 適用於訓練模型之語料庫條件

一個語料庫是否適用於訓練韻律模型，重點在於其單元的多樣化。以中文而言，字(Character)的類別約有 12,000 餘種；若以發聲方式來區分，約有 1,300 種帶聲調的音節(Tonal syllable)；若不考慮聲調的類別，則只有 411 種基本音節(Basesyllable)。

一般認為適用於韻律模擬的語料庫應同時具有「豐富語音」(Phonetically rich)和「豐富韻律」(Prosodically rich)兩個特性。所謂豐富語音是指語料庫具有各種不同音節連接的組合；而豐富韻律則是指語料庫具有多種不同的韻律變化。

2.2 語料庫來源與錄製音檔

本論文使用的語料庫文字部分來自於交通大學語音實驗室「文句分析器辭典」，選擇的條件以聲調的平衡為主，共有 107,936 個單詞或 277,218 個字，分別錄製成 4,321 個音檔。在本節中我們將介紹語料庫文字內容的萃取、統計和錄製語料的環境。

2.2.1 文字內容萃取與統計

首先我們將文句分析器辭典的單詞取出，格式如下：

公司　因此　今年　單位　他們

接著我們分別標記單詞中每個音節的聲碼、音節在詞中的位置(Syllable position in word)和詞性(Part of speech, POS)；其中聲碼由四個數字組成，第一個數字代表聲調(1~5)，後三個數字則為基本音節類別(1~411)；音節在詞中的位置由三個數字組成，[i 0 j]表示該字為 i 字詞中的第 j 個字；詞性在本論文並未用到，故不討論。我們將資訊整理成表格如下：

公	1380	201	12
司	1007	202	12
今	1270	201	16
年	2264	202	16
因	1269	201	40
此	3006	202	40

我們統計整個語料庫的單詞數量，整理於表 2-1 和圖 2-1：

表 2-1：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

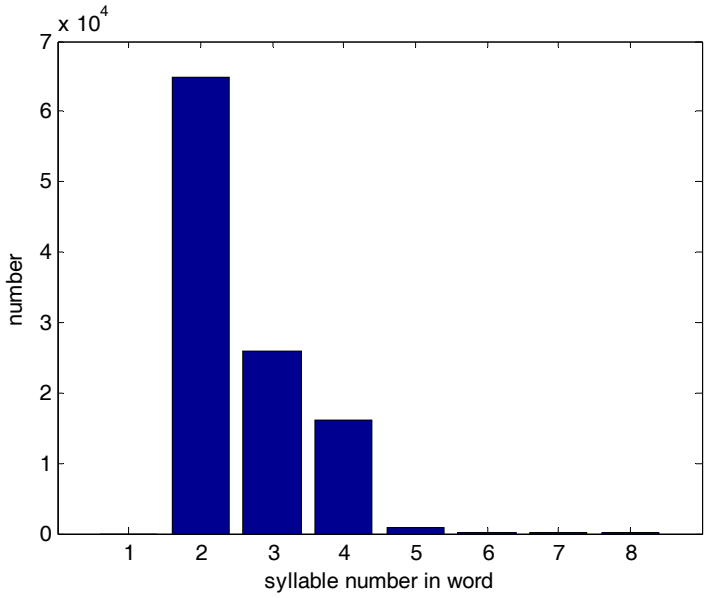


圖 2-1：語料庫詞長分佈圖

2.2.2 錄製音檔

產生文字內容後，我們準備錄製音檔。我們請一位專業的女性廣播人員幫我們錄製音檔，為了能精確模擬語音韻律，我們力求錄音品質的完美，若在錄音過程中有發生口吃、遲疑停頓或唸錯的情況，則將該段重新錄製直到完全正確為止。錄音的環境設定如表 2-2。

表 2-2：錄音環境設定

錄音軟體	Cool Edit Pro 直接錄成聲音檔案
麥克風	單一指向性 (Uni-directional)
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	20kHz
發音速度	每秒約 3.5 個音節
取樣大小	16 位元 (Bits)
聲道	單聲道 (Mono)
檔案格式	PCM

2.3 韻律參數資料庫置

完成語料庫的錄製後，我們對音檔求取語音的韻律參數，以供本論文的模擬使用。在本節中我們先說明音節切割的方法，然後分別討論求取音節的基頻軌跡、長度、能量和其它韻律參數資訊的方法。

2.3.1 音節切割資訊

在錄完音檔後，我們雖然有語音和文字的資料，卻不知道二者之間相對應的位置。此時我們以 HTK(Hidden Markov Model Toolkit)軟體[4]的「強制切割」(Forced alignment)功能對音檔進行切割，細節如下：

首先我們將音框大小(Frame size)和音框位移(Frame shift)分別設定為 20 毫秒(ms)和 5 毫秒，對音檔抽取 12 維的梅爾倒頻譜參數(Mel-frequency cepstral coefficient, MFCC)和能量，組成 13 維的特徵參數向量(Feature vector)，用以描述語音訊號的特性；接著我們以 HTK 中的“Isolated Word Style Training”方式以特徵參數為每種音節的聲母(Initial)和韻母(Final)訓練隱藏式馬可夫模型(Hidden Markov Model, HMM)，反覆訓練直到收斂。接下來我們以 HTK 中的“HVite”指令，利用維特比搜尋(Viterbi search)演算法在音檔中的所有音節聲母和韻母序列中找到相似度(Likelihood)最大的對應位置，以則得到我們要的切割資訊。然而以電腦進行自動切割難免有誤差，我們最後以人工修正較嚴重的錯誤。

2.3.2 求取音節基頻軌跡資訊

音節基頻軌跡(F0 contour)被認為是最重要的韻律參數，我們先以 Wavesurfer 軟體所提供的 ESPS(Entropic Signal Processing System)演算法求出每個音框的基頻數值，再以程式修正，分別去除基頻求取的倍頻(Double pitch)、半頻(Half pitch)、語音開頭和結尾時不穩(On-set & Off-set)等現象。圖 2-2 為 Wavesurfer 軟體的介面。

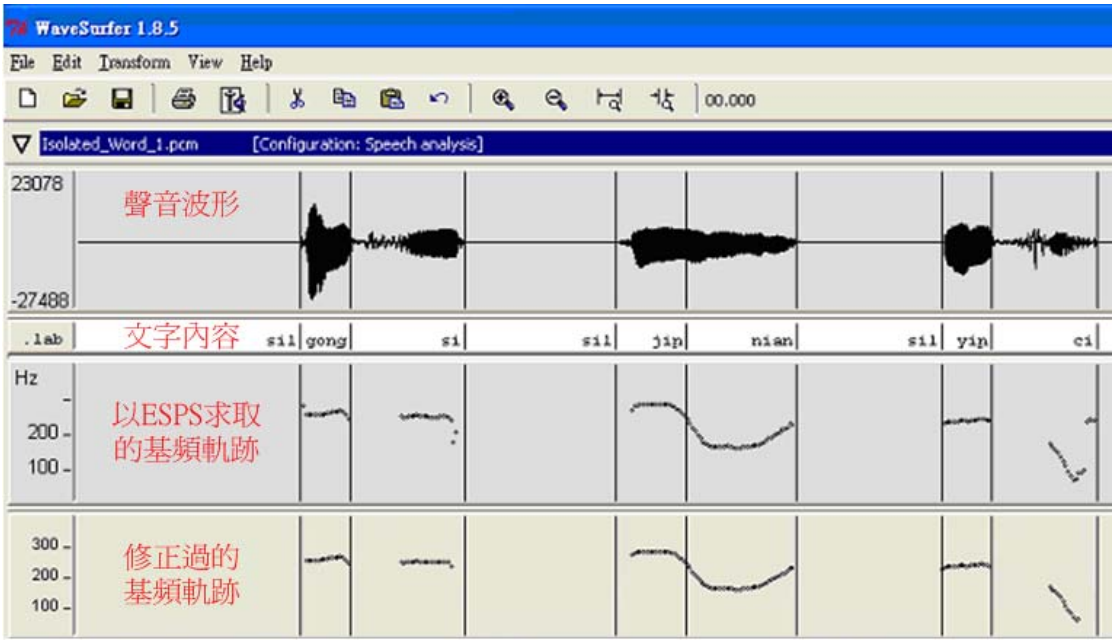


圖 2-2：Wavesurfer軟體介面，其中範例文字分別為「公司、今年、因此」。

此外，由於 Wavesurfer 軟體求取出的基頻數值是以音框為單位，對不同長度的音節其基頻軌跡亦不等長。為了統一量化和方便模擬，我們採用前人的方法將音節基頻軌跡進行正交展開(Orthogonal expansion)，投影到四個勒讓德多項式(Legendre polynomial)基底，以得到四維正交參數[5]。數學式如下：

$$a_j = \frac{1}{N+1}\sum_{i=0}^{N} f(i)\cdot\phi_j\left(\frac{i}{N}\right) \quad \text{for } j = 0,1,2,3 \tag{2-1}$$

其中 $f(i)$ 為原始的基頻軌跡，$0 \le i \le N$ ； $N+1$ 為基頻軌跡的長度； $a_0 \sim a_3$ 為四維正交參數； $\phi_j\left(\frac{i}{N}\right)$ 為四維正交的勒讓德多項式基底，分別如下：

$$\phi_0\left(\frac{i}{N}\right) = 1 \tag{2-2}$$

$$\phi_1\left(\frac{i}{N}\right) = \left[\frac{12\cdot N}{(N+2)}\right]^{\frac{1}{2}}\cdot\left[\left(\frac{i}{N}\right) - \frac{1}{2}\right] \tag{2-3}$$

$$\phi_2\left(\frac{i}{N}\right) = \left[\frac{180\cdot N^3}{(N-1)(N+2)(N+3)}\right]^{\frac{1}{2}}\cdot\left[\left(\frac{i}{N}\right)^2 - \left(\frac{i}{N}\right) + \frac{N-1}{6\cdot N}\right] \tag{2-4}$$

$$\phi_3\left(\frac{i}{N}\right)=\left[\frac{2800N^5}{(N-1)(N-2)(N+2)(N+3)(N+4)}\right]^{\frac{1}{2}} \cdot\left[\left(\frac{i}{N}\right)^3-\frac{3}{2}\left(\frac{i}{N}\right)^2+\frac{6N^2-3N+2}{10N^2}\left(\frac{i}{N}\right)-\frac{(N-1)(N-2)}{20N^2}\right] \tag{2-5}$$

求得四維參數後，我們可以用下式還原基頻軌跡：

$$f^{'}(i)=\sum_{j=0}^{3}a_j\cdot\phi_j\left(\frac{i}{N}\right),\ \text{for } 0\le i\le N \tag{2-6}$$

值得注意的是，在本論文中我們將模擬的韻律參數設定為對數基頻軌跡(Log-F0 contour)，如此並不會影響模擬的數學式，且更符合物理意義。

2.3.3 求取音節長度資訊

由 2.3.1 小節中我們使用 HTK 軟體對音檔進行切割，得到以聲母和韻母為單位的切割單元。接著我們將聲母和韻母合併成音節，即可得依切割結果計算得到音節的長度資訊，如圖 2-3 所示：

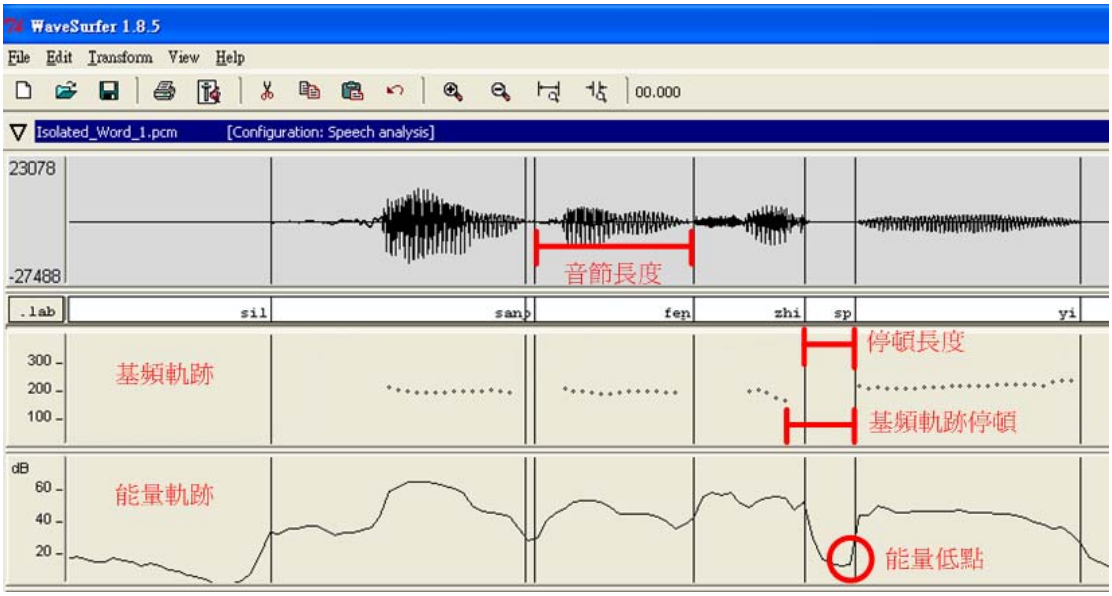


圖 2-3：音節基頻軌跡、長度、能量、停頓長度、基頻軌跡停頓和能量低點示意圖。其中範例文字為「三分之一」。

2.3.4 求取音節能量資訊

由 Wavesurfer 軟體我們可求出所有音節的能量軌跡(Energy contour)，但在本論文中我們模擬的目標為「韻母部分的能量位準最大值」(Maximum energy in final)，因此我們另外寫程式將此值抽取出。

此外，本論文中的能量單位為 dB，求取公式如下：

$$E = 10\log_{10}\frac{\sum(w_i x_i^2)}{N} \tag{2-7}$$

其中 x 為語音訊號；w 為漢明窗(Hamming window)的值；N 為音框大小，在本論文中設為 240 個取樣點；i 表示該音框中語音訊號和漢明窗的索引編號(Index)。

2.3.5 求取其它韻律參數資訊

除了音節基頻軌跡、長度和能量外，在本論文的模擬中我們仍會用到另外四種輔助韻律參數，分述如下：

一、**能量低點(Energy dip)**：二個音節中間的能量最低點，通常能量低點愈低彼此的影響愈弱；可參考圖 2-3。

二、**基頻軌跡停頓(F0 contour pause)**：二個音節中間基頻軌跡暫停的時間長度，通常基頻軌跡停頓愈長彼此的影響愈弱；參考圖 2-3。

三、**正規化基頻差(Normalized F0 jump)**：二個音節減去其聲調影響因素後之基頻軌跡平均值的差距，通常正規化基頻差愈大連音愈弱；聲調影響因素即為語料庫中屬於各種聲調的音節基頻軌跡平均值，後面有更進一步的介紹。

四、**停頓長度(Pause duration)**：二個音節在時間軸上語音訊號的停頓，通常停頓長度愈長彼此的影響愈弱；可參考圖 2-3。

第三章 音節基頻軌跡模型

音節基頻軌跡反應人說話音調的高低起伏變化。在中文裡，字的聲調(Lexical tone)表現在基頻軌跡中；而英文語音的音調高低則能傳達語者說話的重點所在。因此基頻軌跡在所有韻律參數中最重要，亦為影響聽覺舒適度最多的韻律參數。

3.1 模型設計

在本論文中，我們考慮三種影響音節基頻軌跡的因素：聲調、前後音節的連音影響(Coarticulation effect)和音節在詞中的位置。我們假設此三種影響因素(Affecting factor)彼此互相獨立且具加成性，並且對音節在詞首、詞中及詞尾三種不同位置將模型略做修改，如下：

$$\begin{cases} \mathbf{sp}_1 = \mathbf{sp}_1^r + \boldsymbol{\beta}_{t_1} + \boldsymbol{\beta}_{t_1}^f + \boldsymbol{\beta}_{c_1,tp_1}^b + \boldsymbol{\beta}_{w_1} + \boldsymbol{\mu}^p & \text{for } n=1 \\ \mathbf{sp}_n = \mathbf{sp}_n^r + \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{c_{n-1},tp_{n-1}}^f + \boldsymbol{\beta}_{c_n,tp_n}^b + \boldsymbol{\beta}_{w_n} + \boldsymbol{\mu}^p & \text{for } 2 \le n \le N_k - 1 \\ \mathbf{sp}_N = \mathbf{sp}_N^r + \boldsymbol{\beta}_{t_N} + \boldsymbol{\beta}_{c_{N-1},tp_{N-1}}^f + \boldsymbol{\beta}_{t_N}^b + \boldsymbol{\beta}_{w_N} + \boldsymbol{\mu}^p & \text{for } n=N_k \end{cases} \quad (3\text{-}1)$$

參數的說明如下：

N_k：第 k 個詞的音節總數，$k \in (1,2,...,107936)$

$\mathbf{sp}_n$：第 n 個音節的基頻軌跡(Observed F0 contour)四維正交參數(取對數)

$\mathbf{sp}_n^r$：第 n 個音節的殘餘(Residual)基頻軌跡四維正交參數(取對數)

$\boldsymbol{\beta}_{t_n}$：第 n 個音節的聲調影響因素(Tone affecting factor)，$t_n \in (1,2,3,4,5)$

c_n：第 n 個音節和第 n+1 個音節的連音狀態(Coarticulation state)，$c_n \in (1,2,3)$

tp_n：第 n 個音節和第 n+1 個音節的聲調組合(Tone pair)，$tp_n \in \{(1,1),(1,2),...,(5,5)\}$

$\boldsymbol{\beta}_{c_{n-1},tp_{n-1}}^f$：第 n-1 個音節與第 n 個音節之間連音狀態為 c_{n-1}、音節組合為 tp_{n-1} 時，

第 n 個音節受到第 n-1 個音節向前(Forward)影響的因素

$\boldsymbol{\beta}^{f}_{t_1}$：詞首的影響因素，為 $\boldsymbol{\beta}^{f}_{c_{n-1},tp_{n-1}}$ 在詞首的特例

$\boldsymbol{\beta}^{b}_{c_n,tp_n}$：第 n 個音節與第 n+1 個音節之間連音狀態為 c_n、音節組合為 tp_n 時，第 n 個音節受到第 n+1 個音節向後(Backward)影響的因素

$\boldsymbol{\beta}^{b}_{t_N}$：詞尾的影響因素，為 $\boldsymbol{\beta}^{b}_{c_n,tp_n}$ 在詞尾的特例

$\boldsymbol{\beta}_{w_n}$：第 n 個音節在多字詞中位置的影響因素，$w_n \in \{(2,1),(2,2),...,(i,j),...,(8,8)\}$，其中 (i, j) 表示 i 字詞中的第 j 個字

$\boldsymbol{\mu}^{p}$：所有語料的音節基頻軌跡平均值(Global mean)

在本論文中為了表達簡潔，我們將(3-1)式重寫如下：

$$\mathbf{sp}_n = \mathbf{sp}^{r}_n + \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}^{f}_{c_{n-1},tp_{n-1}} + \boldsymbol{\beta}^{b}_{c_n,tp_n} + \boldsymbol{\beta}_{w_n} + \boldsymbol{\mu}^{p} \quad \text{for } 1 \le n \le N \tag{3-2}$$

其中詞首(n =1)的 $\boldsymbol{\beta}^{f}_{c_{n-1},tp_{n-1}}$ 為原式中的 $\boldsymbol{\beta}^{f}_{t_1}$；而詞尾($n$=$N$)的 $\boldsymbol{\beta}^{b}_{c_n,tp_n}$ 為原式中的 $\boldsymbol{\beta}^{b}_{t_N}$。我們可將基頻軌跡模型的幾個影響因素關係表示如圖 3-1。

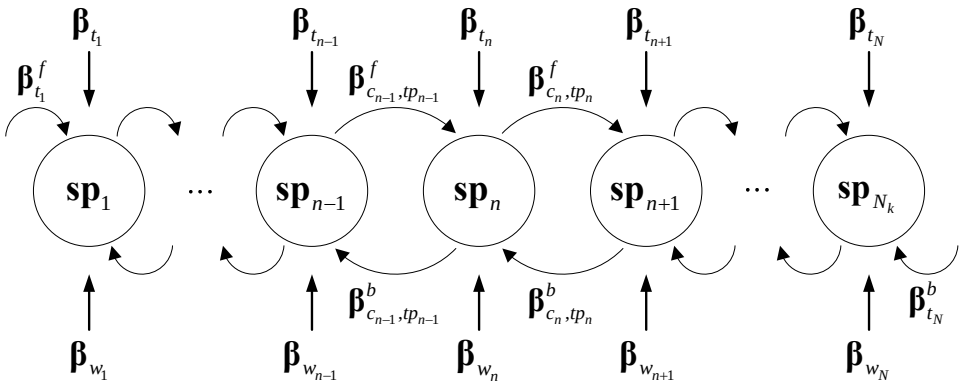


圖 3-1：音節基頻軌跡模型之影響因素關係

3.2 模型訓練

在此論文中我們採用的模型為數據驅動(Data-driven)，模型的參數需要由大量的語音資料訓練得到；在此節中我們討論模型的訓練方法。

3.2.1 影響因素推導

由(3-2)式中可知，音節的基頻軌跡 $\mathbf{sp}_n$ 是由 $\boldsymbol{\beta}_{t_n}$ 、$\boldsymbol{\beta}^{f}_{c_{n-1},tp_{n-1}}$ 、$\boldsymbol{\beta}^{b}_{c_n,tp_n}$ 、$\boldsymbol{\beta}_{w_n}$ 和 $\boldsymbol{\mu}^{p}$ 等影響因素所相加組成，在給定音節的聲調、前後音節的連音狀態、音節在詞中的位置和聲調組合等資訊的情況下，我們可預測出該音節的基頻軌跡(Predicted F0 contour)；而其誤差即為 $\mathbf{sp}^{r}_{n}$ 。我們假設此誤差 $\mathbf{sp}^{r}_{n}$ 呈高斯分佈(Gaussian distribution)，可寫成下面數學式：

$$P(\mathbf{sp}_n \mid t_n, c_{n-1}, tp_{n-1}, c_n, tp_n, w_n) = N(\mathbf{sp}_n; \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}^{f}_{c_{n-1},tp_{n-1}} + \boldsymbol{\beta}^{b}_{c_n,tp_n} + \boldsymbol{\beta}_{w_n} + \boldsymbol{\mu}^{p}, \mathbf{R}^{p}) \qquad (3\text{-}3)$$

其中 $\mathbf{R}^{p}$ 為誤差 $\mathbf{sp}^{r}_{n}$ 的共變異矩陣(Covariance matrix)。

在本研究中，我們採用逐項最佳化程序(Sequential Optimization Procedure)和最大相似度法則(Maximum Likelihood Criterion)來訓練及更新模型參數。首先我們定義對數相似度函數(Log-likelihood function)如下：

$$L = \sum_{k=1}^{K} \sum_{n=1}^{N_k} \log N(\mathbf{sp}_{k,n}; \mathbf{sp}_{t_{k,n}} + \mathbf{sp}^{f}_{c_{k,n-1},tp_{k,n-1}} + \mathbf{sp}^{b}_{c_{k,n},tp_{k,n}} + \mathbf{sp}_{w_{k,n}} + \boldsymbol{\mu}^{p}, \mathbf{R}^{p}) \qquad (3\text{-}4a)$$

其中 N_k 為單詞中音節的數目；$K = 107{,}936$ 為所有語料庫中單詞的數目；$\mathbf{sp}_{k,n}$ 為語料庫中第 k 個單詞中第 n 個音節的基頻軌跡四維正交參數；其餘參數和(3-1)式相似。在本論文中為了表達簡潔，我們將(3-4a)式重寫如下：

$$L = \sum_{n=1}^{N_{all}} \log N(\mathbf{sp}_n; \mathbf{sp}_{t_n} + \mathbf{sp}^{f}_{c_{n-1},tp_{n-1}} + \mathbf{sp}^{b}_{c_n,tp_n} + \mathbf{sp}_{w_n} + \boldsymbol{\mu}^{p}, \mathbf{R}^{p}) \qquad (3\text{-}4b)$$

