國 立 交 通 大 學

電信工程研究所

碩 士 論 文

多語者漢語韻律模型之建立與其在語者韻律轉換之應用

Multi-Speaker Mandarin Speech Prosody Modeling and its Application to Speaker Prosody Conversion

研究生：劉子睿

指導教授：陳信宏 博士

中 華 民 國 一百零二 年 七 月

多語者漢語韻律模型之建立與其在語者韻律轉換之應用

Multi-Speaker Mandarin Speech Prosody Modeling and its Application to Speaker Prosody Conversion

研 究 生：劉子睿　　Student：Tzu-Jui Liu

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

國 立 交 通 大 學

電 信 工 程 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Communication Engineering

College of Electrical and Computer Engineering

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in Communication Engineering

July 2013

Hsinchu, Taiwan, Republic of China

中 華 民 國 一百零二 年 七 月

多語者漢語韻律模型之建立與其在語者韻律轉換之應用

研 究 生：劉子睿　　指導教授：陳信宏 博士

國立交通大學電信工程研究所碩士班

中文摘要

本研究提出以語者韻律模式為基礎的語者韻律轉換方法，其系統架構包含語者韻律模式訓練及聲音韻律轉換兩階段。語者韻律模式訓練階段可再分成語者獨立韻律模型訓練及語者相依韻律模型調適兩個部分，它首先以 PLM 演算法訓練一個語者獨立韻律模型並對訓練語料產生韻律及停頓標記；接著以最大事後機率調適法則將語者獨立韻律模型調適成語者相關韻律模型，並以遞迴方式反覆疊代更新兩類模型直到收斂；聲音韻律轉換階段則包含來源語者韻律分析及目標語者韻律合成，它使用來源語者之語者相關韻律模型來分析輸入語音之韻律信息，以產生韻律標記，然後以目標語者之語者相關韻律模型來合成輸出語音之韻律參數，包括音節基頻軌跡、音節長度、音節能量、及音節間停頓長度。本研究實驗使用自行錄製的部分平行語料庫，包含 9 男 6 女的朗讀語音，實驗結果顯示我們所提出的方法轉換效果略優於傳統的高斯正規化法，並且在部分 Source 語者及 Target 語者韻律狀態的影響數值相差劇烈之處，可以產生補償的效果。

Multi-Speaker Mandarin Speech Prosody Modeling and its Application to Speaker Prosody Conversion

Student：Tzu-Jui Liu　　Advisor：Dr. Sin-Horng Chen

Institute of Communication Engineering
National Chiao Tung University

Abstract

In this thesis, a speech prosody conversion method based on speaker's prosody modeling is proposed. The method comprises a prosody modeling phase and a prosody conversion phase. In the prosody modeling phase, the PLM algorithm proposed previously is firstly employed to train an SI prosodic model from a multi-speaker training dataset and label all training utterances with prosodic states for all syllables as well as break types for all syllable junctures. Then, the maximum a posterior probability (MAP) method is applied to adapt the SI prosodic model to generate a speaker dependent (SD) prosodic model for each speaker. In the prosody conversion phase, the SD prosodic model of the source speaker is firstly used to analyze the input speech to generate prosodic tags. Then, the prosody of the output speech is generated using these prosodic tags by the SD prosodic model of the target speaker. The prosodic information generated includes syllable pitch contour, syllable duration, syllable energy level, and syllable-juncture pause duration. A corpus containing read speeches of six female and nine male speakers was used to examine the validity of the proposed method. Experimental results confirmed that the proposed method performed slightly better than the conventional Z-score normalization method.

致謝

整整兩年的碩士生涯，好不容易從懵懂到略懂，身分準備從老屁股變回社會上的死菜鳥，或多或少都有點擔憂，但是，畢業無庸置疑是件快樂的事！

碩一時聽著小老闆諸多口頭禪，也為咪挺的難熬多添了一份味道；碩二以後有了明確的論文方向，卻也開始出現一個又一個的瓶頸，所幸大老闆總是可以點出重要的地方幫助我又突破難關，真的很感謝兩位老師這兩年的諄諄教誨。