其中 $N_{all}=\sum_{k}^{K}\sum_{n}^{N_k}1=277,218$，即為整個語料庫的音節總數。接著，我們依照最大相似度法則可推導出模型參數的訓練與更新數學式：

$$\boldsymbol{\beta}_t=\frac{\sum_{n=1}^{N_{all}}\left(\mathbf{sp}_n-\boldsymbol{\beta}^f_{c_{n-1},tp_{n-1}}-\boldsymbol{\beta}^b_{c_n,tp_n}-\boldsymbol{\beta}_{w_n}-\boldsymbol{\mu}^p\right)\delta\left(t_n=t\right)}{\sum_{n=1}^{N_{all}}\delta\left(t_n=t\right)} \tag{3-5}$$

$$\boldsymbol{\beta}^f_{c,tp}=\frac{\sum_{n=1}^{N_{all}}\left(\mathbf{sp}_n-\boldsymbol{\beta}_{t_n}-\boldsymbol{\beta}^b_{c_n,tp_n}-\boldsymbol{\beta}_{w_n}-\boldsymbol{\mu}^p\right)\delta\left(c_{n-1}=c,tp_{n-1}=tp\right)}{\sum_{n=1}^{N_{all}}\delta\left(c_{n-1}=c,tp_{n-1}=tp\right)} \tag{3-6}$$

$$\boldsymbol{\beta}^b_{c,tp}=\frac{\sum_{n=1}^{N_{all}}\left(\mathbf{sp}_n-\boldsymbol{\beta}_{t_n}-\boldsymbol{\beta}^f_{c_{n-1},tp_n}-\boldsymbol{\beta}_{w_n}-\boldsymbol{\mu}^p\right)\delta\left(c_n=c,tp_n=tp\right)}{\sum_{n=1}^{N_{all}}\delta\left(c_n=c,tp_n=tp\right)} \tag{3-7}$$

$$\boldsymbol{\beta}_w=\frac{\sum_{n=1}^{N_{all}}\left(\mathbf{sp}_n-\boldsymbol{\beta}_{t_n}-\boldsymbol{\beta}^f_{c_{n-1},tp_{n-1}}-\boldsymbol{\beta}^b_{c_n,tp_n}-\boldsymbol{\mu}^p\right)\delta\left(w_n=w\right)}{\sum_{n=1}^{N_{all}}\delta\left(w_n=w\right)} \tag{3-8}$$

$$\mathbf{R}^p=\frac{\sum_{n=1}^{N_{all}}\mathbf{Y}_n\mathbf{Y}_n^{\mathrm{T}}}{N_{all}} \tag{3-9}$$

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

3.2.2 影響因素初始值

在前一小節我們討論了參數訓練及更新的方法，但在一開始時我們仍需要一個初始模型(Initial model)。一個好的初始模型具有物理意義，它不但應該符合我們對語言學的認知，還要能幫助模型快速收斂。

一般公認聲調的影響因素(Lexical tone affecting factor)最重要，且不同聲調的影響差異最明顯，因此適合當作第一個初始化(Initialization)的影響因素：

$$\boldsymbol{\beta}_t = \frac{\sum_{n=1}^{N_{all}} \left(\mathbf{sp}_n - \boldsymbol{\mu}^p\right)\delta\left(t_n = t\right)}{\sum_{n=1}^{N_{all}} \delta\left(t_n = t\right)} \tag{3-10}$$

接著我們考慮前後音節的連音影響，包括向前和向後的影響因素(Forward and backward affecting factor)；我們把前後音節聲調組合的影響減去該音節自己的聲調當作其初始值：

$$\boldsymbol{\beta}^f_{c,tp} = \frac{\sum_{n=1}^{N_{all}} \mathbf{sp}_n \delta\left(c_{n-1} = c, tp_{n-1} = tp\right)}{\sum_{n=1}^{N_{all}} \delta\left(c_{n-1} = c, tp_{n-1} = tp\right)} - \frac{\sum_{n=1}^{N_{all}} \mathbf{sp}_n \delta\left(c_{n-1} = c, t_n = t\right)}{\sum_{n=1}^{N_{all}} \delta\left(c_{n-1} = c, t_n = t\right)} \tag{3-11}$$

$$\boldsymbol{\beta}^b_{c,tp} = \frac{\sum_{n=1}^{N_{all}} \mathbf{sp}_n \delta\left(c_n = c, tp_n = tp\right)}{\sum_{n=1}^{N_{all}} \delta\left(c_n = c, tp_n = tp\right)} - \frac{\sum_{n=1}^{N_{all}} \mathbf{sp}_n \delta\left(c_n = c, t_n = t\right)}{\sum_{n=1}^{N_{all}} \delta\left(c_n = c, t_n = t\right)} \tag{3-12}$$

有了音節和連音現象的影響因素後，音節在詞中位置的影響因素 $\boldsymbol{\beta}_w$ (Syllable-position-in-word affecting factor)的初始值可直接代入(3-8)式得到。

除了 $\boldsymbol{\beta}_t$ 、$\boldsymbol{\beta}^f_{c,tp}$ 、$\boldsymbol{\beta}^b_{c,tp}$ 和 $\boldsymbol{\beta}_w$ 外，我們亦需要每二個音節間連接處(Juncture)的連音狀態初始值資訊，在本論文中我們以 2.3.5 小節中提到的四種輔助韻律參數作為初始標準：首先將二個音節間「基頻軌跡停頓」長度為 0 的連音狀態初始值設為強(State1)，因為我們發現這一類的音節間連接處的「停頓長度」一定也是 0，「能量低點」偏高，皆符合連音狀態為「強」的條件。接著我們將剩下的音節間連接處以向量量化(Vector quantization, VQ)的方式分成二類，再將「能量低點」較低的那一群的連音狀態初始值定為「弱」(State3)，「能量低點」較高的那一群的連音狀態初始值定為「中」(State2)；連音狀態初始值示意圖如圖 3-2。

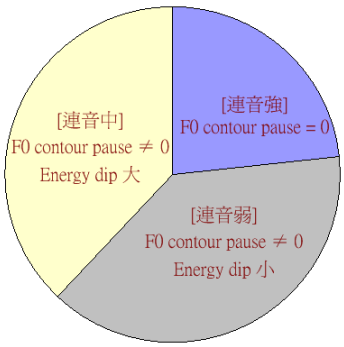


圖 3-2：音節間連音狀態初始值示意圖

3.2.3 訓練流程

如前面所述，本研究採用逐項最佳化程序來訓練和更新各個影響因素的數值，順序分別為$\boldsymbol{\beta}_t$、$\boldsymbol{\beta}^f_{c,tp}$、$\boldsymbol{\beta}^b_{c,tp}$、$\boldsymbol{\beta}_{w_n}$和$\mathbf{R}^p$。值得注意的是，在更新$\mathbf{R}^p$之前我們亦對連音狀態"$c$"進行重新標記(Re-label)，流程如下：

一、定義目標函數(Objective function)：

除了原本(3-4)式之對數相似度函數外，我們重新標記連音狀態時亦考慮每二個音節連接處的四種輔助韻律參數，並分別將「能量低點」、「基頻軌跡停頓」和「正規化基頻差」模擬為高斯分佈$N(ed_n;\mu^{ed}_{c_n},\sigma^{ed2}_{c_n})$、$N(pp_n;\mu^{pp}_{c_n},\sigma^{pp2}_{c_n})$和$N(pj_n;\mu^{pj}_{c_n},\sigma^{pj2}_{c_n})$，且將「停頓長度」模擬伽瑪分佈$G(pd_n;\alpha^{pd}_{c_n},\beta^{pd}_{c_n})$。此外值得注意的是，在第二章我們求取「基頻軌跡停頓」和「停頓長度」時是以音框為單位(Frame-based)，因此數值為離散型(Discrete)；為了模擬真實情況，我們另外加上了偽隨機數(Pseudo-random number)使這二項韻律參數變成連續型(Continuous)。

我們將目標函數定義如下：

$$L' = \sum_{n=1}^{N_{all}} \log N(\mathbf{sp}_n;\boldsymbol{\beta}_{t_n}+\boldsymbol{\beta}^f_{c_{n-1},t_{n-1}}+\boldsymbol{\beta}^b_{c_n,t_n}+\boldsymbol{\beta}_{w_n}+\boldsymbol{\mu}^p,\mathbf{R}^p)$$
$$+\sum_{n=1}^{N_{all}-1}\log[N(ed_n;\mu^{ed}_{c_n},\sigma^{ed2}_{c_n})N(pp_n;\mu^{pp}_{c_n},\sigma^{pp2}_{c_n})N(pj_n;\mu^{pj}_{c_n},\sigma^{pj2}_{c_n})G(pd_n;\alpha^{pd}_{c_n},\beta^{pd}_{c_n})] \quad (3\text{-}13)$$

二、將目標函數最大化：

在進行重新標記時，我們以單詞為單位找出最佳的連音狀態序列(Coarticulation state sequence)，使得目標函數L'最大化，數學式如下：

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}}(L'), \quad \mathbf{c} = c_1, c_2, ..., c_{N_{all}-1} \tag{3-14}$$

為了降低計算量和計算時間，我們採用維特比搜尋演算法來進行重新標記。

我們以疊代(Iteration)的方式更新所有的影響因素、連音狀態及共變異矩陣，直到目標函數收斂為止。在此我們定義收斂條件如下：

$$\frac{L'_m - L'_{m-1}}{L'_{m-1}} \le 10^{-7} \tag{3-15}$$

其中L'_m表式第m次疊代的目標函數值。整個訓練流程可參考圖 3-3；而四種輔助韻律參數的初始和收斂數值的分佈請參考附錄。

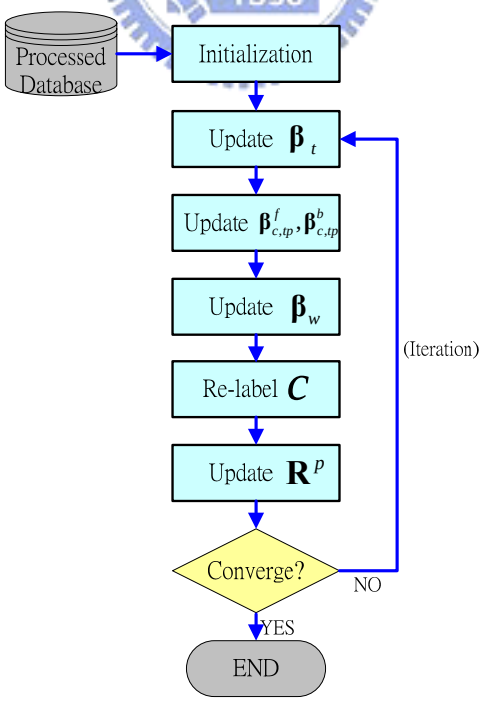


圖 3-3：音節基頻軌基模型參數訓練和更新流程

3.3 模擬結果與分析

在此節中，我們將將分析訓練至收斂的模型參數，並且以訓練好的模型預測音節基頻軌跡。由圖 3-4 可看出，疊代訓練一共重覆了 29 次。

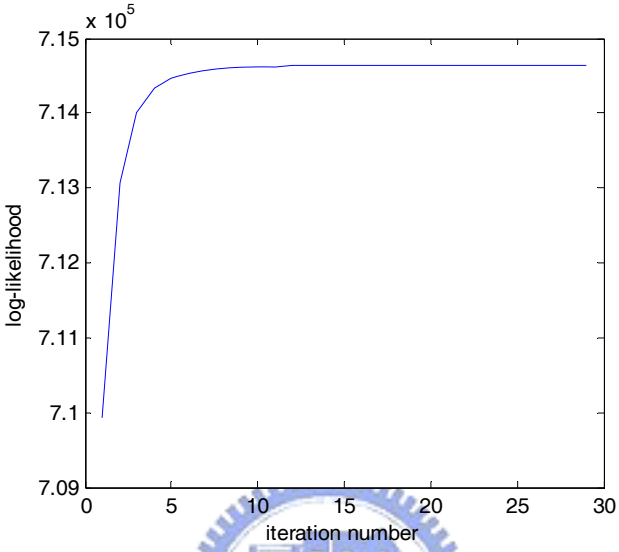


圖 3-4：音節基頻軌跡模型訓練疊代次數及其目標函數值

3.3.1 聲調之影響因素

聲調之影響因素是中文單詞基頻軌跡最重要的影響因素。圖 3-5 為五種聲調的基頻軌跡影響因素，我們將基頻軌跡的長度統一正規化至同樣長度以方便觀察；若其值大於 0 表示該影響因素會拉高基頻軌跡，反之則會降低基頻軌跡。由實驗結果我們發現，一聲整體偏高；二聲約略為由低轉高，但在起始處有稍微抬升；三聲為由中轉低；四聲則為由高轉低；五聲(輕聲)在語音上通常會隨著前面音節的聲調改變，無特殊軌跡形狀。

值得注意的是，三聲在傳統中文文法的觀念為先下降再上升[6]，而在我們模擬的結果卻只有下降而無上升。我們猜測此為台灣人口音習慣所致。為此我們實際錄製許多三聲的音節，發現除非刻意將聲音上揚或學北京腔字正腔圓地唸，

否則多數的基頻軌跡都和圖 3-5 的結果相仿。我們發現在前人的研究中亦有類似的現象[7]。

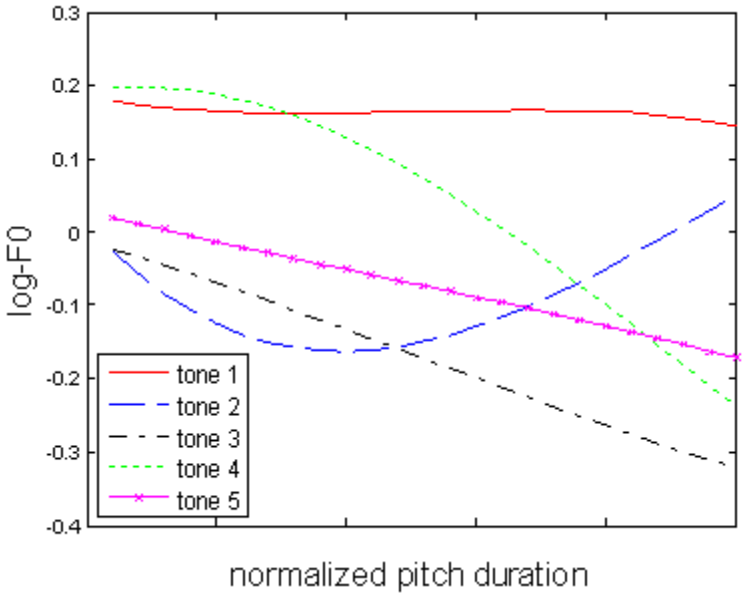


圖 3-5：音節基頻軌跡模型中聲調之影響因素

3.3.2 受前一音節影響因素

中文音節的基頻軌跡會受到前後音節的影響，在前人的研究中著墨甚多[8]。我們將 25 種可能的音節聲調組合畫出來，再以 3 種不同的線表示不同的連音狀態的影響。我們發現連音狀態為「強」的影響較大(軌跡離 0 較遠)，而連音狀態「中」和「弱」的影響較小(軌跡在 0 附近)，此結果符合語言學的知識。

以聲調組合為(1,2)的圖為例，由前一小節的結果我們知道前一音節為一聲，基頻軌跡較高；而此音節為二聲，基頻軌跡是由低往高；而當二個音節一起唸時，若二個音節連音狀態為強(實線)時，此音節的二聲顯然受到前一音節一聲的影響，使得此音節的前端會被拉高。反之，若二個音節連音狀態為中(虛線)或弱(點)時則無此現象。由此可知模擬的結果相當合理，且和前人的理論驗證[9]。

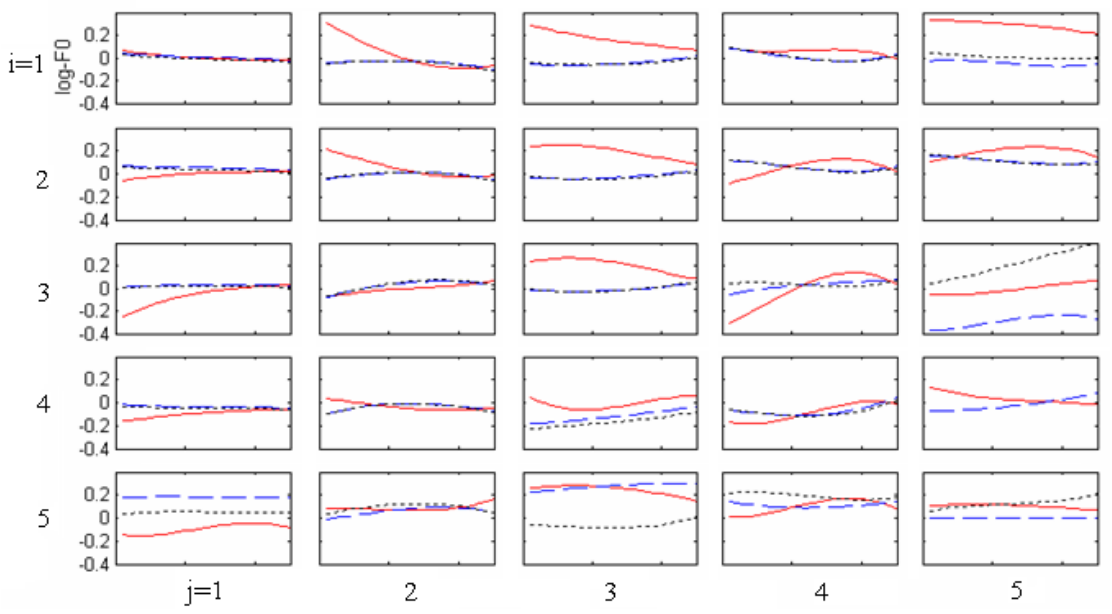


圖 3-6：音節基頻軌跡模型中受前一音節影響因素。其中實線(line)為連音狀態為「強」、虛線(dash)為「中」、點(dots)為「弱」。第(i,j)項表示前一個音節為tone i、此音節為tone j的聲調組合。

如(3-1)式所述，位於詞首的音節並無受前一音節之影響因素 $\boldsymbol{\beta}^{f}_{c_{n-1},tp_{n-1}}$，因此我們以詞首之影響因素 $\boldsymbol{\beta}^{f}_{t_1}$ 與其對應。由圖 3-7 可發現，詞首之影響因素的動態範圍(Dynamic range)和聲調之影響因素相較之下小很多，影響較小。

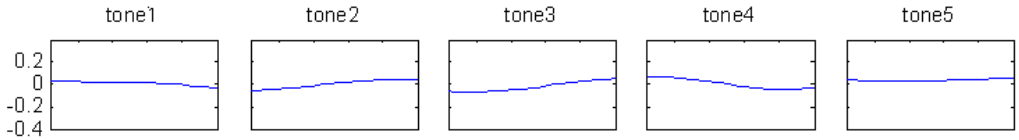


圖 3-7：音節基頻軌跡模型中詞首影響因素。詞首影響因素沒有連音狀態之分別。

3.3.3 受後一音節影響因素

和前一小節相似，音節的基頻軌跡會受到後一音節聲調的影響，其中最著名的例子是當二個連續的音節為三聲接三聲時，第一個音節會變成二聲，此現象稱

為變調(Tone sandhi)。除了變調外，音節的基頻軌跡仍會或多或少受到後一個音節的影響。我們發現，基頻軌跡受後一音節影響的程度較受前一音節影響還小，此現象反應在二種影響因素的動態範圍上。由圖 3-7 我們發現，除了聲調組合(3,3)因為變調有較大的影響外，其它組合的影響較小，基頻軌跡落在 0 的附近。

觀察聲調組合(3,3)，我們發現音節的基頻軌跡在末端會被拉高，這是因為變調使得音節由三聲轉為二聲，而二聲的基頻軌跡末端會向上揚的緣故。

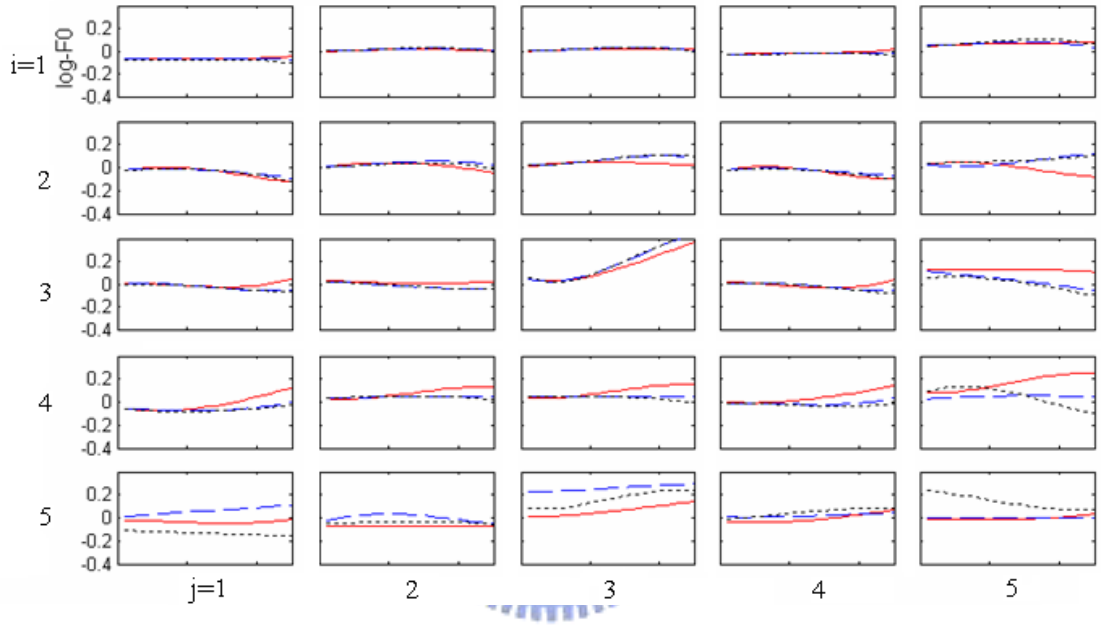


圖 3-8：音節基頻軌跡模型中受後一音節影響因素。其中實線(line)為連音狀態為「強」、虛線(dash)為「中」、點(dots)為「弱」。第(i,j)項表示此音節為tone i、後一音節為tone j的聲調組合。

和前一小節相似，位於詞尾的音節並無受後一音節之影響因素 $\boldsymbol{\beta}^{b}_{c_n,tp_n}$，因此我們以詞首之影響因素 $\boldsymbol{\beta}^{b}_{t_N}$ 與其對應，結果繪於圖 3-9 中。

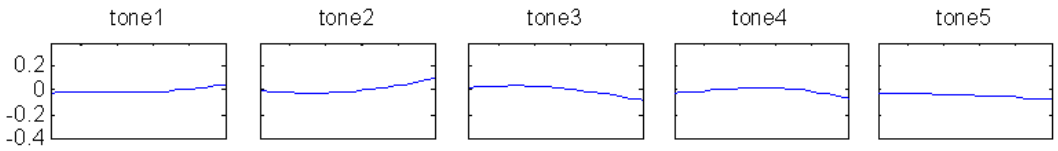


圖 3-9：音節基頻軌跡模型中詞尾影響因素。詞尾影響因素沒有連音狀態之分別。

3.3.4 音節在詞的位置之影響因素

一般認為在語調語言(Intonation language，如英文)中，一句連續語音的基頻軌跡會隨著時間逐漸降低，並且在下一句話(Sentence)或韻律詞(Prosodic word)的開始處躍升，此現象稱為 pitch reset。我們相信在中文單詞中也有類似的現象，並將之模擬為音節在詞的位置之影響因素。