這兩年在 707 認識了很多新的朋友，summer school 就給人留下深深印象的靖觀，恭喜你從那時候的"我單身唷！"到現在幸福的模樣，707 大大小小的事情都跟妳脫離不了關係，沒有這個俏秘書兼開心果，生活一定會黯淡許多！一起運動喝飲料聊天的雷雷夥伴阿龐，你獨特的親和力讓人想不喜歡你都難，和你相處很輕鬆，之後也都還在新竹打拼，一起加油！號稱 707 趙又廷的良基，游泳健身沒一項難的倒你實在太威猛，希望總是缺一個女朋友的你早日找到親愛的另一半！天然傻的婉君，從碩一到現在妳的個性被磨練的圓滑不少，懂得了更多人情世故，以妳的能力未來工作一定可以得心應手，然後開心的做自己喜歡的事！反詐騙專家奕勳，流利的口才把詐騙唬得一愣一愣著實大快人心，辛苦你接到了實驗室最困難的自發性語音，相信你一定可以克服層層困難！

感謝這兩年陪伴我的學長學弟們，韻律之神 kiwi、正妹女朋友護體的睿銓、咦咦咦的深夜小王子 DD、上知天文下之地理的企鵝、強到一年半就可以出國玩樂的子軒、鞋子總是超高調螢光的王柏(一直有個秘密沒告訴你，第一次為了教你修音檔，我被開紅單！)、做事效率一百分的小七忠實顧客小鋒、又高又帥又認真的人生勝利組阿駿、惦惦吃三碗公的仲毛、嘴巴跟腦袋一樣厲害的阿璋、心想事成的 ML、來無影去無蹤的佩樺…還有很多無法一一列舉的人兒，謝謝你(妳)們。

感謝花了大半輩子養育我的父母及和我一起歡樂長大的弟弟，沒有你們在我的背後給予我支持，我絕對無法順利完成學業；最後，總是陪我度過難關低潮，不厭其煩的聆聽我再帶給我笑容和前進的勇氣，雖然被稱做不道德，但成為邏設助教再認識了妳，這一步一步看似不可思議的過程，卻是我最珍貴的回憶，吳小儀謝謝妳。

最後，這本論文帶著我無限的感謝與祝福，獻給你(妳)們。

目錄

表目錄

圖目錄

第一章　緒論

1.1 研究動機

科技始終來自於人性，隨著人類對生活品質的逐步追求，各式各樣的技術與應用紛紛被提出和廣泛的使用，例如：smart phone、smart TV、平板電腦…等等，這些電子產品也潛移默化的改變著整個人類的生活習慣，而語音處理技術憑藉著使用的便利性逐漸取代以往繁瑣的鍵盤輸入及文字輸出，其重要性不言而喻。

語音合成系統是為了能讓機器發出像人類說話的聲音而發展出的技術；隨著隨藏式馬可夫模型為基礎(HMM-based)的文字轉語音(Text-To-Speech, TTS)技術的興起，現今已經可以合成出品質頗佳的聲音，而其中語音韻律部分的掌握是十分重要的環節。

若能在語音合成系統上，運用聲音轉換的技術，並透過目標語者的語料，使電腦能隨意地轉換成不同語者說話韻律之特性，不僅能夠使合成的聲音富有多樣性，也將有助於許多語音相關的應用發展。

1.2 文獻回顧

最常見的聲音轉換就是語者聲音轉換[1]，其目的為希望 Source 語者的聲音經由轉換後可以接近 Target 語者的聲音，整個轉換過程可以簡單分成頻譜轉換及韻律轉換兩部份；近年由於文字轉語音技術的發展，聲音轉換技術也開始應用於其後端，藉由這兩項技術，當需要合成不同語者的聲音時，不再需要重新取得大量新語者的語料來建構新的 TTS 系統，取而代之的是僅需要少量的訓練語料，並利用聲音轉換的方式重新合成新語者的聲音[2,3]。除此之外，聲音轉換的技術也被廣泛的應用在許多方面，例如歌唱聲音的轉換[4]，以窄頻訊號預估寬頻訊號[5]，以及情緒語音的轉換[6,7]。

在過去已有許多有關於聲音轉換的方法被提出[8-10]，[8]提出使用向量量化(Vector

Quantization, VQ)為基礎的轉換，同時對 Source 及 Target 語者的訓練語料做 VQ 並建立碼本(codebook)及統計每個碼字對應所占的權重，轉換時只要利用碼字及權重，以線性組合建立轉換即可。[1]提出以高斯混合模型(Gaussian Mixture Model, GMM)為基礎的轉換，後續許多研究都是基於此方法提出改進，其中以 A. Kain 等人[9]所提出的改進方法最常被引用，其想法為利用 GMM 描繪 Source 及 Target 語者的特徵參數，並以此建立轉換函式。[3,10]提出了以隱藏式馬可夫模型(Hidden Markov Model, HMM)為基礎的轉換，此方法引進了時間的概念，利用信號在 HMM 狀態上的變換，以不同的轉換函式進行聲音轉換。