由圖 3-10 我們發現，若音節在詞的前面部分則其基頻軌跡較高，若音節在詞的後面則較低，並且有逐步下降的趨勢。此外我們發現一個有趣的現象：對於同樣是詞首音節的影響因素而言，三字詞首比二字詞首還要高，四字詞首又比三字詞高；至於詞尾音節則剛好相反。

3.3.5 音節基頻軌跡預測

如前面所述，在給定音節的聲調、前後連音資訊、音節在詞中位置等資訊時，我們可由此模型預測音節的基頻軌跡。圖 3-11 為一些實際(Observed)和預測(Predicted)基頻軌跡的例子。

此外，在我們取全部語料的 9/10 做為訓練語料(Training data)和 1/10 做為測試語料(Testing data)。我們將實際和殘餘基頻軌跡的共變異矩陣整理如表 3-1 (內部測試，Inside test)和表 3-2 (外部測試，Outside test)，左邊為實際訊號，右邊為殘餘訊號，可發現共變異矩陣的數值大幅縮小，表示基頻軌跡的模擬效能很好。

最後我們亦將實際訊號和以模型預測的訊號作比較，計算出相關係數(Correlation coefficient)如表 3-3。

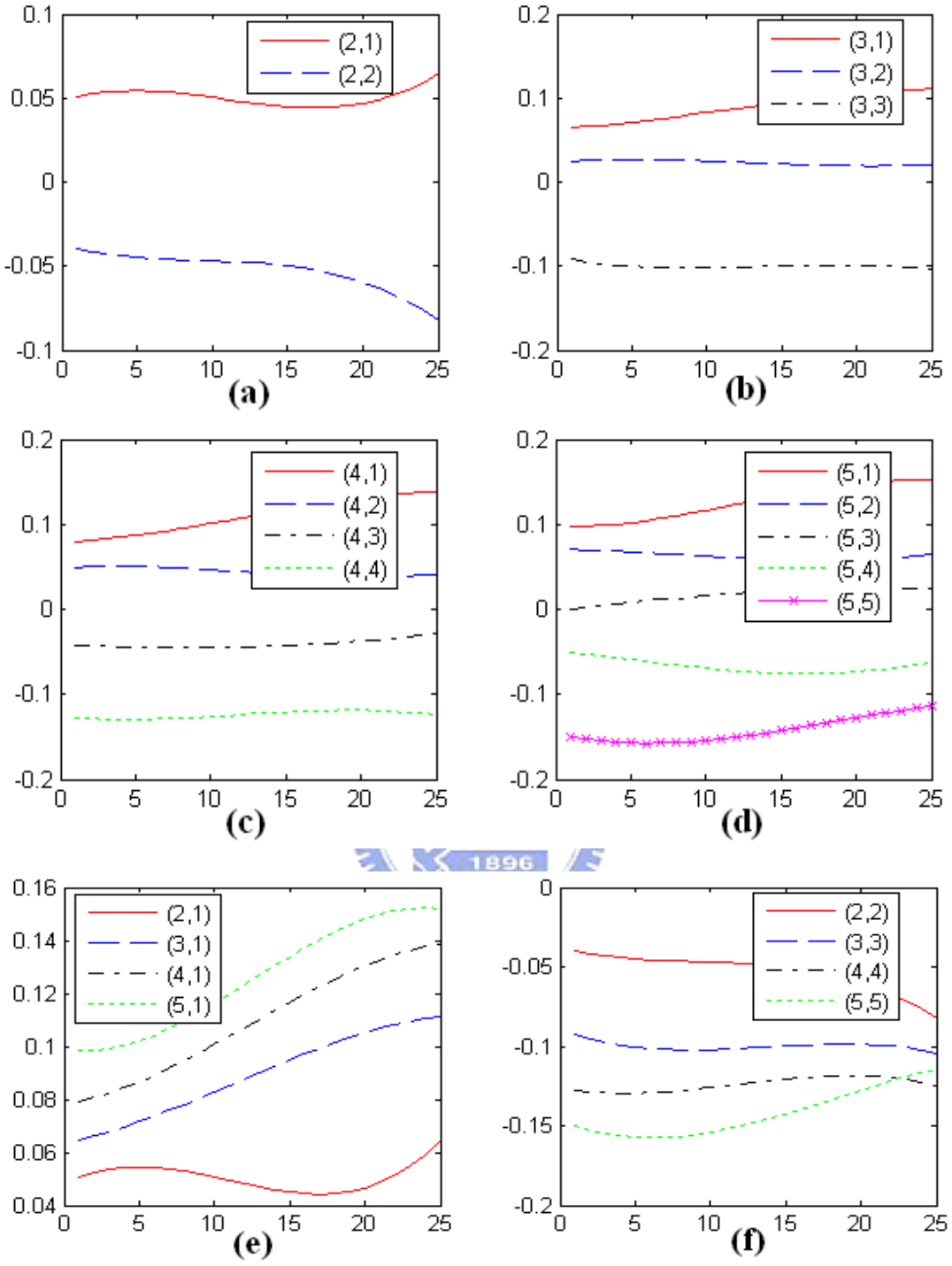


圖 3-10：音節基頻軌跡模型中音節在詞的位置影響因素。其中(a)~(d)分別為二字詞、三字詞、四字詞和五字詞；(e)和(f)為詞首和詞尾的整合比較。圖中的(i,j)代表i字詞第j個字的基頻軌跡影響因素。

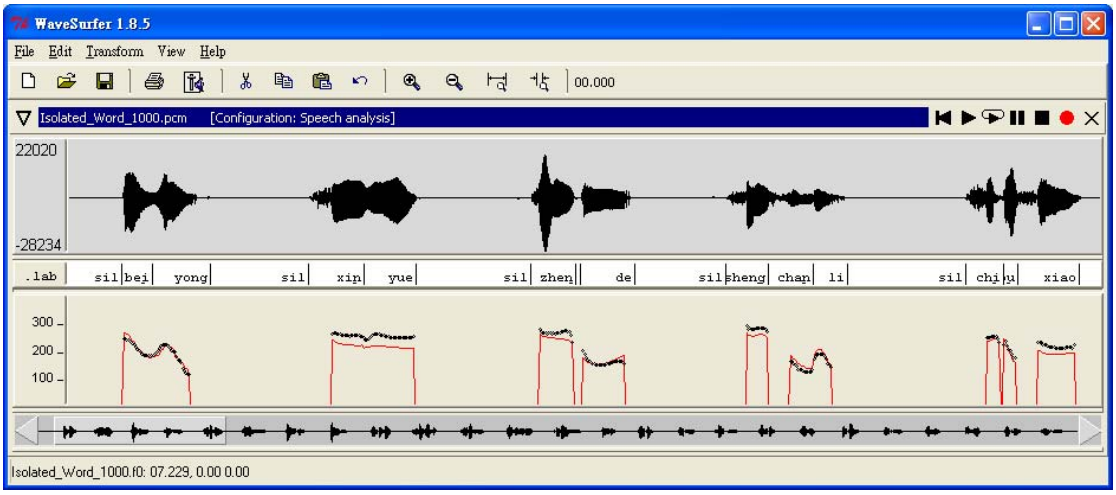


圖 3-11：以模型預測音節基頻軌跡。其中圓點(dots)為實際基頻軌跡，而實線(line)為預測基頻軌跡；文字內容為「備用、新約、貞德、生產力、吃不消」。

表 3-1：實際(左)和殘餘(右)基頻軌跡之正交參數共變異矩陣(Inside test)

$$\begin{bmatrix} \mathbf{0.0311} & 0.0002 & -0.0017 & 0.0001 \\ 0.0002 & \mathbf{0.0099} & 0.0022 & -0.0010 \\ -0.0017 & 0.0022 & \mathbf{0.0020} & -0.0001 \\ 0.0001 & -0.0010 & -0.0001 & \mathbf{0.0004} \end{bmatrix} \Rightarrow \begin{bmatrix} \mathbf{0.0082} & 0.0008 & -0.0003 & -0.0001 \\ 0.0008 & \mathbf{0.0023} & 0.0003 & -0.0002 \\ -0.0003 & 0.0003 & \mathbf{0.0007} & 0.0001 \\ -0.0001 & -0.0002 & 0.0001 & \mathbf{0.0003} \end{bmatrix}$$

表 3-2：實際(左)和殘餘(右)基頻軌跡之正交參數共變異矩陣(Outside test)

$$\begin{bmatrix} \mathbf{0.0313} & 0.0004 & -0.0017 & 0.0001 \\ 0.0004 & \mathbf{0.0099} & 0.0022 & -0.0010 \\ -0.0017 & 0.0022 & \mathbf{0.0021} & -0.0001 \\ 0.0001 & -0.0010 & -0.0001 & \mathbf{0.0004} \end{bmatrix} \Rightarrow \begin{bmatrix} \mathbf{0.0090} & 0.0010 & -0.0003 & -0.0001 \\ 0.0010 & \mathbf{0.0025} & 0.0002 & -0.0002 \\ -0.0003 & 0.0002 & \mathbf{0.0008} & 0.0000 \\ -0.0001 & -0.0002 & 0.0000 & \mathbf{0.0003} \end{bmatrix}$$

表 3-3：實際和以模型預測基頻軌跡四維正交參數之相關係數

	a0	**a1**	**a2**	**a3**
Inside test	0.8519	0.8710	0.7764	0.5305
Outside test	0.8449	0.8657	0.7759	0.5130

3.3.6 模擬誤差分析

最後，我們進一步分析模擬的誤差，分別將多字詞中每個音節的基頻軌跡平均值(F0 contour mean)的模擬誤差求出，再求取該詞中整體誤差的平均值和斜率(Slope)，發現其相關性極小，並無一致性的錯誤。我們將詞中誤差的平均值和斜率繪成散佈圖(Scatter plot)和二維直方圖(2D histogram)以方便觀察，如圖 3-12。

我們亦分析各個影響因素的貢獻度，以「整體殘餘誤差」(Total Residual Error, TRE)表示。整體殘餘誤差的計算方法為實際訊號基頻軌跡扣除影響因素後的變異數，除以實際訊號基頻軌跡的變異數。由表 3-4 可發現，對音節基頻軌跡來說，聲調的影響因素最重要，光是此項即可將誤差減少至 53.9%；其次分別為前後音節的連音影響和音節在詞中的位置。

表 3-4：影響因素與殘餘誤差分析表

影響因素	TRE
+ 聲調	53.9%
+ 連音影響	40.4%
+ 音節在詞中位置	28.9%

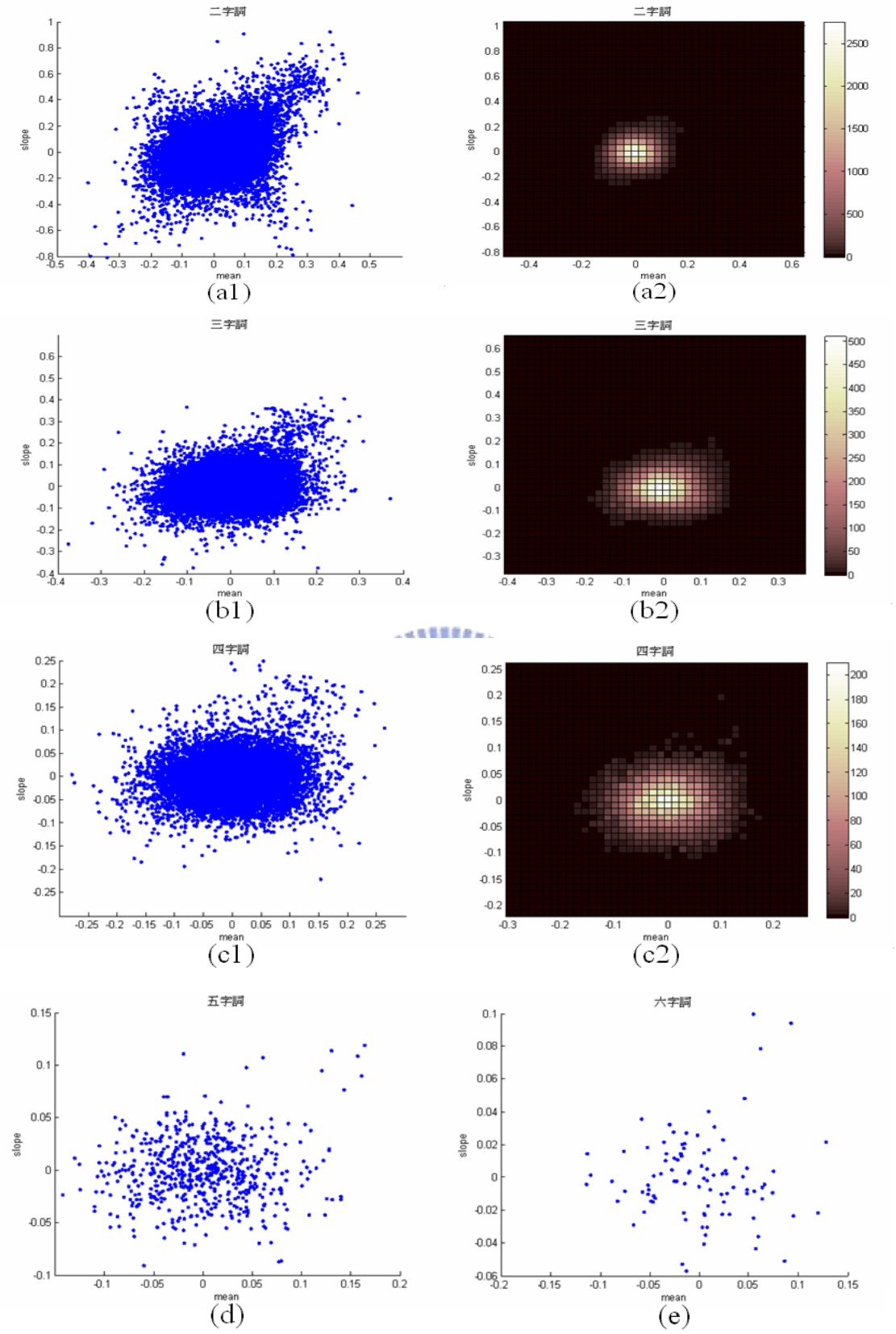


圖 3-12：詞中模擬誤差之平均值與斜率散佈圖與二維直方圖。其中(a)~(e)分別為二字詞至六字詞。五字詞和六字詞因數量較少，故二維直方圖在此省略。

第四章 音節長度和能量模型

在第三章我們討論了音節基頻軌跡模型，而第四章我們繼續討論另二個重要的韻律參數：音節長度(Duration)和能量(Energy)。音節的長度反應出說話速度的快慢，而音節的能量則代表聲音的大小聲，二者都影響聽者的感覺。由於在本研究中音節長度和能量的模型相似，在模型設計與數學推導部分僅介紹音節長度模型，而在模擬結果與分析部分則分開討論。

4.1 模型設計

在音節基頻軌基模型中，我們考慮三種影響因素，分別為聲調、前後音節的連音影響和音節在詞中的位置。在音節長度模型中，我們考慮第四種影響因素：基本音節類別(Basesyllable type)。我們假設此四種影響因素彼此互相獨立且具加成性，並且對音節在詞首、詞中及詞尾三種不同位置將模型略做修改，如下：

$$sd_n = \begin{cases} sd_1^r + \gamma_{t_1} + \gamma_{in_1}^f + \gamma_{c_1,tp_1}^b + \gamma_{w_1} + \gamma_{sy_1} + \mu^d & \text{for } n=1 \\ sd_n^r + \gamma_{t_n} + \gamma_{c_{n-1},tp_{n-1}}^f + \gamma_{c_n,tp_n}^b + \gamma_{w_n} + \gamma_{sy_n} + \mu^d & \text{for } 2 \le n \le N_k - 1 \\ sd_N^r + \gamma_{t_N} + \gamma_{c_{N-1},tp_{N-1}}^f + \gamma_{t_N}^b + \gamma_{w_N} + \gamma_{sy_N} + \mu^d & \text{for } n = N_k \end{cases}$$

(4-1)

參數的說明如下：

N_k ：第 k 個詞的音節總數，$k \in (1,2,...,107936)$

sd_n ：第 n 個音節的長度(Observed duration)

sd_n^r ：第 n 個音節的殘餘長度

γ_{t_n}：第 n 個音節的聲調影響因素，$t_n \in (1,2,3,4,5)$

c_n：第 n 個音節和第 n+1 個音節的連音狀態，$c_n \in (1,2,3)$

tp_n：第 n 個音節和第 n+1 個音節的聲調組合，$tp_n \in \{(1,1),(1,2),...,(5,5)\}$

$\gamma^f_{c_{n-1},tp_{n-1}}$：第 n-1 個音節與第 n 個音節之間連音狀態為 c_{n-1}、音節組合為 tp_{n-1} 時，第 n 個音節受到第 n-1 個音節向前影響的因素

$\gamma^f_{t_1}$：詞首的影響因素，為 $\gamma^f_{c_{n-1},tp_{n-1}}$ 在詞首的特例

$\gamma^b_{c_n,tp_n}$：第 n 個音節與第 n+1 個音節之間連音狀態為 c_n、音節組合為 tp_n 時，第 n 個音節受到第 n+1 個音節向後影響的因素

$\gamma^b_{t_N}$：詞尾的影響因素，為 $\gamma^b_{c_n,tp_n}$ 在詞尾的特例

γ_{w_n}：第 n 個音節在多字詞中位置的影響因素，$w_n \in \{(2,1),(2,2),...,(i,j),...,(8,8)\}$，其中 (i, j) 表示 i 字詞中的第 j 個字

γ_{sy_n}：第 n 個音節的基本音節類別，$sy_n \in (1,2,...,411)$

μ^d：所有語料的音節長度平均值

在本論文中為了表達簡潔，我們將(4-1)式重寫如下：

$$sd_n = sd_n^r + \gamma_{t_n} + \gamma^f_{c_{n-1},tp_{n-1}} + \gamma^b_{c_n,tp_n} + \gamma_{w_n} + \gamma_{sy_n} + \mu^d \quad \text{for } 1 \le n \le N_k \qquad (4\text{-}2)$$

其中詞首(n=1)的 $\gamma^f_{c_{n-1},tp_{n-1}}$ 為原式中的 $\gamma^f_{t_1}$；而詞尾(n=N)的 $\gamma^b_{c_n,tp_n}$ 為原式中的 $\gamma^b_{t_N}$。此外值得注意的是，在音節基頻軌跡模型中的參數為四維正交參數(向量)，而音節長度模型中的參數則為一維(純量)。我們可將基頻軌跡模型的幾個影響因素關係表示如圖 4-1。

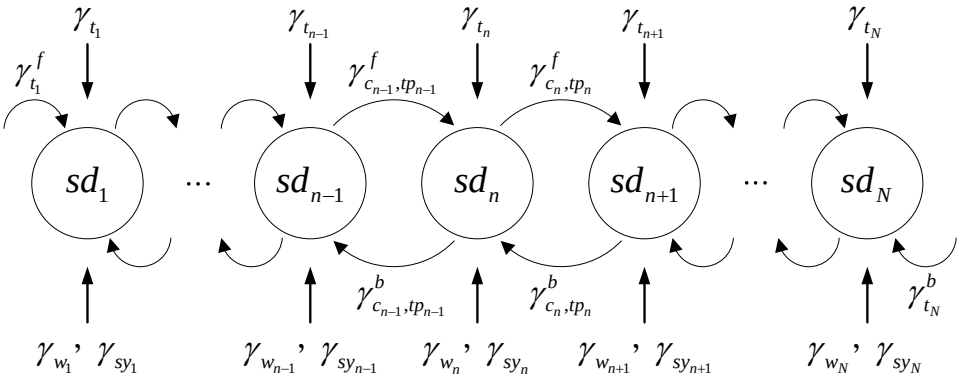


圖 4-1：音節長度模型之影響因素關係

4.2 模型訓練

如同第三章，在此論文中我們採用的模型為數據驅動，模型的參數需要由大量的語音資料訓練得到；在此節中我們討論模型的訓練方法。

4.2.1 影響因素推導

由(4-2)式中可知，音節的基頻軌跡 sd_n (Observed duration)是由 γ_{t_n}、$\gamma^f_{c_{n-1},tp_{n-1}}$、$\gamma^b_{c_n,tp_n}$、$\gamma_{w_n}$、$\gamma_{sy_n}$ 和 μ^d 等影響因素所相加組成，在給定音節的聲調、前後音節的連音狀態和聲調組合、音節在詞中的位置和基本音節類別等資訊的情況下，我們可預測出該音節的長度(Predicted duration)；而其誤差即為 sd^r_n。我們假設此誤差 sd^r_n 呈高斯分佈，可寫成下面數學式：

$$P(sd_n \mid t_n, c_{n-1}, tp_{n-1}, c_n, tp_n, w_n, sy_n) = N(sd_n; \gamma_{t_n} + \gamma^f_{c_{n-1},tp_{n-1}} + \gamma^b_{c_n,tp_n} + \gamma_{w_n} + \gamma_{sy_n} + \mu^d, \sigma^{d2}) \quad (4\text{-}3)$$

其中 σ^{d2} 為誤差 sd^r_n 的變異數。

同樣地，在音節長度模型中我們採用逐項最佳化程序和最大相似度法則來訓

練及更新模型參數。首先我們定義對數相似度函數如下：

$$L=\sum_{n=1}^{N_{all}}\log N(sd_n;sd_{t_n}+sd^{f}_{c_{n-1},tp_{n-1}}+sd^{b}_{c_n,tp_n}+sd_{w_n}+sd_{sy_n}+\mu^d,\sigma^{d2}) \tag{4-4}$$

其中 N_{all} 為語料庫中的音節總數；sd_n 為語料庫中第 n 個音節的長度；其餘參數和式(4-1)相似。接著，我們依照最大相似度法則可推導出模型參數的訓練與更新數學式：

$$\gamma_t=\frac{\sum_{n=1}^{N_{all}}\left(sd_n-\gamma^{f}_{c_{n-1},tp_{n-1}}-\gamma^{b}_{c_n,tp_n}-\gamma_{w_n}-\gamma_{sy_n}-\mu^d\right)\delta(t_n=t)}{\sum_{n=1}^{N_{all}}\delta(t_n=t)} \tag{4-5}$$

$$\gamma^{f}_{c,tp}=\frac{\sum_{n=1}^{N_{all}}\left(sd_n-\gamma_{t_n}-\gamma^{b}_{c_n,tp_n}-\gamma_{w_n}-\gamma_{sy_n}-\mu^d\right)\delta(c_{n-1}=c,tp_{n-1}=tp)}{\sum_{n=1}^{N_{all}}\delta(c_{n-1}=c,tp_{n-1}=tp)} \tag{4-6}$$

$$\gamma^{b}_{c,tp}=\frac{\sum_{n=1}^{N_{all}}\left(sd_n-\gamma_{t_n}-\gamma^{f}_{c_{n-1},tp_{n-1}}-\gamma_{w_n}-\gamma_{sy_n}-\mu^d\right)\delta(c_n=c,tp_n=tp)}{\sum_{n=1}^{N_{all}}\delta(c_n=c,tp_n=tp)} \tag{4-7}$$

$$\gamma_w=\frac{\sum_{n=1}^{N_{all}}\left(sd_n-\gamma_{t_n}-\gamma^{f}_{c_{n-1},tp_{n-1}}-\gamma^{b}_{c_n,tp_n}-\gamma_{sy_n}-\mu^d\right)\delta(w_n=w)}{\sum_{n=1}^{N_{all}}\delta(w_n=w)} \tag{4-8}$$

$$\gamma_{sy}=\frac{\sum_{n=1}^{N_{all}}\left(sd_n-\gamma_{t_n}-\gamma^{f}_{c_{n-1},tp_{n-1}}-\gamma^{b}_{c_n,tp_n}-\gamma_{w_n}-\mu^d\right)\delta(sy_n=sy)}{\sum_{n=1}^{N_{all}}\delta(sy_n=sy)} \tag{4-9}$$

$$\sigma^{d2}=\frac{1}{N_{all}}\sum_{n=1}^{N_{all}}\left(sd_n-\gamma_{t_n}-\gamma^{f}_{c_{n-1},tp_{n-1}}-\gamma^{b}_{c_n,tp_n}-\gamma_{w_n}-\gamma_{sy_n}-\mu^d\right)^2 \tag{4-10}$$

4.2.2 影響因素初始值

如前面所述，我們採用的逐項最佳化程序需要一個模型參數的初始值，而一個好的初始模型具有物理意義，它不但應該符合我們對語言學的認知，還要能幫

助模型快速收斂。因此我們仿照 3.2.2 小節產生各個影響因素的初始值：

$$\gamma_t = \frac{\sum_{n=1}^{N_{all}} \left(sd_n - \mu^d \right) \delta\left(t_n = t \right)}{\sum_{n=1}^{N_{all}} \delta\left(t_n = t \right)} \tag{4-11}$$

$$\gamma_{c,tp}^f = \frac{\sum_{n=1}^{N_{all}} sd_n \delta\left(c_{n-1} = c, tp_{n-1} = tp \right)}{\sum_{n=1}^{N_{all}} \delta\left(c_{n-1} = c, tp_{n-1} = tp \right)} - \frac{\sum_{n=1}^{N_{all}} sd_n \delta\left(c_{n-1} = c, t_n = t \right)}{\sum_{n=1}^{N_{all}} \delta\left(c_{n-1} = c, t_n = t \right)} \tag{4-12}$$

$$\gamma_{c,tp}^b = \frac{\sum_{n=1}^{N_{all}} sd_n \delta\left(c_n = c, tp_n = tp \right)}{\sum_{n=1}^{N_{all}} \delta\left(c_n = c, tp_n = tp \right)} - \frac{\sum_{n=1}^{N_{all}} sd_n \delta\left(c_n = c, t_n = t \right)}{\sum_{n=1}^{N_{all}} \delta\left(c_n = c, t_n = t \right)} \tag{4-13}$$

$$\gamma_w = \frac{\sum_{n=1}^{N_{all}} \left(sd_n - \gamma_{t_n} - \gamma_{c_{n-1},tp_{n-1}}^f - \gamma_{c_n,tp_n}^b - \mu^p \right) \delta\left(w_n = w \right)}{\sum_{n=1}^{N_{all}} \delta\left(w_n = w \right)} \tag{4-14}$$

$$\gamma_{sy} = \frac{\sum_{n=1}^{N_{all}} \left(sd_n - \gamma_{t_n} - \gamma_{c_{n-1},tp_{n-1}}^f - \gamma_{c_n,tp_n}^b - \gamma_{w_n} - \mu^d \right) \delta\left(sy_n = sy \right)}{\sum_{n=1}^{N_{all}} \delta\left(sy_n = sy \right)} \tag{4-15}$$

4.2.3 訓練流程

如前面所述，本研究採用逐項最佳化程序來訓練和更新各個影響因素的數值，順序分別為 γ_t、$\gamma_{c,tp}^f$、$\gamma_{c,tp}^b$、γ_{w_n}、γ_{sy_n} 和 σ^{d2}。值得注意的是，在音節長度模型中，我們暫時採用由第三章音節基頻軌跡模型訓練收斂的連音狀態，並且在疊代訓練的過程中固定住，暫不進行重新標記。我們重覆此疊代訓練的流程，直到模型參數收斂為止。在此收斂條件為：

$$\frac{L_m - L_{m-1}}{L_{m-1}} \le 10^{-7} \tag{4-16}$$

其中 L 為(4-4)式中定義的對數相似度函數，而 L_m 表示第 m 次疊代的函數數值。

音節長度的模型訓練流程如圖 4-2。

音節能量模型和音節長度模相似，惟能量模型模擬的對象為音節韻母部分的能量位準最大值，在此僅列出音節能量模型的數學式：

$$se_n = \begin{cases} se_1^r + \alpha_{t_1} + \alpha^f_{in_1} + \alpha^b_{c_1,tp_1} + \alpha_{w_1} + \alpha_{sy_1} + \mu^e & \text{for } n=1 \\ se_n^r + \alpha_{t_n} + \alpha^f_{c_{n-1},tp_{n-1}} + \alpha^b_{c_n,tp_n} + \alpha_{w_n} + \alpha_{sy_n} + \mu^e & \text{for } 2 \le n \le N_k - 1 \\ se_N^r + \alpha_{t_N} + \alpha^f_{c_{N-1},tp_{N-1}} + \alpha^b_{t_N} + \alpha_{w_N} + \alpha_{sy_N} + \mu^e & \text{for } n = N_k \end{cases} \quad (4\text{-}17)$$

其中 se_n 為第 n 個音節的能量(Observed energy)，se_n^r 為第 n 個音節的殘餘能量，其餘參數可參考(4-1)式。同樣地，我們可將(4-17)式簡化如下：

$$se_n = se_n^r + \alpha_{t_n} + \alpha^f_{c_{n-1},tp_{n-1}} + \alpha^b_{c_n,tp_n} + \alpha_{w_n} + \alpha_{sy_n} + \mu^d \quad \text{for } 1 \le n \le N_k \quad (4\text{-}18)$$

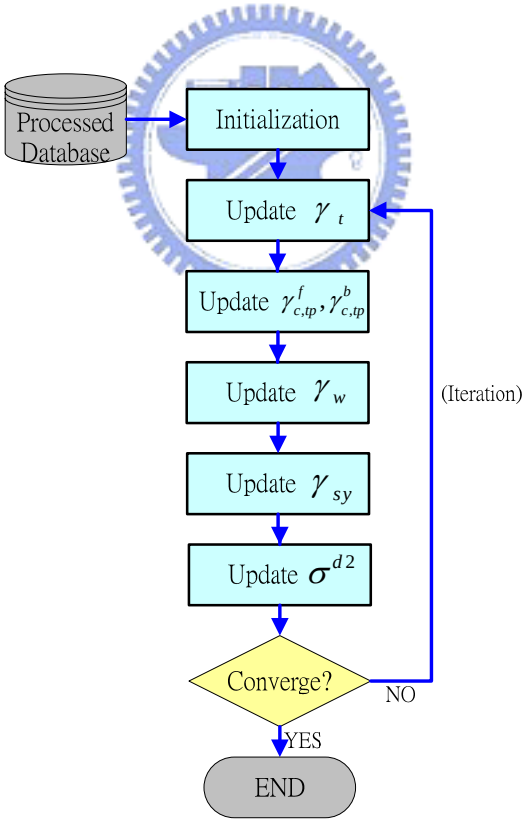


圖 4-2：音節長度模型中參數訓練和更新流程

4.3 音節長度模擬結果與分析

在此節中，我們將試圖分析訓練至收斂的模型參數，並且以訓練好的模型預測音節長度。由圖 4-3 可看出，疊代訓練一共重覆了 9 次，比音節基頻軌跡少許多，我們推測原因是音節長度模型的參數為一維(音節基頻軌跡模型為四維)，且省略了重新標記連音狀態的步驟。

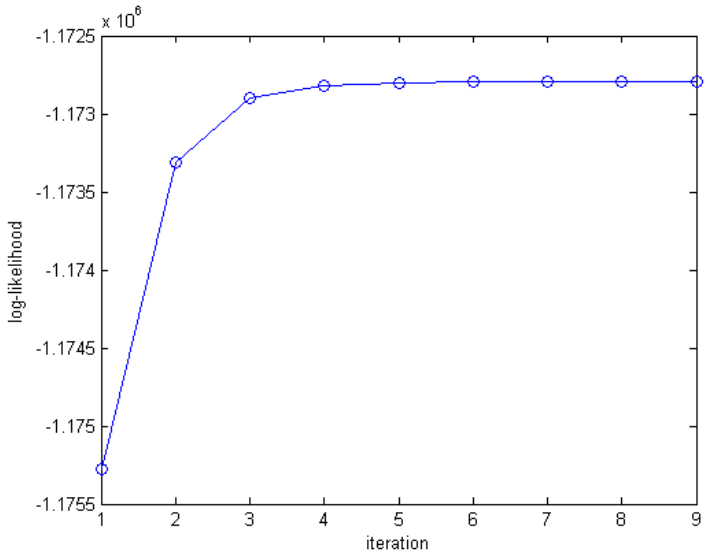


圖 4-3：音節長度模型訓練疊代次數及其目標函數值

4.3.1 聲調之影響因素

聲調對音節長度的影響非常大。圖 4-4 畫出聲調的影響因素，若其值大於 0 表示會拉長該音節的長度，反之則會縮短音節的長度。從圖中我們發現二聲的長度最長，而五聲(輕聲)最短，符合我們的認知。

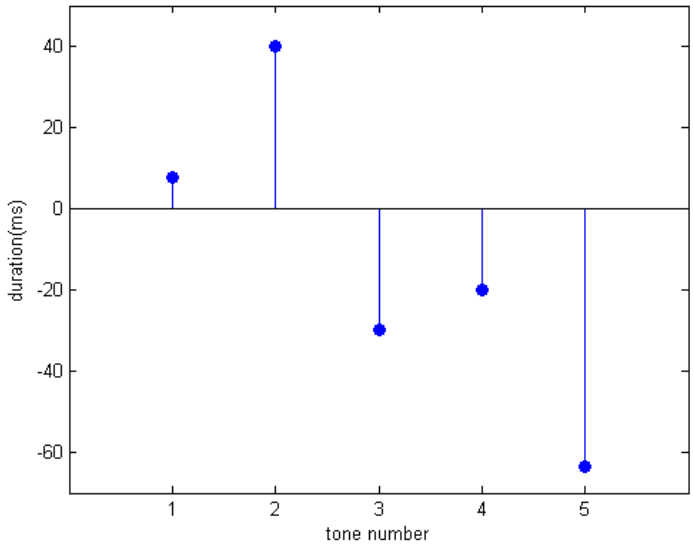


圖 4-4：音節長度模型中聲調之影響因素

4.3.2 受前一音節影響因素

音節長度亦會受前後音節所影響。圖 4-5 為受前一音節影響的結果，我們發現二個音節間的連音現象為「強」時影響較大，而連音現象為「中」和「弱」時影響較小，此結果和基頻軌跡的模擬結果相符。

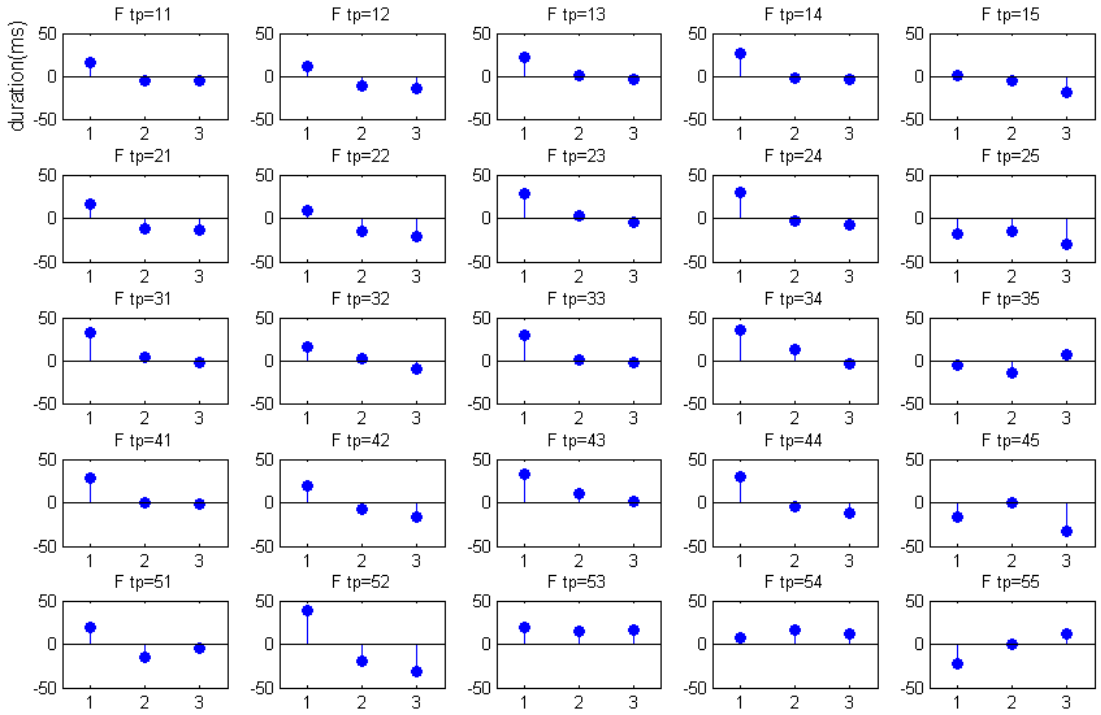


圖 4-5：音節長度模型中受前一音節影響因素。其中橫軸的“1,2,3”分別表示連音狀態「強」、「中」和「弱」。第(i,j)項表示前一個音節為tone i、此音節為tone j的聲調組合。

如(4-1)式所述，位於詞首的音節並無受前一音節之影響因素 $\gamma^{f}_{c_{n-1},tp_{n-1}}$，因此我們以詞首之影響因素 $\gamma^{f}_{t_1}$ 與其對應，結果繪於圖 4-6 中。

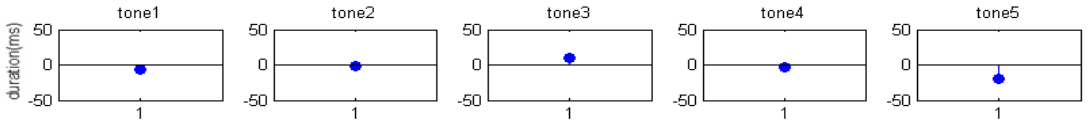


圖 4-6：音節長度模型中詞首影響因素。詞首影響因素沒有連音狀態之分別。

4.3.3 受後一音節影響因素

圖 4-7 為音節長度受後一音節影響的結果。值得注意的是(3,3)的聲調組合，由於中文的(3,3)聲調組合會唸成(2,3)聲，並且由 4.3.1 小節中我們知道二聲的長度比三聲長；而此圖 4-7 中第(3,3)項皆為正值的結果相當合理。

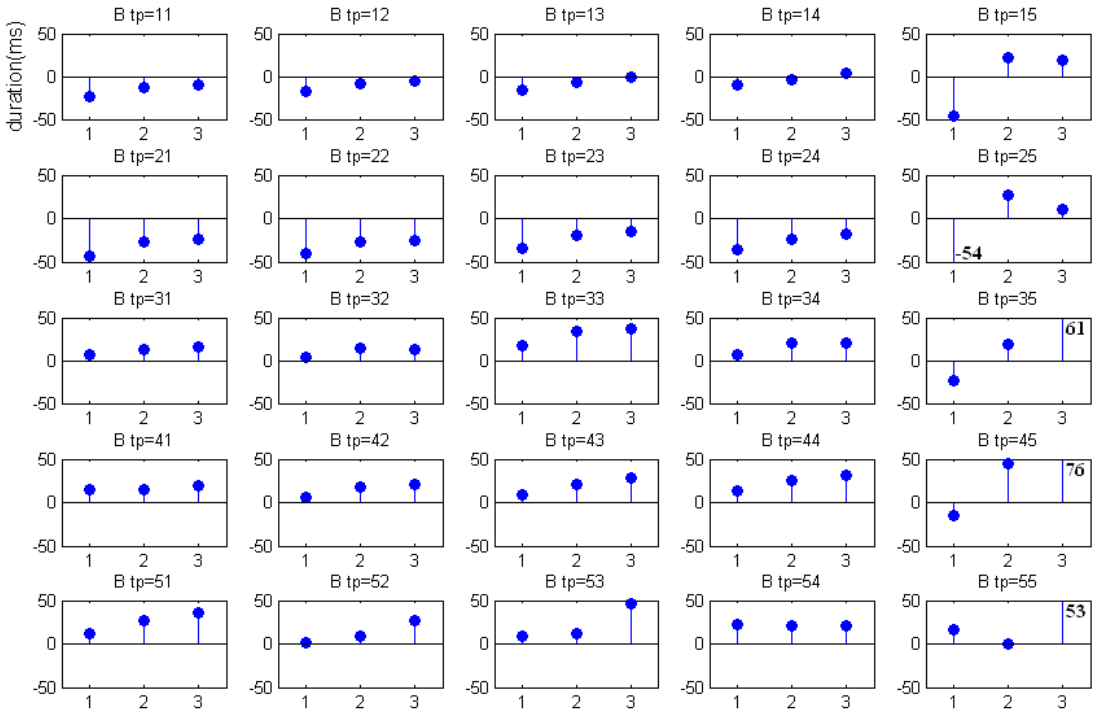


圖 4-7：音節長度模型中受後一音節影響因素。其中橫軸的“1,2,3”分別表示連音狀態「強」、「中」和「弱」。第(i,j)項表示此音節為tone i、後一個音節為tone j的聲調組合。

和前一小節相似，位於詞尾的音節並無受後一音節之影響因素 $\gamma^{b}_{c_n,tp_n}$，因此我們以詞尾之影響因素 $\gamma^{b}_{t_N}$ 與其對應，結果繪於圖 4-8 中。

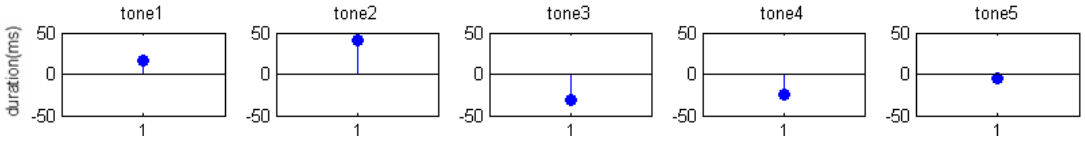


圖 4-8：音節長度模型中詞尾影響因素。詞尾影響因素沒有連音狀態之分別。

4.3.4 音節在詞的位置之影響因素

圖 4-9 為音節在詞的位置影響因素。我們發現從二字詞到六字詞的詞尾音節都有被拉長的現象，詞尾音節的平均長度比其它音節的平均長度約多了 100ms。此現象在語言學中已被廣泛討論(Final lengthening)，通常在一句話的尾端會特別明顯；而此論文所模擬的對象為單詞，在語料庫中單詞和單詞間有充分的停頓，因此詞尾可等同於句尾。

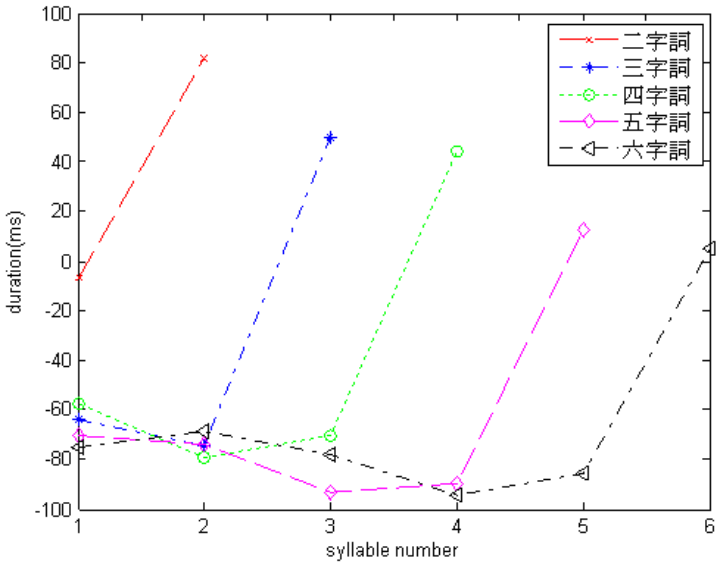


圖 4-9：音節長度模型中音節在詞的位置影響因素

4.3.5 音節類別之影響因素

中文有 411 種基本音節類別，不同的音節由不同的聲母和韻母組合而成，也對音節的長度產生直接的影響。為了進一步瞭解音節類別所造成的影響，我們使用決策樹(Decision tree)來分析。

決策樹可分為二大類，分別為分類樹(Classification tree)和回歸樹(Regression tree)；其中分類樹適用於類別式(Categorical)的分類結果，而回歸樹則適用於數值式(Numeric)的分類結果，決策樹相關的介紹可參考[10]。在此小節中我們欲分析的音節類別影響因素是連續的實數，因此採用回歸樹。

在進行決策樹的分裂前，我們先準備一套問題集，內容和各種基本音節在語言學的特性有直接的關係，如表 4-1。

表 4-1：決策樹問題集

編號	問題內容
Q1	聲母是否為空聲母

Q2	聲母是否為爆破音、不送氣(ㄅ,ㄉ,ㄍ)
Q3	聲母是否為爆破音、送氣(ㄆ,ㄊ,ㄎ)
Q4	聲母是否為鼻音、濁音(ㄇ,ㄋ,ㄌ,ㄖ)
Q5	聲母是否為摩擦音、清音(ㄈ,ㄏ,ㄒ,ㄕ,ㄙ)
Q6	聲母是否為塞擦音、送氣(ㄑ,ㄔ,ㄘ)
Q7	聲母是否為塞擦音、不送氣(ㄐ,ㄓ,ㄗ)
Q8	韻母是否為單韻母(single vowel)
Q9	韻母是否為複韻母(compound vowel)
Q10	韻母是否為鼻音結尾(nasal ending vowel)
Q11	韻母是否有介音(medial)
Q12	韻母是否有開口
Q13	韻母是否為 ㄧ 開頭
Q14	韻母是否為 ㄨ 開頭
Q15	韻母是否為 ㄩ 開頭

接著我們將所有 411 種基本音節的影響因素放在根節點(Root node)，再一一詢問表 4-1 中的問題，可依「是」或「否」的回答將基本音節分成二類，稱之為左子節點(Left son node)和右子節點(Right son node)。然後我們以高斯分佈分別去模擬根節點、左子節點和右子節點的資料，得到相似度 L_{root} 、 L_{left_son} 和 L_{right_son} 。在 15 個問題中，我們分別計算其相似度增益(Likelihood gain)：

$$L_{gain} = L_{left\_son} + L_{right\_son} - L_{root} \tag{4-20}$$

其中 L_{gain} 最大的問題則為此節點決定要問的問題。接著我們重覆此步驟，讓節點不斷分裂，直到收斂為止。一般收斂的條件有二，分別為相似度增益太小或左右子節點所擁有的資料量太少。

我們將音節類別之影響因素的決策樹分裂結果整理於圖 4-10，其中每個節點中的資訊分別為：(1)此節點被問的問題；(2)此節點中所有資料的平均值(Mean)；(3)此節點中所有資料的標準差(Standard deviation)；和(4)此節點的資料量(Data number)。值得注意的是，中文的基本音節共有 411 種，但有 9 種在我們的語料庫中未出現，因此在根節點中只有 402 種基本音節。