在早期的聲音轉換研究上，主要探討以頻譜轉換為主，而近年來開始有越來越多學者對於韻律轉換的研究提出許多新的方法。根據韻律研究的文獻，語音的韻律結構是由階層式的架構(Hierarchical structure)所組成[11,12]，因此近年來學者開始運用韻律及語言學的知識輔助韻律的轉換。

1.3 研究方法概述

聲音轉換主要可分成頻譜轉換和韻律轉換兩個部份；本研究提出一個以語者韻律模式為基礎之語者韻律轉換的新作法，它以非監督式中文語音韻律標記及韻律模式(unsupervised joint Prosody Labeling and Modeling, PLM)演算法[12]為基礎，進行最大事後機率 (Maximum A Posterior Probability, MAP) [13-15]之韻律模型調適，以產生語者相依之韻律模型，作為語者語音韻律轉換之用，其系統架構可分為語者獨立韻律模型訓練、韻律模型調適及語者韻律參數轉換三個部分。

在語者獨立模型訓練部分，為避免個別語者的韻律特性太突顯，使得語者獨立韻律模型不夠中性，我們先做韻律聲學參數正規化的動作，再以 PLM 演算法對語料標示韻律狀態及停頓標記並建立韻律模型；在模型調適部分，我們使用最大事後機率調適法則調適語者獨立模型中的各韻律聲學參數，以產生語者相依之韻律模型，並利用所有語者調適過的參數疊代更新韻律模型中的部分子模型；最後，利用調適過後的語者相關韻律模型進行語者韻律參數轉換，以

Source 語者之韻律模型分析輸入語音之韻律，產生韻律標記，再使用 Target 語者之韻律模型合成輸出語音的韻律參數，並以實驗來驗證此方法的有效性。

1.4 章節概要說明

本論文一共分為五章，其各章節內容分配如下：

第一章：緒論

第二章：基於韻律模型之語者調適

第三章：漢語語者韻律轉換

第四章：實驗結果與分析

第五章：結論與未來展望

第二章　基於韻律模型之語者調適

本章討論以前人所提出之 HPM [12]為基礎，利用最大事後機率將語者獨立的韻律模型調適為語者相依的韻律模型。

2.1 漢語語音階層式韻律架構

依據語言學家的研究，語音的韻律結構是呈階層式架構，前人依此概念提出韻律標記的概念並定義了階層式多短語韻律句群(Hierarchical Prosodic Phrase Grouping, HPG)架構[11]，如圖 2.1 所示，最底層為音節層次(SYLlable layer, SYL)，也代表漢語最基本的字義，其中聲調為最強烈的影響因素，不只影響音節基頻軌跡之走向，也影響了音節長度及能量位階；往上依序為韻律詞層次(Prosodic Word layer, PW)，由雙音節或多音節所構成的詞組，通常在句法和語意上關係緊密，結尾常會帶有不明顯但可察覺之停頓；韻律短語層次(Prosodic Phrase layer, PPh)，由一或多個韻律詞所組成，結尾會帶有明顯可察覺之停頓；呼吸組層次(Breath Group, BG)，由單一或數個韻律短語組成的句子，其結尾通常帶有明顯停頓；最上層為韻律組句(Prosodic phrase Group, PG)，由一個或數個呼吸組構成。

停頓標記是用來區分韻律組成份子的邊界，*B0* 和 *B1* 區分了 SYL 的邊界，其中 *B0* 表示 reduced syllabic boundary，而 *B1* 表示 normal syllabic boundary，這兩種停頓類別通常都不具明顯停頓；*B2* 和 *B3* 分別是韻律詞和韻律短語的邊界；*B4* 則代表了呼吸組的邊界，和 *B2*、*B3* 比較起來會有較明顯的停頓；至於 *B5* 定義了韻律句組邊界，代表一個完整的段落結束，通常句尾會有音節長度拉長(final lengthening)及能量減弱等現象。