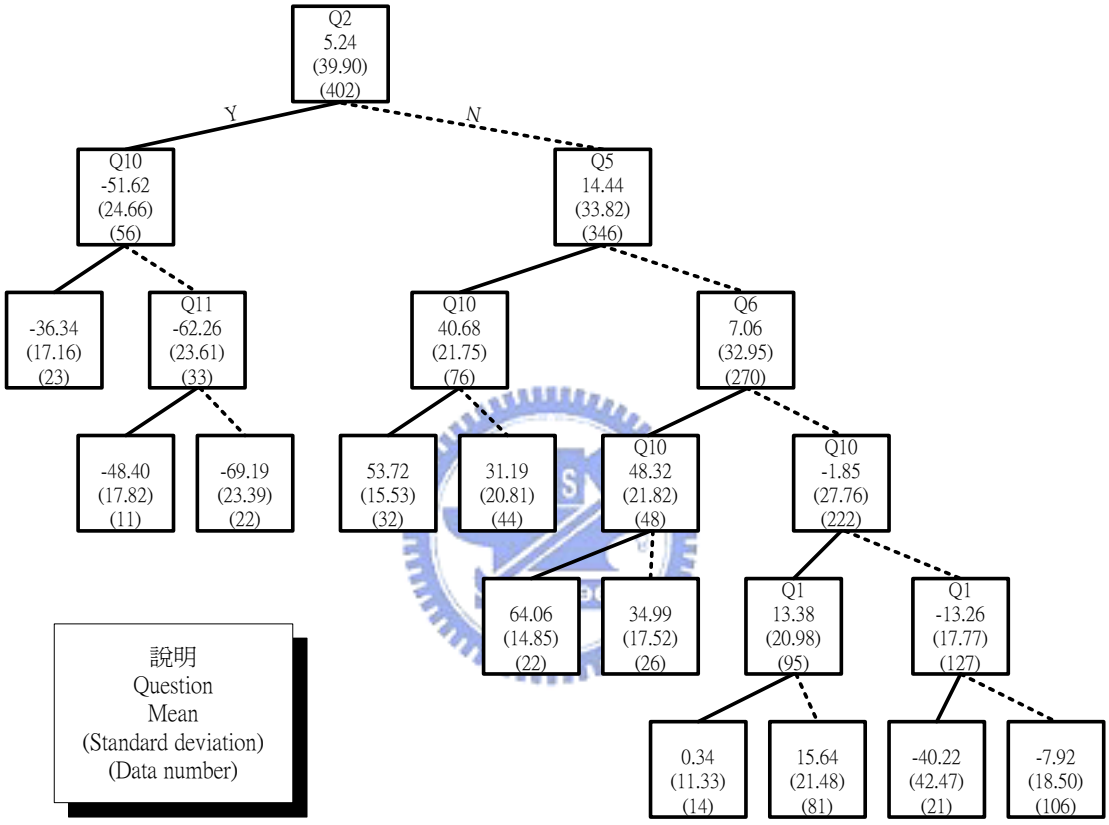


圖 4-10：音節長度模型中音節類別影響因素之決策樹分析結果

由圖 4-10 我們可得到以下結論：

一、聲母為【爆破音、不送氣】的音節長度較短(Q2)

二、韻母為【鼻音結尾】的音節長度較長(Q10)

三、聲母為【摩擦音、清音】的音節長度較長(Q5)

四、韻母為【有介音】的音節長度較長(Q11)

五、聲母為【擦塞音、送氣】的音節長度較長(Q6)

六、聲母為【空聲母】的音節長度較短(Q1)

以上結果符合我們對語言學的認識，亦證明此模型不但能有效模擬音節長度，也有助於分析。

4.3.6 音節長度預測與模擬誤差分析

如前面所述，在給定音節的聲調、前後連音資訊、音節在詞中位置和基本音節類別等資訊時，我們可由此模型預測音節的長度。圖 4-11 為一些實際和預測音節長度的例子。

為了方便觀察模擬結果，我們畫出測試語料實際和正規化(Normalized，即原音節長度減去四種音節長度模型影響因素後的音節長度)後的直方圖(Histogram)於圖 4-12。我們發現經正規化後的音節長度分佈有效向內集中，變異數較原本小許多。

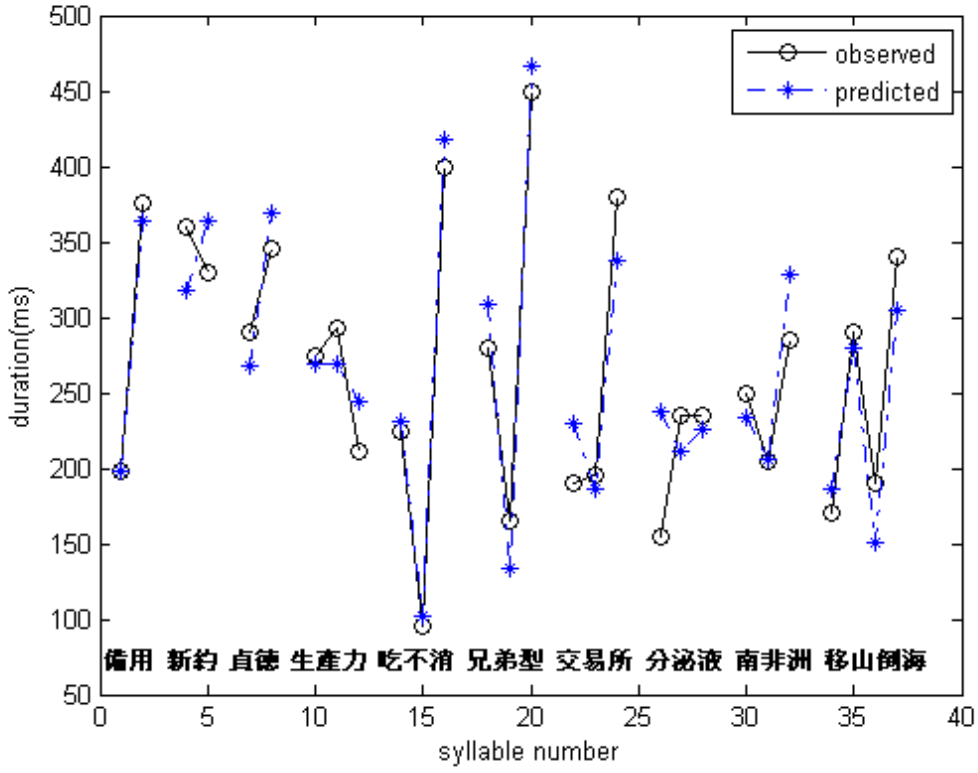


圖 4-11：以模型預測音節長度。其中實線為實際長度，而虛線為預測長度。

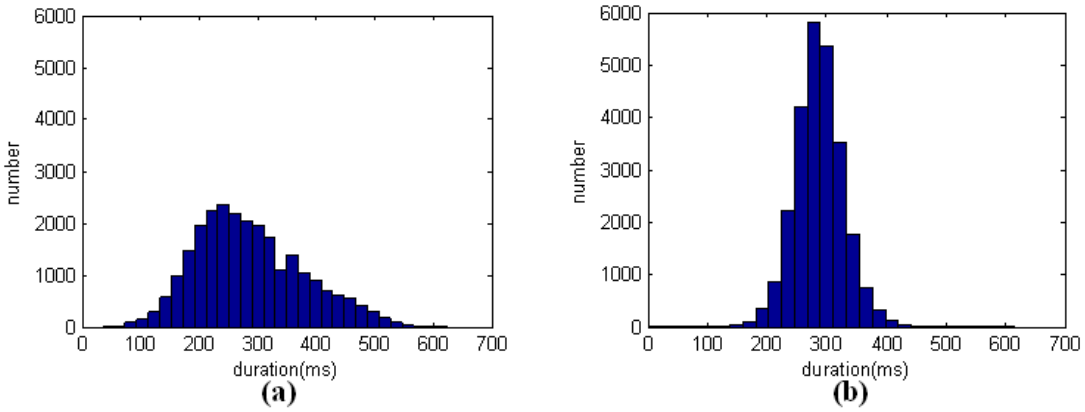


圖 4-12：(a)實際與(b)正規化音節長度之分佈

我們將內部測試與外部測試的變異數、預測均方誤差(Mean Square Error, MSE)和相關係數整理如表 4-2。

表 4-2：音節長度模型內部測試與外部測試結果(聲調組合)

	Observed Data Variance	**Predicted Data MSE**	**Correlation Coefficient**
Inside	8372 (ms^2)	1637 (ms^2)	0.897
Outside	8389 (ms^2)	1729 (ms^2)	0.891

如同 3.3.6 小節，我們亦計算音節長度模型各種影響因素的貢獻度；由表 4-3 可發現音節在詞中位置之影響因素的貢獻度最大，單是此項可將模擬誤差降至 55.9%；其次分別為基本音節類別、聲調和連音影響等影響因素。

表 4-3：影響因素與殘餘誤差分析表

影響因素	**TRE**

+ 音節在詞中位置	55.9%
+ 基本音節類別	35.3%
+ 聲調	28.8%
+ 連音影響	20.6%

值得補充的是，在本論文中的「連音影響」是考慮二個音節的「聲調組合」，我們亦嚐試考慮「韻母-聲母類別組合」(Final-initial pair)，即二個音節間第一個音節的韻母類別和第二個音節的聲母類別之組合，但模擬效果並不如預期；從表 4-4 中我們發現「韻母-聲母類別組合」的模型在預測誤差的大小和相關係數上的表現都不如「聲調組合」。

表 4-4：音節長度模型內部測試與外部測試結果(韻母-聲母組合)

	Observed Data Variance	**Predicted Data MSE**	**Correlation Coefficient**
Inside	8372 (ms^2)	2230 (ms^2)	0.857
Outside	8389 (ms^2)	2279 (ms^2)	0.854

4.4 音節能量模擬結果與分析

在本論文中，音節能量模型和音節長度模型相似，因此我們參考 4-2 節的模型，模擬音節的能量(韻母部分的能量位準最大值)，並且試圖分析訓練至收斂的模型參數，再以訓練好的模型預測音節能量。由圖 4-13 可看出，疊代訓練一共重覆了 10 次，和音節長度模型相近。

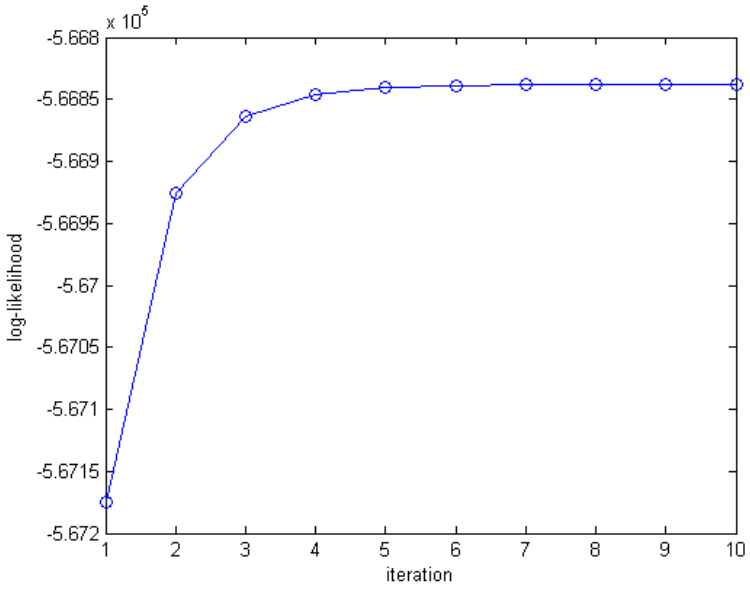


圖 4-13：音節能量模型訓練疊代次數及其目標函數值

4.4.1 聲調之影響因素

聲調對音節長度的影響非常大。圖 4-14 畫出聲調的影響因素，若其值大於 0 表示會提高該音節的能量，反之則會降低音節的能量。從圖中我們發現一聲和四聲的能量較大，二聲和三聲較小，而五聲(輕聲)最小，符合我們的認知。

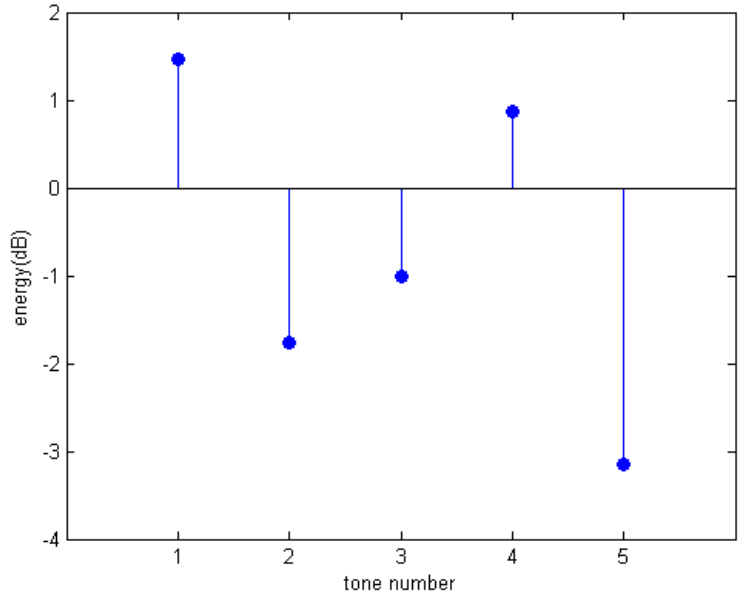


圖 4-14：音節能量模型中聲調之影響因素

4.4.2 受前一音節影響因素

音節長度亦會受前後音節所影響。圖 4-15 為受前一音節影響的結果，我們發現整體而言二個音節間的連音現象為「強」時影響較大，而連音現象為「中」)和「弱」時影響較小，此結果和音節長度的模擬結果相符。我們發現此現象在音節為輕聲(第五行)時特別明顯，推測是因為輕聲本身能量較小，因此在連音現象「強」時，受到前一音節影響而使能量增大的幅度較大之故。

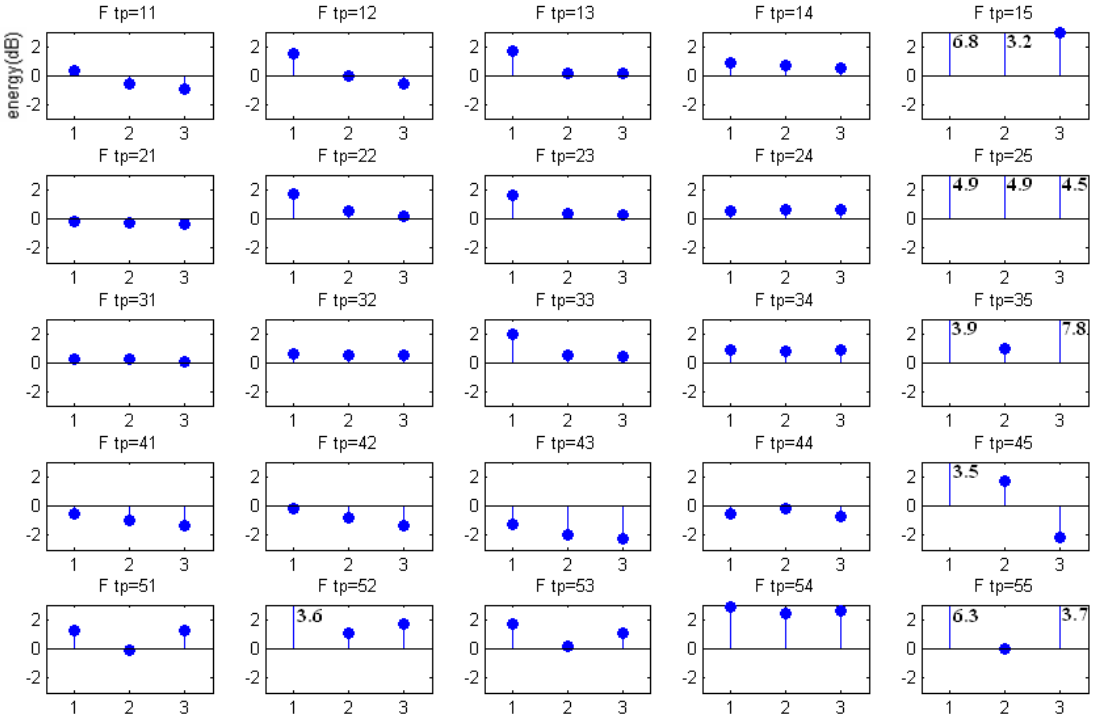


圖 4-15：音節能量模型中受前一音節影響因素。其中橫軸的“1,2,3”分別表示連音狀態「強」、「中」和「弱」。第(i,j)項表示前一個音節為tone i、此音節為tone j的聲調組合。

如(4-17)式所述，位於詞首的音節並無受前一音節之影響因素 $\alpha^{f}_{c_{n-1},tp_{n-1}}$ ，因此我們以詞首之影響因素 $\alpha^{f}_{t_1}$ 與其對應，結果繪於圖 4-16 中。

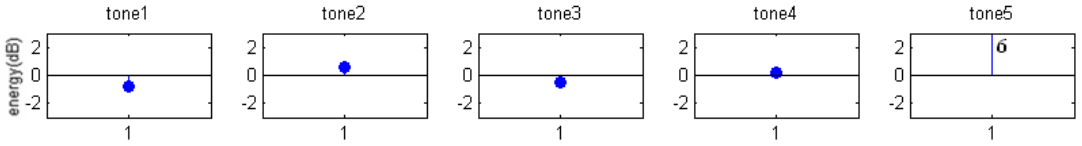


圖 4-16：音節能量模型中詞首影響因素。詞首影響因素沒有連音狀態之分別。

4.4.3 受後一音節影響因素

同樣地，我們將音節長度受後一音節影響的結果繪於圖 4-17，發現此影響因素普遍較受前一音節影響因素來得小。

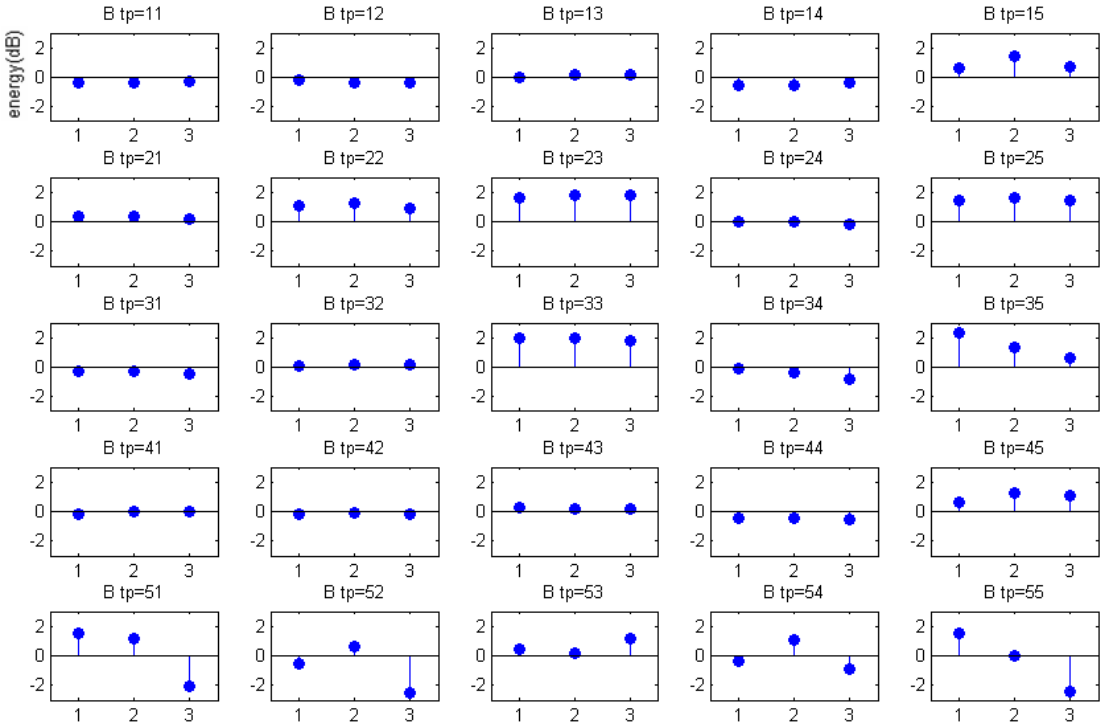


圖 4-17：音節能量模型中受後一音節影響因素。其中橫軸的“1,2,3”分別表示連音狀態「強」、「中」和「弱」。第(i,j)項表示此音節為tone i、後一個音節為tone j的聲調組合。

和前一小節相似，位於詞尾的音節並無受後一音節之影響因素 $\alpha^{b}_{c_n,tp_n}$ ，因此

我們以詞首之影響因素 $\alpha_{t_N}^{b}$ 與其對應，結果繪於圖 4-8 中。同樣地，我們發現無論是詞首或詞尾之影響因素，在動態範圍上均比聲調之影響因素小許多。

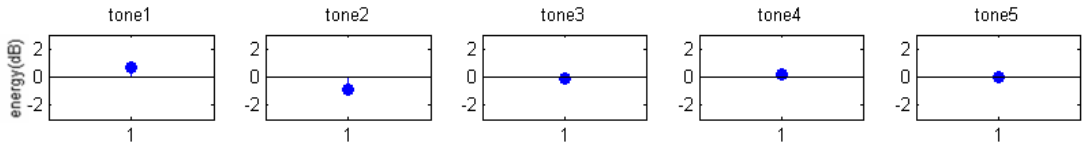


圖 4-18：音節能量模型中詞尾影響因素。詞尾影響因素沒有連音狀態之分別。

4.4.4 音節在詞的位置之影響因素

圖 4-19 為音節在詞的位置影響因素。我們發現從二字詞到六字詞音節的能量都由前向後逐漸遞減，此結果約略和音節基頻軌基模型的模擬結果相似；此現象表示人在說話時不僅音調會慢慢下降，能量也會漸漸變小，直到下一句話或下一個韻律詞出現[11]。此外，我們發現若詞中的字數愈多，詞首的能量會愈高，此結果亦和音節基頻軌基模型相似。

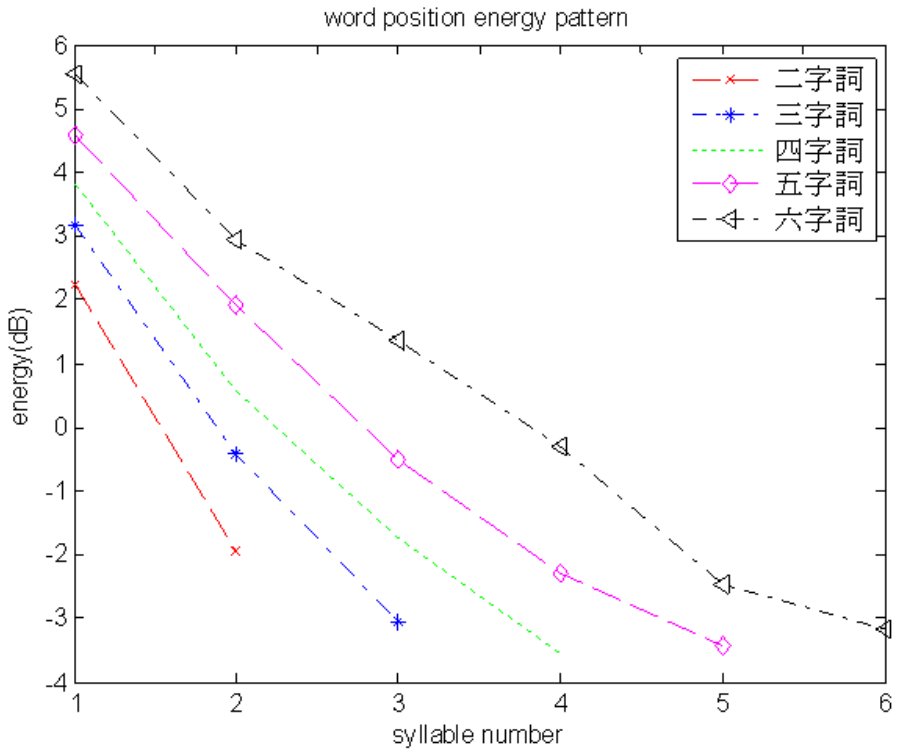


圖 4-19：音節能量模型中音節在詞的位置影響因素

4.4.5 音節類別之影響因素

和 4.3.5 小節相同，我們以決策樹(回歸樹)分析音節能量模型中的基本音節類別影響因素。在此的問題集同表 4-1，而決策樹分裂的結果如圖 4-20。

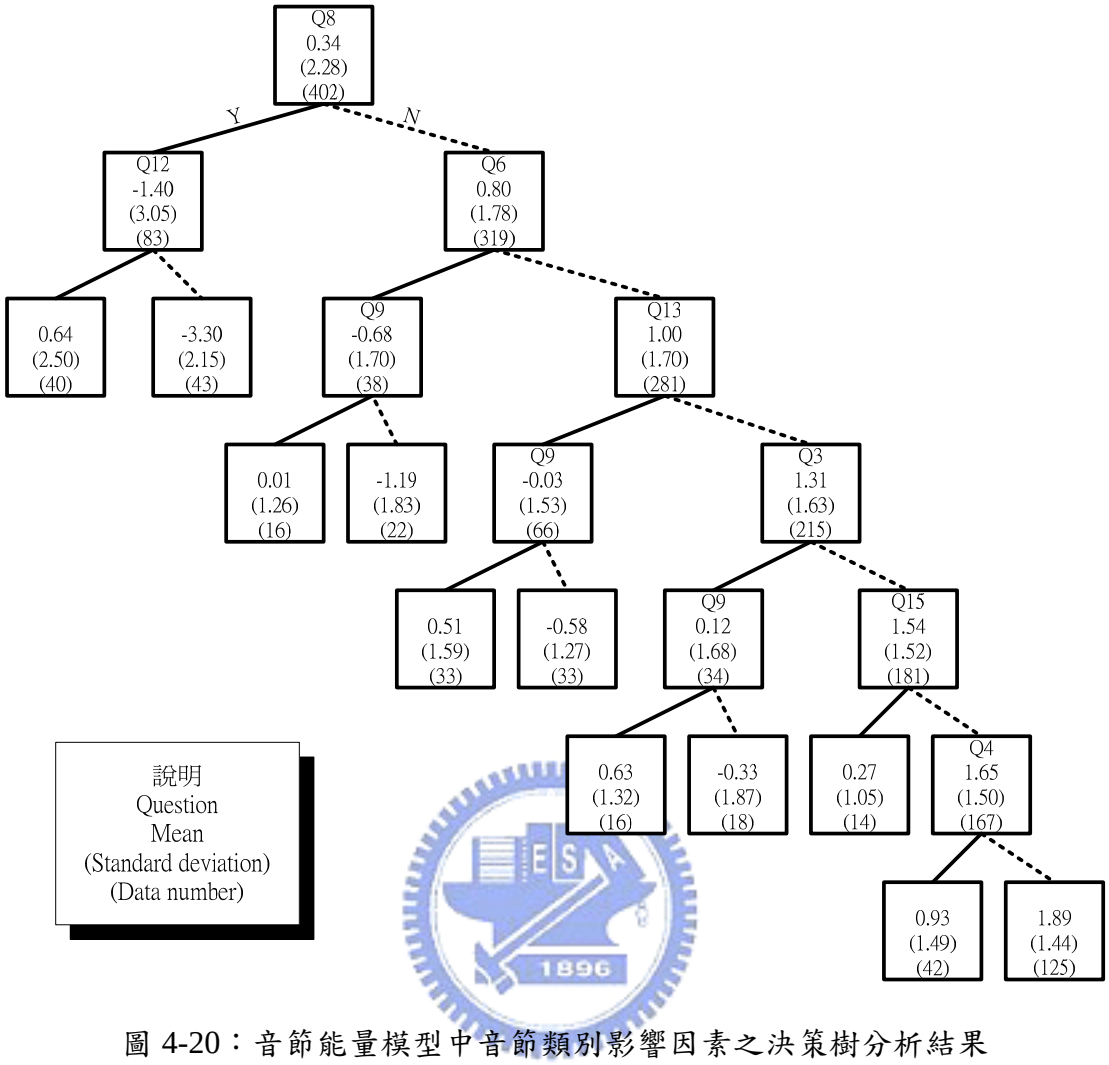


圖 4-20：音節能量模型中音節類別影響因素之決策樹分析結果

我們進一步分析決策樹，可得到以下結論：

一、韻母為【單韻母】的音節能量較小(Q8)

二、韻母為【開口】的音節能量較大(Q12)

三、聲母為【塞擦音、送氣】的音節能量較小(Q6)

四、韻母為【複韻母】的音節能量較大(Q9)

五、韻母為【以 一 開頭】的音節能量較小(Q13)

六、韻母為【以 ㄩ 開頭】的音節能量較小(Q15)

七、聲母為【鼻音、濁音】的音節能量較小(Q4)

4.4.6 音節能量預測

如前面所述，在給定音節的聲調、前後連音資訊、音節在詞中位置和基本音節類別等資訊時，我們可由此模型預測音節的能量。圖 4-21 為一些實際和預測音節能量的例子。

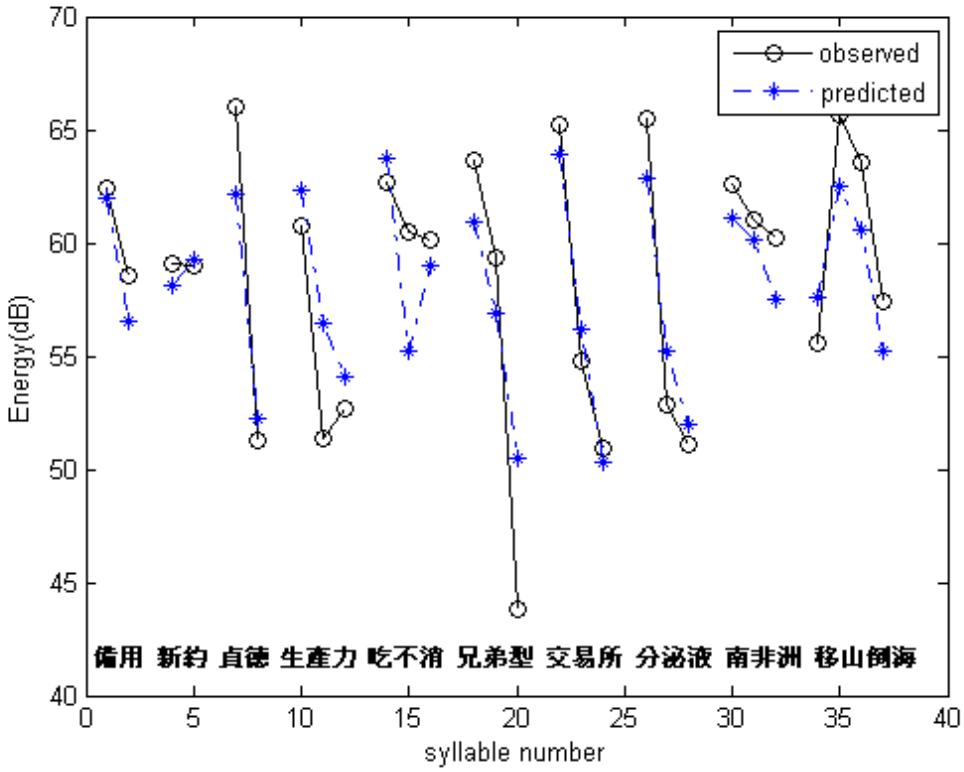


圖 4-21：以模型預測音節能量。其中實線為實際能量，而虛線為預測能量。

和 4.3.6 小節相似，為了方便觀察模擬結果，我們畫出測試語料實際和正規化後的長條圖於圖 4-22。我們發現經正規化(原音節能量減去四種音節能量模型影響因素後的音節能量)後的音節能量分佈明顯向內集中，變異數較原本小許多。

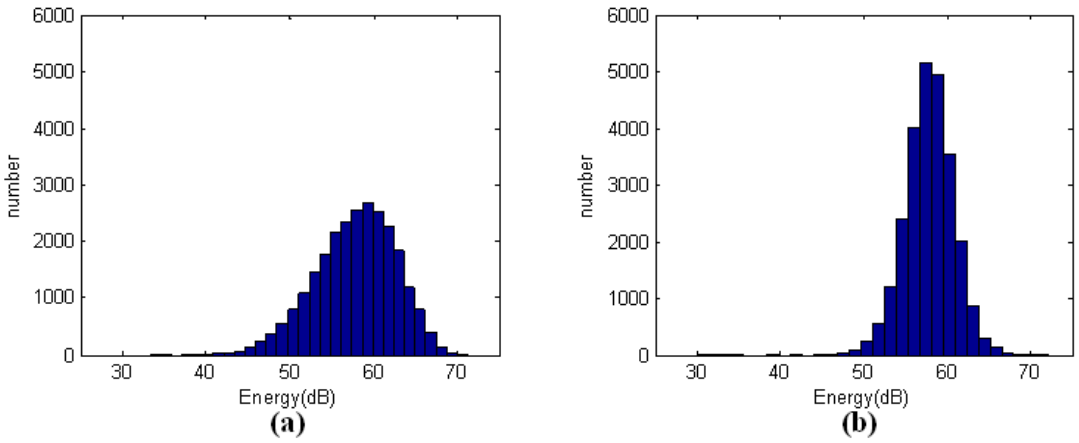


圖 4-22：(a)實際與(b)正規化音節能量之分佈

同樣地，我們將內部測試與外部測試的變異數、預測均方誤差和相關係數整理如表 4-5。

表 4-5：音節能量模型內部測試與外部測試結果

	Observed Data Variance	**Predicted Data MSE**	**Correlation Coefficient**
Inside	22.77 (dB^2)	8.25 (dB^2)	0.799
Outside	22.97 (dB^2)	8.56 (dB^2)	0.792

最後，我們亦計算音節能量模型各種影響因素的貢獻度；由表 4-6 可發現基本音節類別之影響因素的貢獻度最大，可將模擬誤差降至 71.2%；其次分別為音節在詞中位置、聲調和連音影響等影響因素。

表 4-6：影響因素與殘餘誤差分析表

影響因素	**TRE**
＋ 基本音節類別	71.2%
＋ 音節在詞中位置	46.9%
＋ 聲調	42.5%
＋ 連音影響	37.3%

4.5 音節間停頓與連音狀態分析

在前面幾個章節我們分別介紹了音節基頻軌跡、長度和能量的模型，但在把模型實際運用在 TTS 系統時，我們仍需要音節間停頓長度和連音狀態的資訊，因此在此節中我們延續前面的韻律模型，以決策樹來訓練此二種模型。

4.5.1 音節間停頓之分析

我們以決策樹分析語料庫中所有訓練語料的音節連接處停頓的長度(單位為 ms)，並以表 4-1 的問題集做為此決策樹的分裂依據，並將結果繪於圖 4-23。值得注意的是，在前面音節長度與能量的模型中，問題集中的聲母和韻母是指該音節的聲母和韻母；而在此小節和下一小節中，問題集中的聲母和韻母分別是指音節連接處右邊音節的聲母和左邊音節的韻母。

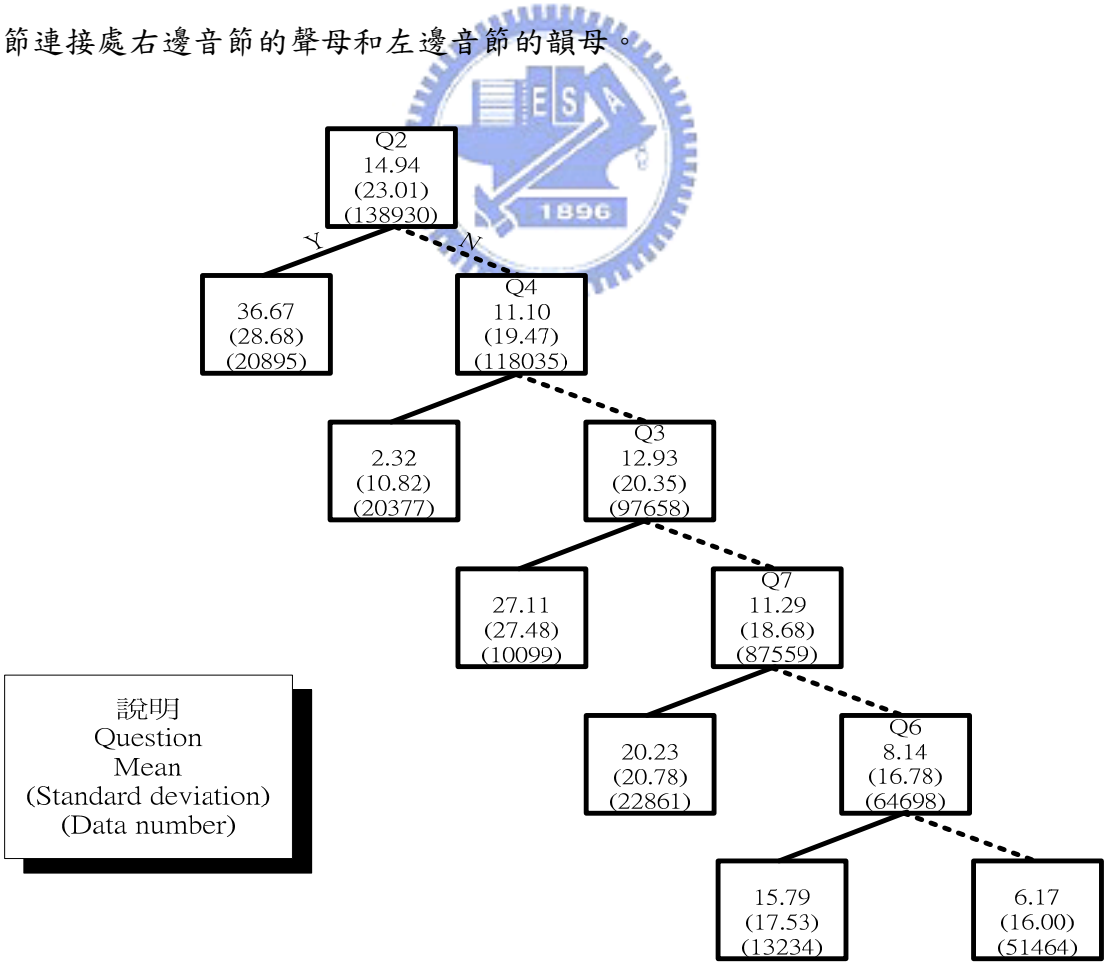


圖 4-23：音節間停頓長度之決策樹分析結果

由圖 4-23 我們可知音節聲母為【爆破音、不送氣】時其前面停頓最長，其次是【爆破音、送氣】，接著依序為【塞擦音、不送氣】和【塞擦音、送氣】；而音節聲母為【鼻音、濁音】時其前面停頓最短。

4.5.2 音節間連音狀態之分析

音節間的連音狀態分析和前面有些不同，因為在本論文中連音狀態為離散的，分為「強」、「中」和「弱」三種，因此適用於類別式的決策樹，即分類樹。在前面的回歸樹中，我們以高斯模擬資料的分佈，並求取其相似度和相似度增益做為節點分裂的標準；而在分類樹中，以高斯模擬有限類別的分佈顯然不適合，因此我們改取資料的熵(Entropy)和資訊增益(Information gain)來表示某一問題是否能將節點中的資料有效分成二類。

熵能描述一群資料分佈的亂度和資訊量，若資料分佈愈平均則熵愈大，反之若資料愈集中於少數幾種類別則熵愈小。熵的數學式如下：

$$\text{Entropy}(S)=\sum_{i=1}^{N} -p_i \log_2(p_i) \tag{4-21}$$

其中 S 表示某一群資料的集合，在決策樹中可視為某一個節點中的所有資料；N 表示資料的類別，在此論文中為三種連音狀態；p_i 則表示該集合中資料類別為 i 的機率。

資訊增益能描述某問題對一個節點分裂前後熵的減少程度，資訊增益愈大表示此問題在這個節點的分裂愈有效率。資訊增益的數學式如下：

$$\text{Gain}(S,A)=\text{Entropy}(S)-\sum_{v\in \text{Values}(A)} \frac{|S_v|}{|S|}\text{Entropy}(S_v) \tag{4-22}$$

其中 A 表示某一群資料的屬性，在決策樹中即為問題集中的問題；Values(A) 表示屬性的內容，在決策樹中即為問題的答案(是或否)，S_v 為 S 的一個子集合，$|S_v|$

和$|S|$分別為二個集合的資料數量。

重覆(4-22)式我們可讓決策樹不斷分裂，直到收斂為止。和前面的回歸樹相同，收斂的條件為資訊增益太小或左右子節點所擁有的資料量太少。我們將連音狀態的決策樹分裂結果整理於圖 4-24，其中每個節點中的資訊分別為：(1)此節點被問的問題；(2)此節點的資料量；和(3)此節點中三種連音狀態占有比例圖。

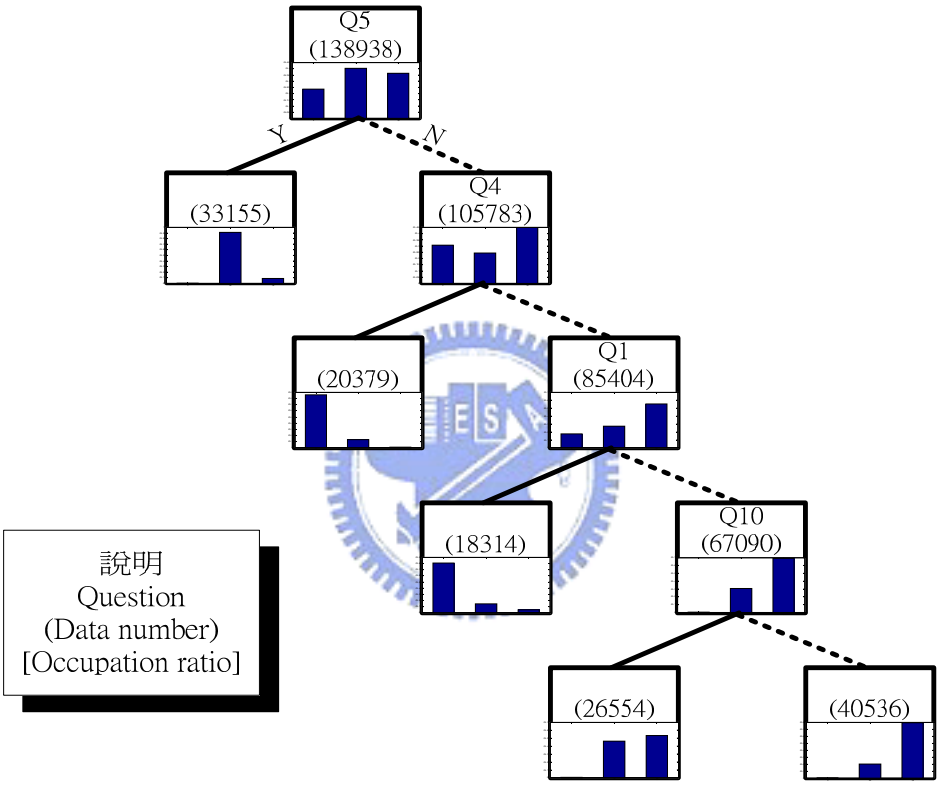


圖 4-24：連音狀態之決策樹分析結果

我們將圖 4-24 之結果整理如下：

一、音節聲母為【摩擦音、清音】時其和前一音節之連音狀態多為「中」(Q5)

二、音節聲母為【鼻音、濁音】時其和前一音節之連音狀態多為「強」(Q4)

三、音節聲母為【空聲母】時其和前一音節之連音狀態多為「強」(Q1)

四、音節韻母不為【鼻音結尾】時其和後一音節之連音狀態多為「弱」(Q10)

為了將此韻律模型提供 TTS 系統使用，我們將 4.5.1 和 4.5.2 小節的結果整合成表 4-7，待 TTS 欲合成中文單詞語音時，可直接查表取得相關資訊。

表 4-7：整合音節間停頓長度和連音狀態資訊

聲母類別	內容	停頓長度	連音狀態
1	聲母為空聲母	0 ms	強
2	聲母為爆破音、不送氣 (ㄅ,ㄉ,ㄍ)	37 ms	弱
3	聲母為爆破音、送氣 (ㄆ,ㄊ,ㄎ)	27 ms	弱
4	聲母為鼻音、濁音 (ㄇ,ㄋ,ㄌ,ㄖ)	0 ms	強
5	聲母為摩擦音、清音 (ㄈ,ㄏ,ㄒ,ㄕ,ㄙ)	6 ms	中
6	聲母為塞擦音、送氣 (ㄑ,ㄔ,ㄘ)	16 ms	弱
7	聲母為塞擦音、不送氣 (ㄐ,ㄓ,ㄗ)	20 ms	弱

第五章 中文韻律學習系統

在第三章和第四章中我們分別建立了音節基頻軌跡、長度和能量模型，在第五章我們將運用此韻律模型建立一個基本的韻律學習系統。對許多外國人而言，中文學習最大的難處在於韻律，如聲調不分等，常常會造成聽者對語意的誤解。對於初學者而言，單詞比整句話容易學習，因此我們建立一套中文單詞韻律學習系統，期望能在語言學習上有所幫助。

在 5.1 節中我們將介紹系統架構與細節，而 5.2 節則會展示此系統。

5.1 系統架構

本系統可分為以下四部分：

一、**輸入單詞**：由使用者輸入中文單詞，而系統會由我們訓練好的韻律模型產生適當的韻律參數，包括音節基頻軌跡、長度、能量和連音狀態，並顯示在螢幕上供使用者參考。

二、**播放範例**：系統內建一個簡易版的 TTS 系統，可以依使用者輸入的單詞進行單元選取(Unit selection)並合成出範例單詞語音，讓使用者模仿學習。

三、**錄音與放音**：使用者可以將自己的聲音錄下來，除了可播放出來外，系統亦會求取此段錄音的韻律參數並顯示在螢幕上。

四、**比較與學習**：在完成以上流程後，使用者可同時看到系統合成的目標(Target)韻律參數和使用者錄音的源(Source)韻律參數，經比較後反覆修正和練習，達成韻律學習的目的。

圖 5-1 為整個系統的流程與架構圖；而在接下來幾個小節中，我們將介紹系統實作的事前準備工作和細節。

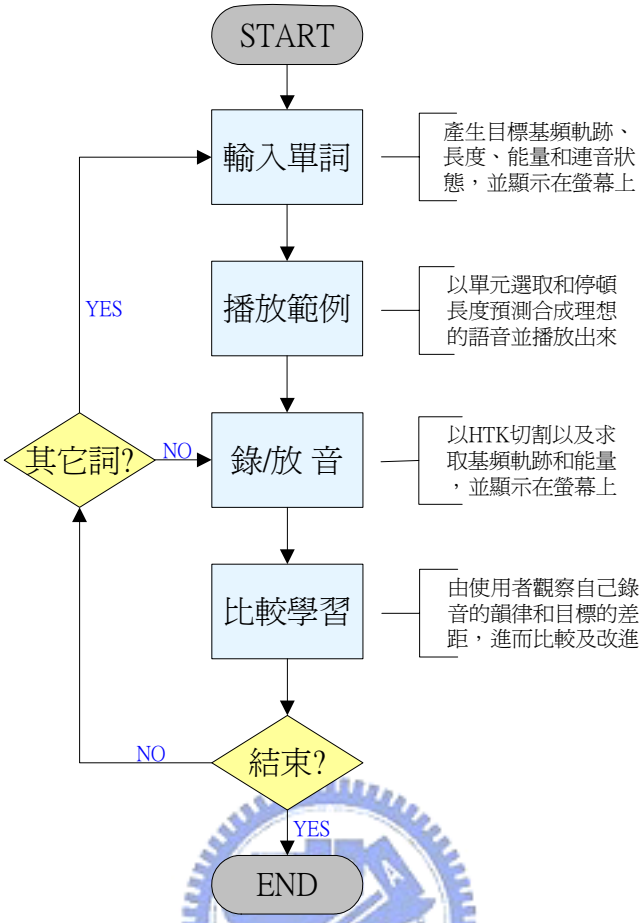


圖 5-1：韻律學習系統流程與架構圖

5.1.1 樣本選取(事前準備工作)