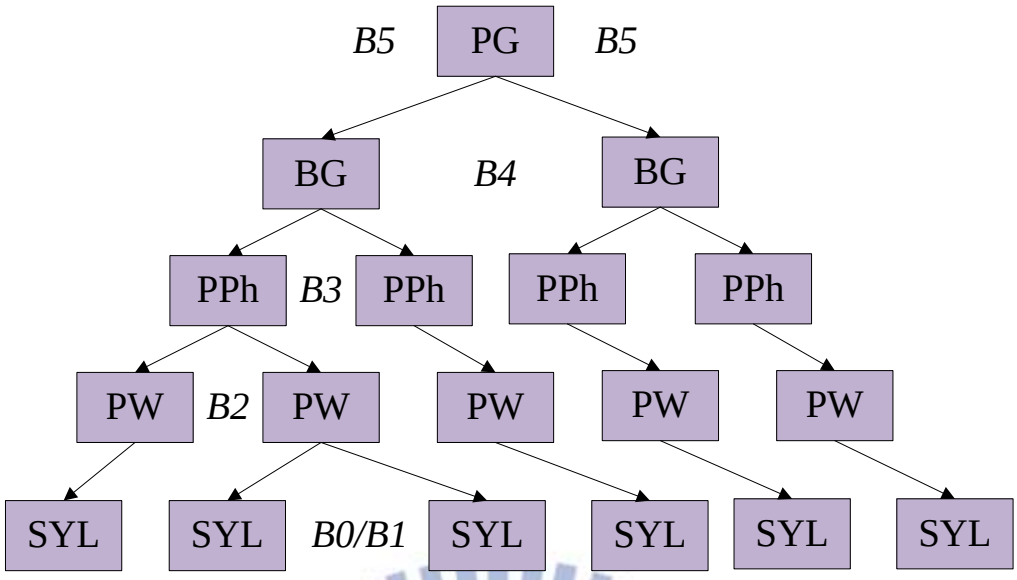


圖 2.1：中文語音韻律階層式架構概念 [16]

本研究使用之語料庫為大段落的語音，因此以 HPG 架構為基礎，經過進一步的修改後，利用此韻律階層架構來建立本研究所提出之韻律模型。首先將 *B2* 再細分為 *B2-1*、*B2-2*、*B2-3*，分別代表明顯音高重置(pitch reset)、短停頓(short pause)及含有音節拉長效應(duration lengthening)之韻律詞邊界等不同現象。接著因為 BG 和 PG 所描述的韻律特性相近，將這兩層合併為同一層，*B4* 則和 *B5* 合成為 *B4*。整個架構從 5 層變成 4 層，如圖 2.2 所示。最後採用的 7 種韻律邊界停頓(break type)為 **B**={ *B0*, *B1*, *B2-1*, *B2-2*, *B2-3*, *B3*, *B4*}，以此來標記四種韻律單元：音節(SYL)、韻律詞(PW)、韻律短語(PPh)、呼吸組/韻律句組(BG/PG)。

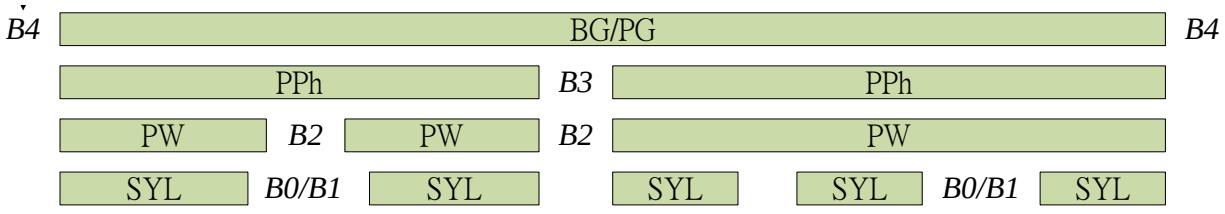


圖 2.2：本研究所採用之階層式韻律架構 [12]

2.2 語者獨立韻律模型之建立方法

圖 2.3 為本研究所提出之語者獨立韻律模型(Speaker Independent prosodic model, SI prosodic model)訓練流程圖，其中 x'_k 為第 k 位語者的原始訓練音檔，$k=1\sim K$ ；x_k 為第 k 位語者正規化後的音檔；Λ_{SI} 為語者獨立韻律模型。

首先，我們分別從每位語者的聲音資料中抽取個別的韻律聲學特徵參數；接著分別針對不同參數對每位語者進行韻律聲學特徵參數正規化，目的是避免因個別語者的韻律特性太突顯，使得語者獨立韻律模型不夠中性；最後使用 PLM 演算法來訓練語者獨立韻律模型，同時產生韻律標記。

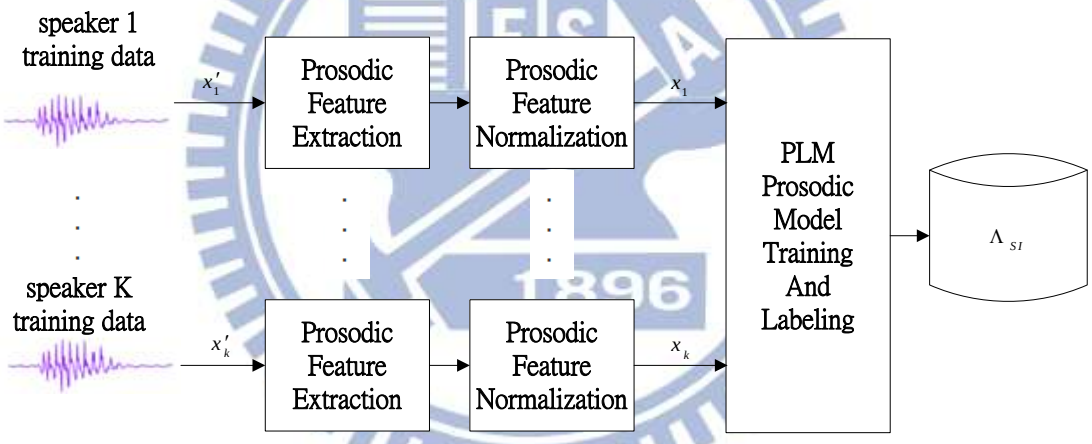


圖 2.3：本研究所提出之語者獨立韻律模型訓練流程圖

2.3 韻律聲學特徵參數正規化

為消除不同語者個別的韻律特性以訓練一個語者獨立韻律模型，因此對於每位語者，分別針對其音節長度 sd 、停頓時長 pd 、音節基頻軌跡 $Logf0$ 以及音節能量 se 提出以下的正規化方法。

2.3.1 音節長度正規化

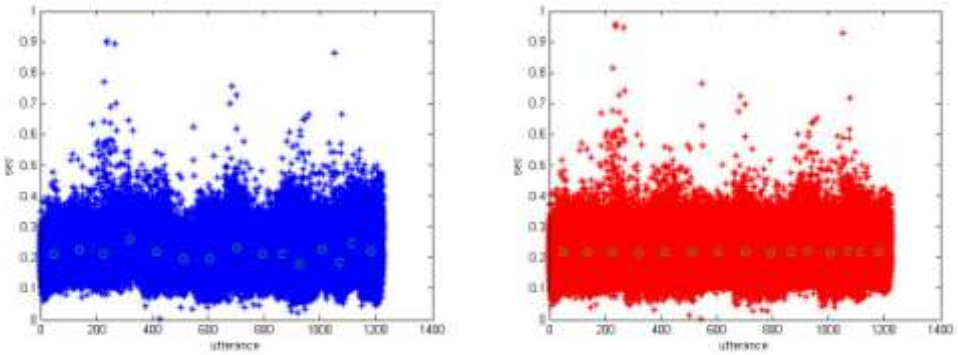
在韻律聲學特徵參數中，漢語音節長度可近似於高斯分布(Gaussian distribution)，故我們使用高斯正規化法，另外為了避免 outliers 的影響，我們採取強建性的正規化方法，詳細步驟如下：

1. 將個別語者的音節長度做排序，扣除所有音節長度中最短 5%的音節長度及最長 5%的音節長度資料
2. 利用 1.所得到的音節長度，求取個別語者的音節長度之 mean 和 standard deviation
3. 利用 2.所求得的所有語者的音節長度之 mean 和 standard deviation，將其分別再取平均值，當作 average speaker 的 mean 和 standard deviation
4. 依據 2.和 3.求出的 mean、standard deviation、average speaker mean、average speaker standard deviation 做正規化，其正規化函數如下所示：

$$sd = \frac{sd' - \mu_s^{sd}}{\sigma_s^{sd}} \sigma_g^{sd} + \mu_g^{sd} \qquad (2\text{-}1)$$

其中 sd' 及 sd 是修正前後的音節長度，μ_s^{sd} 和 σ_s^{sd} 是語者音節長度的 mean 以及 standard deviation，μ_g^{sd} 和 σ_g^{sd} 是音節長度的 average speaker mean 和 average speaker standard deviation.