在此韻律學習系統中，我們希望合成出範例語音供系統使用者參考，因此需內建簡易的 TTS 系統。建立 TTS 系統的首要工作為建立一套好的樣本語料庫(Speech corpus)，所以在此節中我們討論如何由整個中文單詞語料庫進行樣本選取，以得到一套好的樣本語料庫。

建立 TTS 系統時，合成語音品質好壞的關鍵之一在於樣本語料庫的大小，樣本語料庫愈豐富則合成的音質愈好，但也相對會增加系統的大小和運算複雜度。在我們的中文單詞語料庫中帶有聲調的音節有 1,255 種，而我們為了兼顧系統的音質和運算複雜度，從整個中文單詞語料庫中在本系統中選取 20,000 個左右的帶聲調音節樣本，以供 TTS 系統使用。由於此樣本需具有代表性，我們採

用分群(Clustering)方法和貪婪演算法(Greedy algorithm)使得樣本和整體語料庫的均方誤差(Mean Square Error, MSE)最小。

首先我們統計每個帶聲調音節在整體語料庫中的個數，然後將此統計值依照區間對應到一個「基數」，如表 5-1。接著我們分配給每種帶聲調音節一個基數的樣本數 n_k，$k \in (1,2,...,1255)$，作為樣本數的初始值。有了此樣本數後，我們使用 LBG 演算法(LBG Algorithm) [12]依照音節基頻軌跡和長度將帶聲調音節分成 n_k 群，再取各群最接近群中心(Clustering center)的音節做為代表此群集的樣本，由表 5-1 計算後可發現此時的總樣本數為 9,275 個。最後我們計算每一種帶聲調音節樣本和該群集的均方誤差，做為下一次疊代的標準。

表 5-1：帶聲調音節樣本數與基數對應表

樣本個數 *n*	**1**	**2**	**3~10**	**11~100**	**101~500**	**>501**
樣本個數為 *n* 的帶聲調音節數	44	36	96	450	504	125
基數 n_k	1	2	3	5	10	15

為了有效率地分配樣本數量給不同的帶聲調音節，我們採用貪婪演算法，逐步增加樣本數量給均方誤差最大的帶聲調音節。在每次分配樣本數量給帶聲調音節後，我們將 1,255 種帶聲調音節的均方誤差依大小排序(Sorting)，取誤差最大的前 10 名增加一個基數的樣本分配，接著重覆前面步驟直到樣本總數超過 20,000 個為止。圖 5-2 為疊代的流程圖。

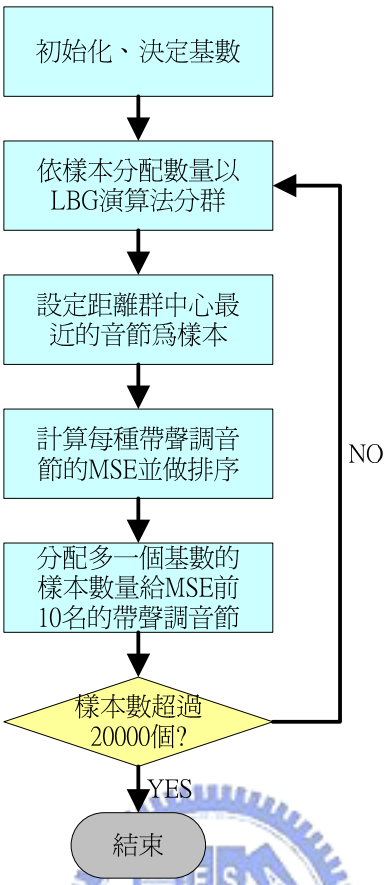


圖 5-2：樣本選取流程圖

進一步分析此樣本選取演算法的效能，我們將每種帶聲調音節的均方誤差繪成直方圖如圖 5-3，發現在初始化時的樣本總數為 9,275 個，均方誤差介於 0 和 0.45 之間，總均方誤差約為 0.16；而收斂時的樣本總數為 20,115 個，均方誤差則介於 0 和 0.14 之間，總均方誤差約為 0.137，誤差明顯縮小。值得注意的是，在本系統中我們將音節基頻軌基和長度正規化，使二者權重相同且變異數為 1。

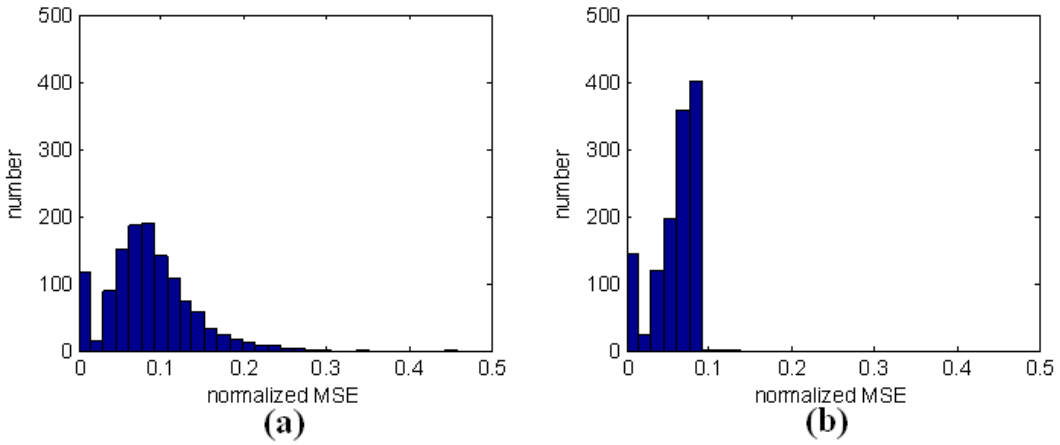


圖 5-3：帶聲調音節樣本均方誤差的(a)初始化和(b)收斂結果長條圖

此外，我們將樣本總數和總均方誤差的關係繪於圖 5-4，可看出總均方誤差隨著樣本總數增加而下降的關係。

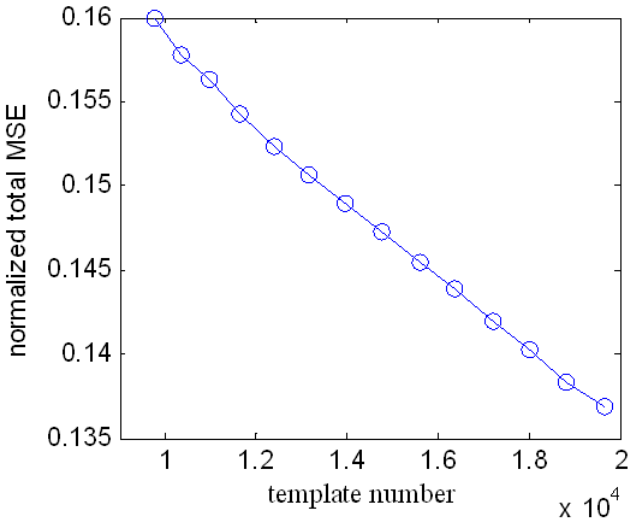


圖 5-4：樣本總數和樣本總均方誤差之關係

5.1.2 韻律產生

在使用者輸入中文單詞後，系統會對其進行分析，將詞中每個字所對應的參數找出，如音調、和前後音節的音調組合、字在詞中的位置、基本音節類別等，

然後依前面訓練好的模型產生預測的韻律參數。預測韻律參數的數學式如下：

$$\mathbf{sp}_n = \mathbf{sp}_n^r + \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}^f_{c_{n-1},tp_{n-1}} + \boldsymbol{\beta}^b_{c_n,tp_n} + \boldsymbol{\beta}_{w_n} + \boldsymbol{\mu}^p \tag{3-2}$$

$$sd_n = sd_n^r + \gamma_{t_n} + \gamma^f_{c_{n-1},tp_{n-1}} + \gamma^b_{c_n,tp_n} + \gamma_{w_n} + \gamma_{sy_n} + \mu^d \tag{4-2}$$

$$se_n = se_n^r + \alpha_{t_n} + \alpha^f_{c_{n-1},tp_{n-1}} + \alpha^b_{c_n,tp_n} + \alpha_{w_n} + \alpha_{sy_n} + \mu^d \tag{4-18}$$

對於音節間連音狀態的預測，我們已在 4.5.2 小節中用決策樹分析，可直接由表 4.7 得到預測結果。

值得一提的是，中文的破音字常造成輸入的問題，例如「長度」和「兄長」二個詞中的「長」分別發音為「ㄔㄤˊ」和「ㄓㄤˇ」，如果直接以詞中每個字分別對應的音碼輸入系統，可能會產生錯誤。為了避免這個問題，我們將帶有破音字的詞整理出來，並且在查詢音碼時優先以詞為單位進行查詢比對，即可為每個破音字找到正確的音碼。

此外，在本論文中我們模擬的語音為女聲，而女生的韻律參數和男生有些許不同，特別是女聲的基頻軌跡普遍會比男聲高。因此在產生目標基頻軌跡時，我們可供使用者選擇，將基頻軌跡的平均值和標準差調適至男性的標準，數學式如下：

$$Y = \frac{X - \mu_X}{\sigma_X} \times \sigma_Y + \mu_Y \tag{5-1}$$

其中 X 和 Y 分別為原(女聲)基頻和調適後(男聲)基頻的數值；μ_X 和 μ_Y 分別為女聲和男聲基頻數值的平均值；而 σ_X 和 σ_Y 分別為女聲和男聲基頻數值的標準差。女聲的 μ_X 和 σ_X 可由單詞語料庫直接求取，而男聲的 μ_Y 和 σ_Y 則需另外準備：我們事先錄製四位男生各 5,520 字、長約 30 分鐘的平行語料，並分別求取其基頻的平均值和標準差；最後再取四位男生統計數值的平均值做為 μ_Y 和 σ_Y。由表

5-2 我們可得知 μ_X = 206.18 Hz，σ_X = 40.43 Hz，μ_Y = 112.51 Hz，σ_Y = 19.39 Hz。圖 5-5 為基頻調適的示意圖。

表 5-2：男女聲基頻數值統計特性(Hz)

	男生 1	男生 2	男生 3	男生 4	男生平均	女生
μ	122.38	95.06	110.41	122.20	112.51	206.18
σ	17.73	19.98	20.49	19.35	19.39	40.43

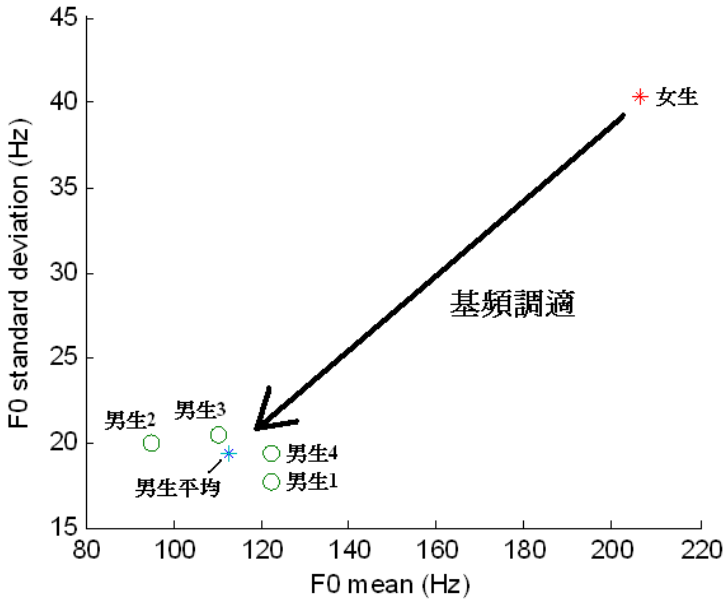


圖 5-5：基頻調適示意圖

5.1.3 單元選取

由上一小節產生韻律參數後，我們進行 TTS 系統中的單元選取。在本論文中我們合成的單詞語音是由一個個帶聲調音節所串接而成，而單元選取的目的在於從多個樣本中選出最適合的單元，以合成出自然悅耳的語音。

傳統的單元選取方式訂定差異函式(Cost function)，主要分為合成單元目標差異(Target cost)和合成單元間轉移差異(Transition cost)二種，其中前者表示合成單

元和目標韻律參數之間的差距；而後者則顯示出合成語音是否流暢自然。為了使整體合成語音的品質達到最好，我們訂定各種差異的權重(Weight)，並靠大量的計算找出整體差異最小的單元組合。

在本論文中，我們利用前面建立的韻律模型取代傳統的做法，除了具有物理意義外，亦能大量降低計算量。如式(3-2)和式(4-2)，我們的音節基頻軌跡和長度模型考慮了聲調(如 $\boldsymbol{\beta}_t$ 和 γ_t)、前後音節的連音影響(如 $\boldsymbol{\beta}^f_{c,t}$ 、$\boldsymbol{\beta}^b_{c,t}$ 、$\gamma^f_{c,tp}$ 和 $\gamma^b_{c,tp}$)、音節在詞中的位置(如 $\boldsymbol{\beta}_w$ 和 γ_w)和基本音節類別(如 γ_{sy})等四種影響因素；而在 TTS 系統中單元選取的樣本為帶聲調音節，已包含聲調和基本音節類別；所以在選取單元時只需考慮前後音節的連音影響、音節在詞中的位置即可。因此我們依最小差異法則訂定以下選取標準：

$$\begin{aligned}\Phi_n = \arg\min_m \{ &(\boldsymbol{\beta}^f_{c_{m-1},t_{m-1}} - \boldsymbol{\beta}^f_{c_{n-1},t_{n-1}})^{\mathrm{T}} (\boldsymbol{\beta}^f_{c_{m-1},t_{m-1}} - \boldsymbol{\beta}^f_{c_{n-1},t_{n-1}}) + (\boldsymbol{\beta}^b_{c_m,t_m} - \boldsymbol{\beta}^b_{c_n,t_n})^{\mathrm{T}} (\boldsymbol{\beta}^b_{c_m,t_m} - \boldsymbol{\beta}^b_{c_n,t_n}) \\ &+ (\boldsymbol{\beta}_{w_m} - \boldsymbol{\beta}_{w_n})^{\mathrm{T}} (\boldsymbol{\beta}_{w_m} - \boldsymbol{\beta}_{w_n}) + (\gamma^f_{c_{m-1},tp_{m-1}} - \gamma^f_{c_{n-1},tp_{n-1}})^2 + (\gamma^b_{c_m,tp_m} - \gamma^b_{c_n,tp_n})^2 + (\gamma_{w_m} - \gamma_{w_n})^2 \}\end{aligned} \tag{5-1}$$

其中 Φ_n 為欲合成單詞中第 n 個帶聲調音節所選取的單元；m 為該種帶聲調音節的樣本總數；$\boldsymbol{\beta}_{w_n}$ 為欲合成音節在詞中位置的目標影響因素之樣式(Target affecting factor pattern)；$\boldsymbol{\beta}_{w_m}$ 為第 m 個樣本在原資料庫中音節在詞中位置的源目標影響因素之樣式(Source affecting factor pattern)；其餘項則以此類推。

由(5-1)式我們可分別挑選出欲合成單詞中每個音節的最佳樣本，並且同時考慮到我們韻律模型中所有的影響因素。除此之外，由於此方法中所有的目標影響樣式和所有樣本的源影響樣式在進行單元選取前皆為已知，因此所有的相似度計算都可以事先算好，並將數值建立成一個差異表(Cost table)；在進行單元選取時直接查表將幾個數值相加即可完成相似度的計算。

5.1.4 合成單元後製處理

在前一小節我們完成了單元選取，在此小節中我們為合成單元進行後製處理，分別為變調處理、隨機亂數機制、能量調整、淡入與漸消、音節間停頓長度預測，分述如下：

一、**變調處理**：中文是聲調語言(Tone language)，聲調的高低會直接影響語意內容。在中文裡有一個很著名的變調規則，即連續兩個三聲的音節在一起時，第一個三聲音節會轉為二聲，如「雨傘」、「總統」等。為了模擬此現象，我們將變調的規則加入單元選取的限制，亦即碰到連續兩個三聲的音節時，只有同樣發生變調的單元會被選擇。

二、**隨機亂數機制**：在進行韻律模擬時，我們發現二個音節即使擁有相同的影響樣式(包含聲調、連音影響和基本音節類別等)，亦會在音節基頻韻律軌跡、長度和能量上有不同的表現；我們認為這是語者在說話時可控制的隨機因素(Random factor)，可能受到語意或其它環境因素所影響。為了模擬此現象，我們在 TTS 系統進行單元選取時為目標韻律參數加上一個呈高斯分佈的亂數，而此亂數的數值則和第 3.3.6 小節的模擬誤差相當。經由此機制，每次 TTS 系統合成的語音皆有些微差異，在音質損失極少的情況下使合成語音具有多樣性。

三、**能量調整**：在我們的 TTS 系統中單元選取只考慮單元的基頻軌跡和長度；而能量部分則是對單元乘上一個倍率進行即時調整，使得單元的能量和韻律模型預測出的目標能量(Target energy)相當。由以下數學式我們可推導出能量調整倍率 α ：

$$E_t = 10\log_{10}\frac{\sum(w_i x_{t,i}^2)}{N}$$

$$E_s = 10\log_{10}\frac{\sum(w_i x_{s,i}^2)}{N}$$

$$E_{s'} = 10\log_{10}\frac{\sum(w_i(\alpha x_{s,i})^2)}{N} = 10\log_{10}\alpha^2 + E_s \qquad (5\text{-}2)$$

$$let\ \ E_t = E_{s'}:$$

$$\alpha = 10^{\frac{E_t - E_s}{20}}$$

其中 w 為的漢明窗的大小；x 為聲音波型訊號；i 表示計算能量的音框中第 i 個點；N 為音框的大小；E_t 為目標能量；E_s 為選取單元原本的能量；$E_{s'}$ 則為選取單元乘上調整倍率 α 的能量。我們令 $E_t = E_{s'}$ 則可導出調整倍率的值。

四、**淡入與漸消**：由於進行單元選取的樣本皆為從中文單詞語料庫所截取而來，因此在單元的起始和結尾不一定為 0，特別是該單元在原本語料庫中和前後音節連音現象嚴重的時候。此現象會導致合成語音時因波型不連續而讓人有聽覺突兀的感覺，因此我們分別對合成單元的前 1/10 和後 1/10 進行淡入(Fade-in)和漸消(Fade-out)處理以改善合成音質。

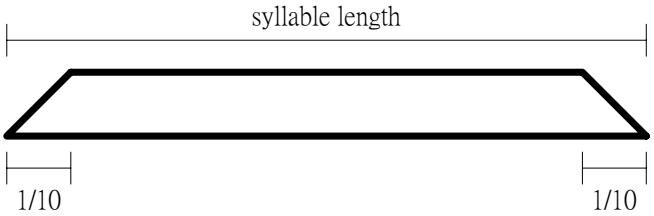


圖 5-6：對合成單元做淡入與漸消示意圖

五、**音節間停頓長度預測**：在 4.5.1 小節中我們對音節間停頓長度進行決策樹分析，並將結果整理於表 4.7，供 TTS 系統直接查詢使用。統合此小節所有步驟後，我們即完成單詞語音的合成。

5.1.5 錄音訊號處理

在合成範例語音後，使用者可以用麥克風錄下自己的聲音，並且由系統抽取音節基頻軌跡、長度和能量等韻律參數，顯示於螢幕上提供使用者參考，讓使用者經由比較目標韻律參數和源韻律參數的差異後，經由反饋(Feedback)反覆調整，達成學習的效果。

首先我們以前述的 ESPS 演算法進行線上(On-line)的基頻求取，並進一步驗證基頻的可靠性。接著我們以 HTK 的「強制切割」(Forced alignment)功能依照輸入的單詞內容對語音進行切割，以求取每一個音節的長度。在進行強制切割前，我們事先以 TCC300 語料庫訓練好一個語者非相關(Speaker independent)的語音 HMM 模型，該模型以音節為單位，每個音節有八個狀態(State)，每個狀態是平均由八個高斯分佈所組成的高斯混合模型(Gaussian Mixture Model, GMM)。最後，我們以(2-7)式求取能量，其音框大小設為 240 個取樣點。

由於在本論文中所模擬的能量是指音節「韻母部分的能量位準最大值」，而我們假設在前一步驟切割音節時，八個狀態中的前三個是聲母，後五個是韻母，因此音節後五個狀態的能量最大值即為所求。

5.2 系統展示

在前一節中我們介紹了系統的架構與實作細節，在此節中我們將展示系統的介面。如圖 5-7，我們將系統介面分為六大部分，詳細內容如下：

一、**主功能區**：此區包含系統的所有功能選擇，包括文字的輸入(含目標語音韻律參數求取)、播放 TTS 範例、錄使用者的聲音、播放使用者的聲音(含源語音韻律參數求取)。此外，在 5.1.2 和 5.1.4 小節中我們討論了模型誤差模擬和男女生模型轉換的功能，在此系統中可自由選擇。

二、**文字與連音狀態區**：此區會顯示使用者輸入的文字內容，並以表 4-7 預測出單詞中音節間的連音狀態。

三、**音節長度區**：此區會顯示目標語音音節長度和源語音音節長度。

四、**音節能量區**：同上，此區會顯示目標語音音節能量和源語音音節能量。

五、**語音波形區**：在合成範例語音時，此區會繪出該範例語音的波形；而在使用者錄音後，此區會顯示該錄音的波形。

六、**音節基頻軌跡區**：同上，在合成範例語音時，此區會繪出目標語音基頻軌跡；而在使用者錄音後，此區會同時顯示目標語音基頻軌跡和源語音基頻軌跡。

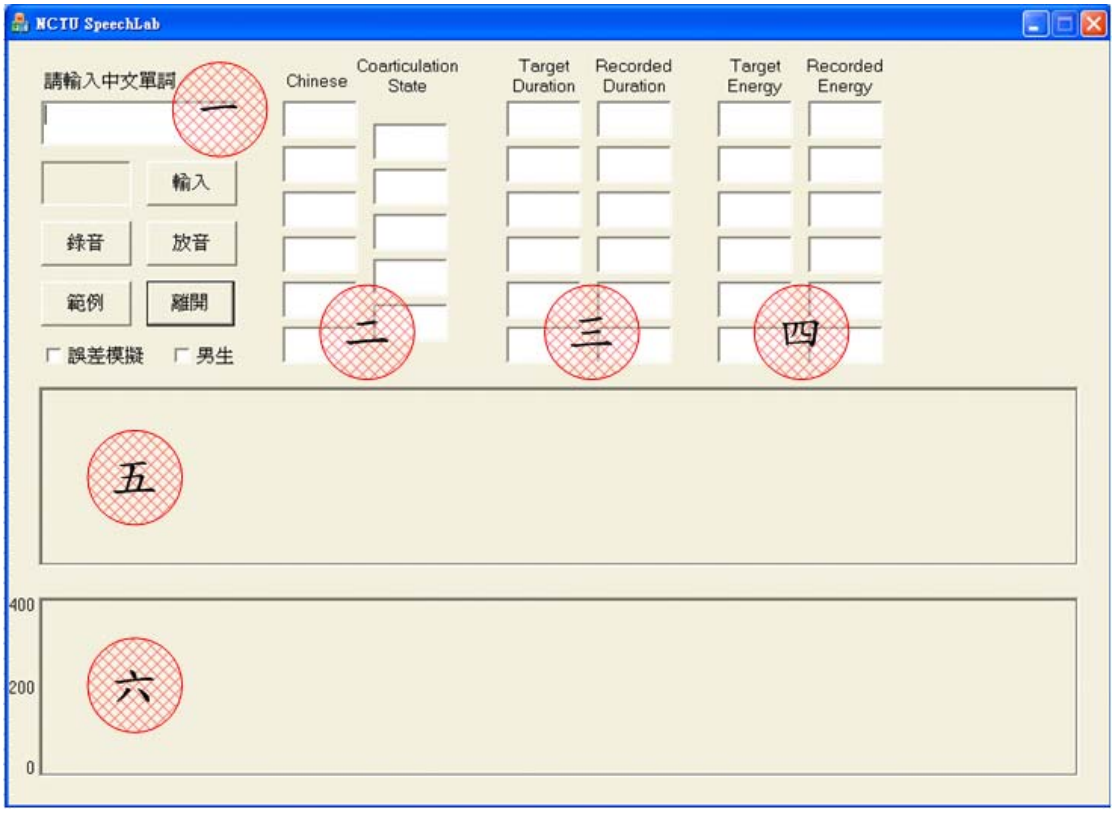


圖 5-7：系統展示圖之一(介面概觀)

我們以「交通大學」一詞為例，當輸入單詞並按「輸入」鍵後，螢幕上會顯示目標語音的音節基頻軌跡、長度和能量。接著按「範例」鍵後，系統會播放合成語音，並且把語音波形繪於螢幕上，如圖 5-8。

接著我們按「錄音」鍵開始錄音，結束後再按一次即停止；此時系統會自動對源語音進行強制切割得到音節長度，並求取基頻軌跡和能量。最後按「放音」鍵後系統會將源語音的波形和韻律參數顯示在螢幕上，供使用者比較學習，如圖5-9。值得注意的是，此時基頻軌跡部分有二種，「實線」為目標語音的基頻軌跡，而「圓點」為源語音的基頻軌跡。

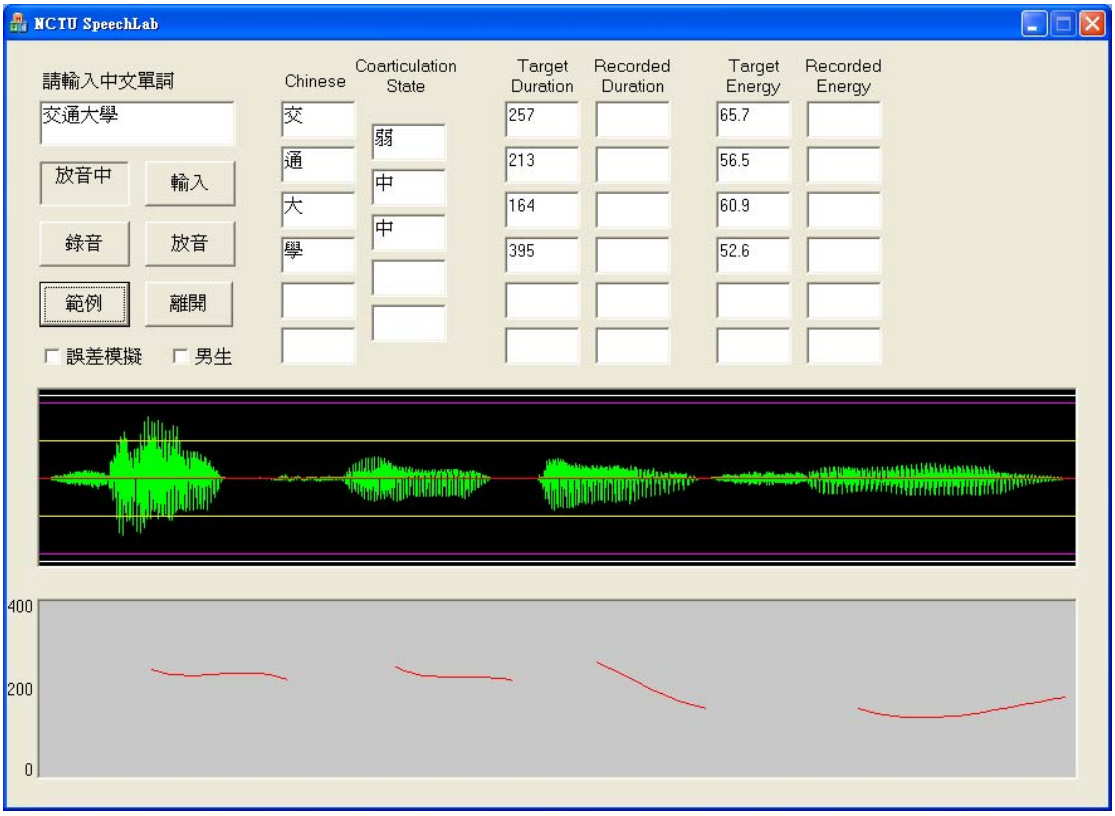


圖 5-8：系統展示圖之二(目標語音與韻律參數合成)

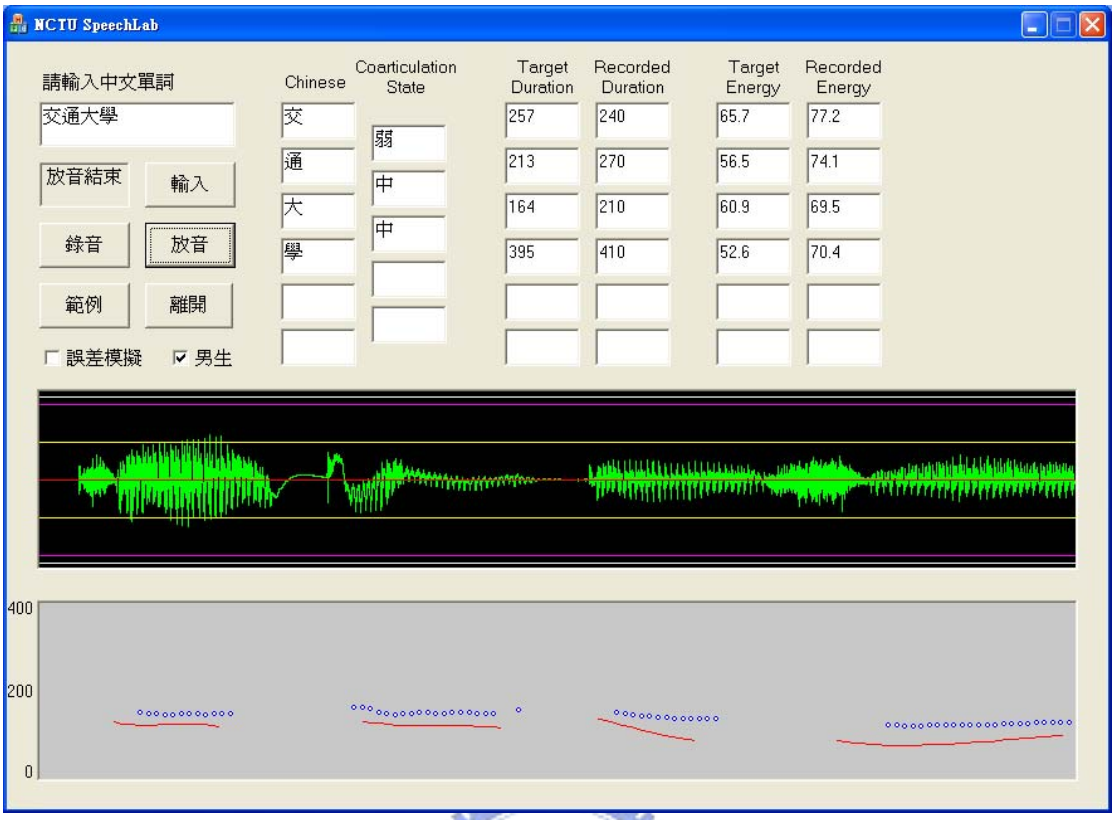


圖 5-9：系統展示圖之三(源語音與韻律參數求取)

第六章 結論與未來展望

6.1 結論

本論文主要分為二個部分，在論文的前半部我們考慮音節的聲調、與前後音節的連音影響、音節在詞中的位置和基本音節類別等影響因素，建立了中文單詞的音節基頻軌跡、長度和能量等三種韻律模型，並分析各種影響因素的物理意義和模擬的誤差；實驗結果顯示此模型能有效模擬此三種韻律參數。

在論文的後半部我們以韻律模型建立一套中文韻律學習系統，供非中文母語的使用者學習。使用者可依需要輸入單詞，系統會自動合成該單詞之語音和三種韻律參數讓使用者模仿；並且在使用者口說錄音後進行強制切割等處理，提供使用者相關韻律資訊回饋學習。

6.2 未來展望

此中文韻律系統仍有許多地方可以擴充，例如我們可提供英文單詞的輸入功能，再依中英文詞典翻譯成中文單詞，讓不會中文輸入的外國人亦能使用系統；我們亦可建立評分系統，依使用者輸入語音的韻律和品質給分，並進一步提供修正指示；此外，我們亦能加入基頻同步疊加法(Pitch Synchronous Overlap and Add, PSOLA)，將使用者的錄音經調整至目標韻律後再播放，讓使用者能聽自己的聲音來學習，更添效果；最後我們可以找幾位中文學習人士進行試用，並依其反應和回饋改進系統。

參考文獻

[1] 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, no.9, p1309-1319, Sep. 1989.

[2] L. S. Lee, C. Y. Tseng, and C. J. Hsieh, ”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.

[3] S. H. Chen, 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.

[4] The HTK Book (for HTK Version 3.4)

[5] S. H. Chen and Y. R. Wang, “Vector Quantization of Pitch Information in Mandarin Speech,” *IEEE Trans. Communications*, vol. 38, no.9, pp.1317-1320, Sept. 1990.

[6] Y. R. Chao, *A Grammar of Spoken Chinese*. Berkeley Press, University of California, Berkeley, CA, 1968.

[7] C. Y. Chiang, H. M. Yu, Y. R. Wang and S. H. Chen, “An Automatic Prosody Labeling Method for Mandarin Speech,” *Proc. of Interspeech 2007*, Antwerp, Belgium, pp. 494-497.

[8] Y. Xu, “Contextual Tonal Variations in Mandarin,” *J. Phonetics*, vol. 25, no. 1, pp. 61-83, 1997.

[9] Y. Xu, “Sources of Tonal Variations in Connected Speech,” *Journal of Chinese Linguistics*, monograph series #17. 1-31, 2001.

[10] Y. Yuan, M. J. Shaw, “Induction of fuzzy decision trees,” *Fuzzy Sets and Systems*, vol.69, pp.125-139, 1995.

[11] 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

[12] Y. Linde, A. Buzo and R. Gray, “An Algorithm for Vector Quantizer Design,” *IEEE Transactions on Communications*, vol. 28, pp. 84-94, 1980.

[13] S. H. Chen, W. H. Lai and Y. R. Wang, “A Statistics-based Pitch Contour Model for Mandarin Speech" , *J. Acoust. Soc. Am.*, 117 (2), pp.908-925, 2005.

[14] S. H. Chen, W. H. Lai, and Y. R. Wang, “A New Duration Modeling Approach for Mandarin Speech” , *IEEE Trans. On Speech and Audio Processing*, vol. 11, no. 4, 2003.

[15] C. Y. Chiang, H. M. Yu, Y. R. Wang and S. H. Chen, “Exploration of High-level Prosodic Patterns for Continuous Mandarin Speech”, Proc. of *ICASSP 2008*, Las Vegas, U.S.A., pp.3977-3980.

[16] C. F. Chen, C. Y. Chiang, Y. R. Wang and S. H. Chen, “A Study on Prosodic Modeling for Isolated Mandarin Words,” *Proc. of ROCLING XVIV*, Taipei, Taiwan, pp. 273-286, 2007.

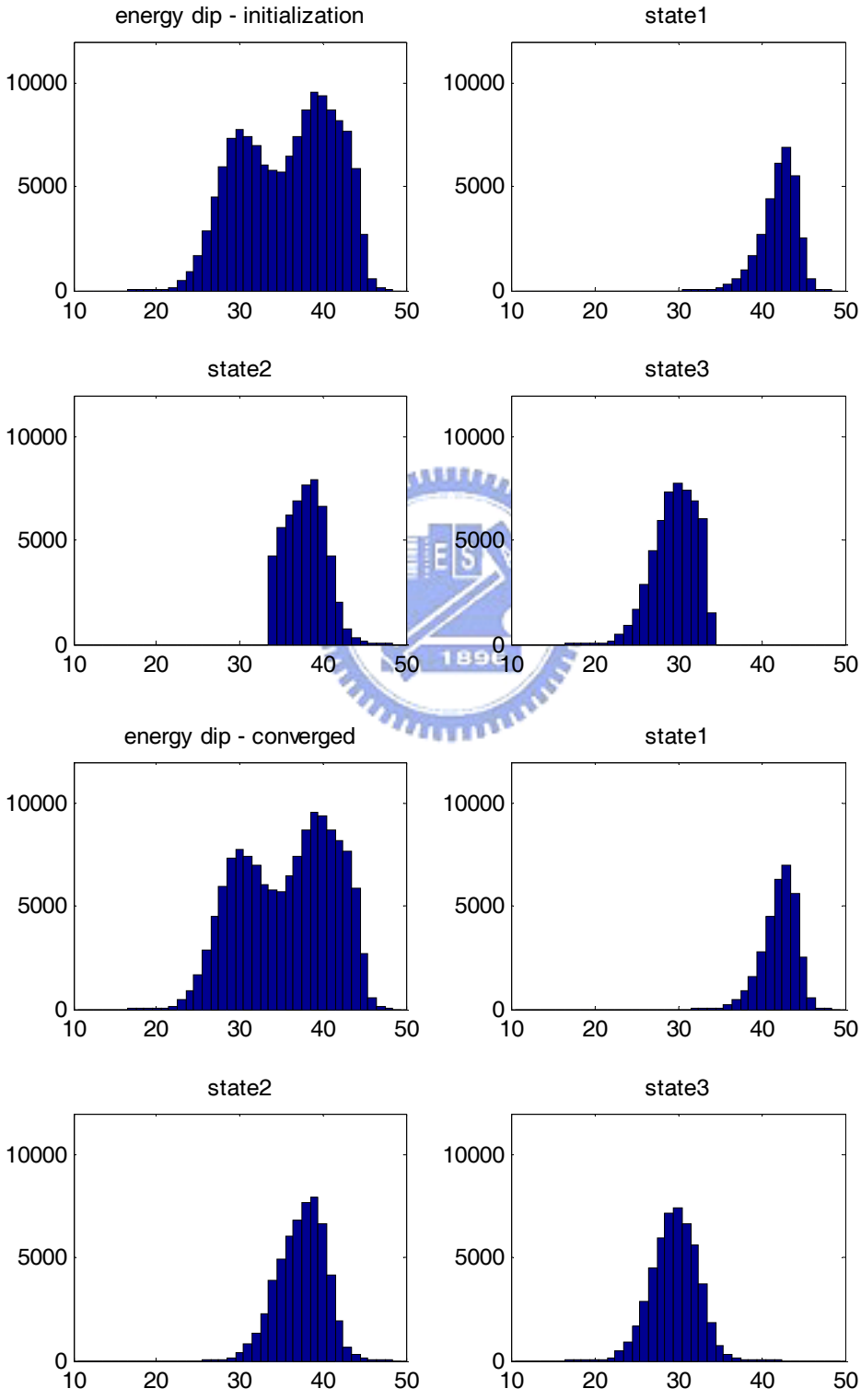
[17] 陳啟風，“中文單詞之韻律模式研究”，國立交通大學碩士論文，民國九十六年七月。

[18] 洪國興，“以語料庫為基礎之中文文句翻語音系統實現”，國立交通大學碩士論文，民國九十五年八月。

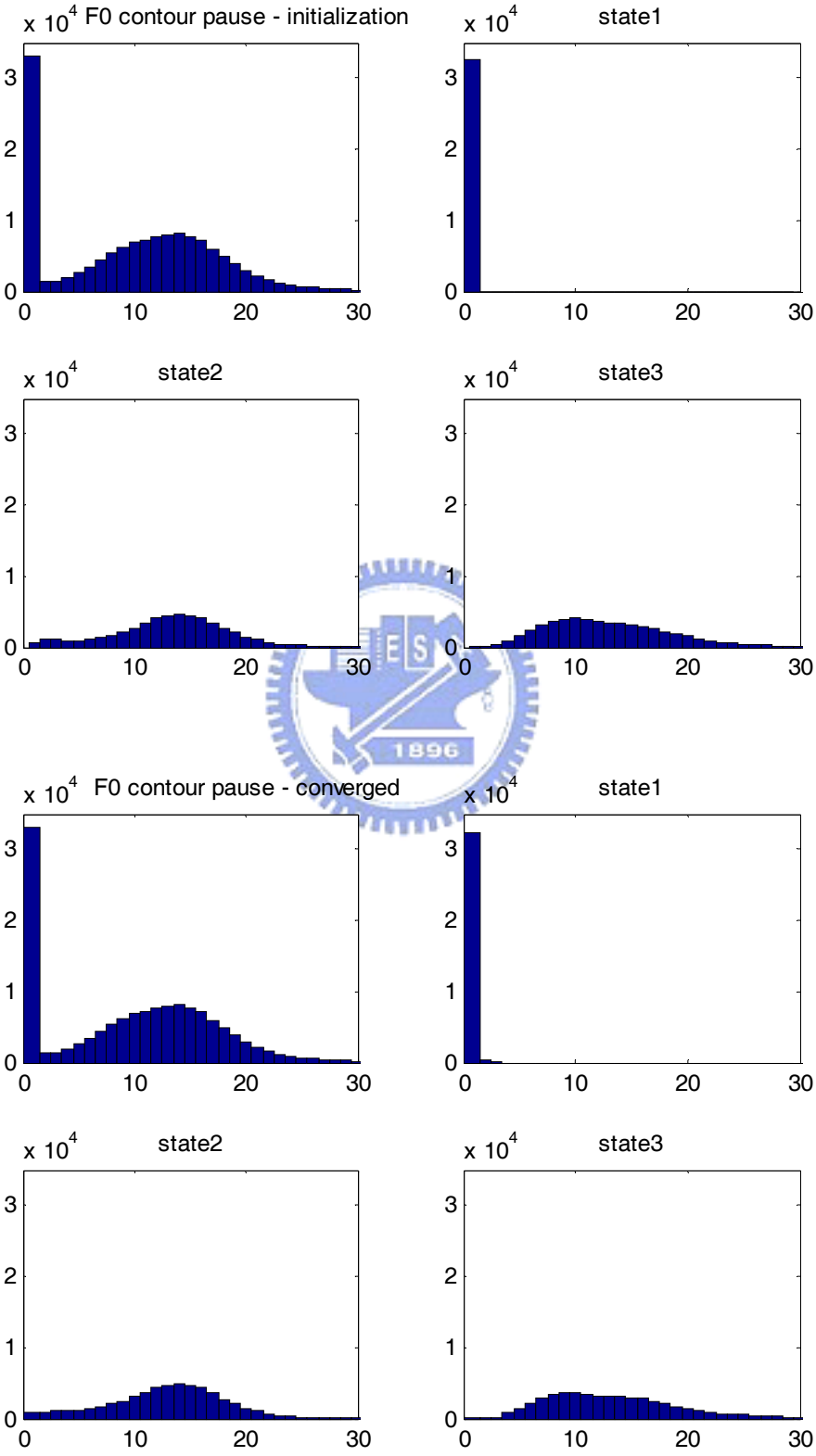
附錄

四種輔助韻律參數初始和收斂數值分佈圖

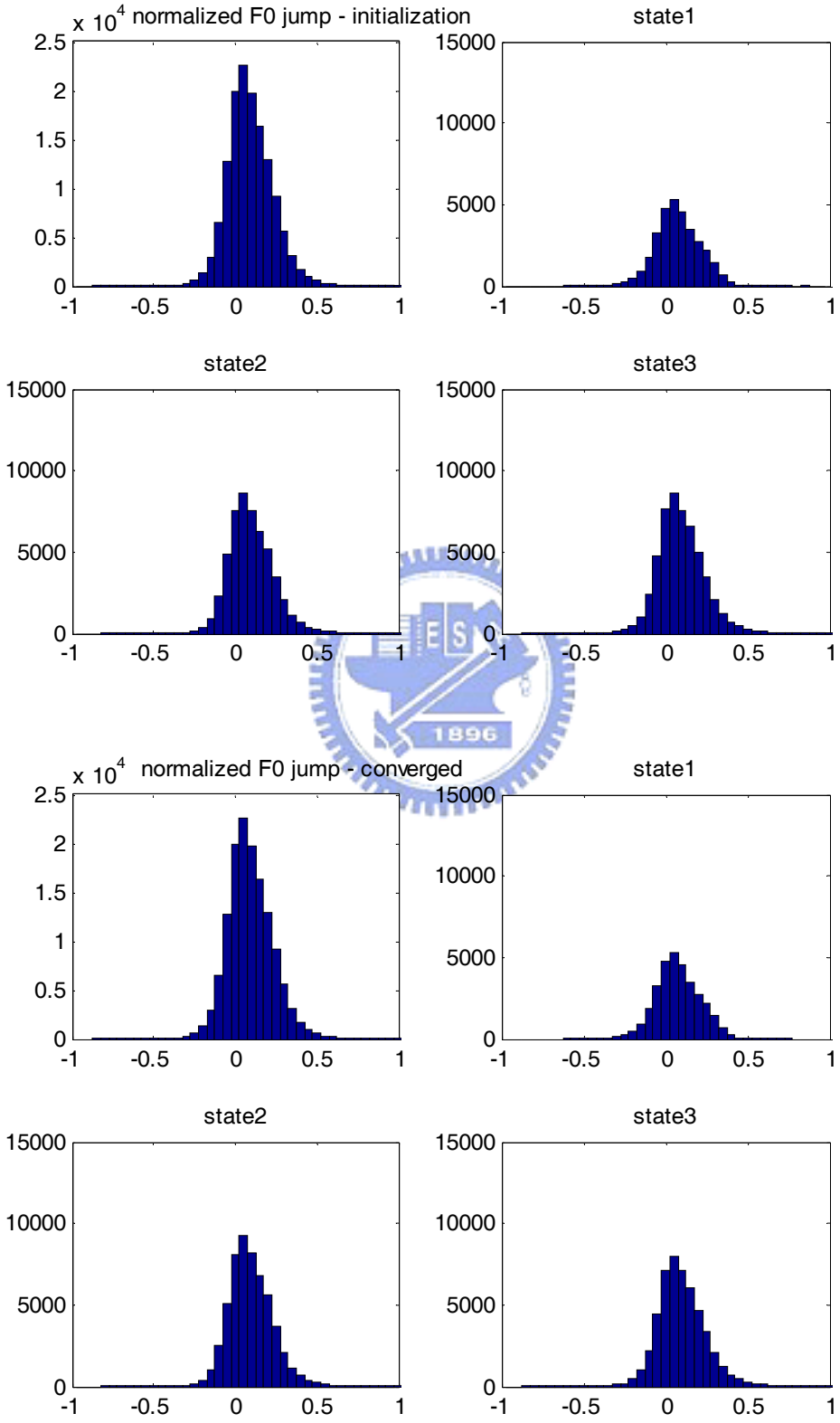
一、能量低點(Energy dip) unit:dB



二、基頻軌跡停頓(F0 contour pause) unit:10ms



三、正規化基頻差(Normalized F0 jump) unit:log-Hz



四、停頓長度(Pause duration) unit:ms