透過 2-1 式可以將全部語者的音節長度都對應到同樣的平均和範圍。圖 2.4(a)為未正規化前所有語者音節長度的分布圖，圖 2.4(b)為正規化後所有語者音節長度的分布圖。



(a) (b)

圖 2.4：(a)未正規化前所有語者音節長度分布，(b)正規化後所有語者音節長度分布

2.3.2 停頓時長正規化

經由觀察停頓時長的分布，我們發現伽瑪分布(Gamma distribution)比高斯分布更適於模擬停頓時長的分布，故我們使用伽瑪正規化法。另外為了避免 outliers 的影響，我們採取強建性的正規化方法，詳細步驟如下：

1. 經由觀察發現，極短的停頓時長因數量相對龐大會大幅主導整個正規化參數，導致無法有效的正規化，故我們扣除 5ms 以下的停頓時長
2. 將扣除 5ms 以下的個別語者停頓時長做排序，再扣除所有停頓時長中最長 5%的停頓時長資料
3. 利用 2.所得到的停頓時長，求取個別語者的停頓時長之 mean 和 standard deviation
4. 利用 3.所求得的所有語者的停頓時長之 mean 和 standard deviation，將其分別再取平均值，當作 average speaker 的 mean 和 standard deviation
5. 利用 mean 和 standard deviation 就可以算出 gamma distribution 的 alpha 和 beta 如下式：

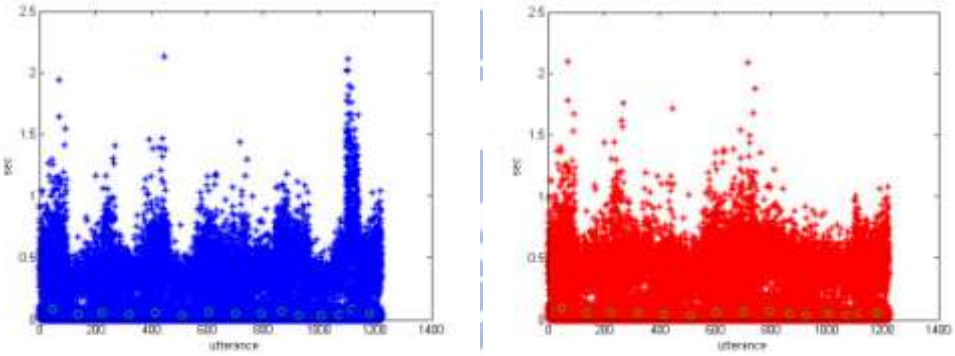
$$\alpha = \frac{\mu^2}{\sigma^2} \quad , \quad \beta = \frac{\sigma^2}{\mu} \tag{2-2}$$

6. 利用 2-2 式所求出的 alpha 和 beta 參數做正規化，其正規化函數如下所示：

$$pd = G^{-1}(G(pd';\alpha_s^{pd},\beta_s^{pd});\alpha_g^{pd},\beta_g^{pd}) \tag{2-3}$$

其中 pd' 及 pd 是修正前後的停頓時長，$G(pd';\alpha,\beta)$ 為伽碼分佈累積密度函數(Cumulative Density Function, CDF)，G' 為 G 之反函數，α_s^{pd} 和 β_s^{pd} 是利用語者停頓時長的 mean 以及 standard deviation 所求出的 gamma 參數，α_g^{pd} 和 β_g^{pd} 是利用停頓時長的 average speaker mean 和 average speaker standard deviation 所求出的 gamma 參數

透過 2-3 式可以將全部語者的停頓時長都對應到同樣的平均和範圍。圖 2.5(a)為所有語者原始停頓時長的分布圖，圖 2.5(b)為正規化後所有語者停頓時長的分布圖，圖中的圈圈代表每位語者的停頓時長平均值



(a) (b)

圖 2.5：(a)未正規化前所有語者停頓時長分布，(b)正規化後所有語者停頓時長分布

2.3.3 音節基頻軌跡正規化

與音節長度正規化類似，我們使用高斯正規化法對音節基頻軌跡做正規化；而與音節長度稍有不同的是，在此我們先把原始 frame-based 的基頻轉成對數形式後，再對此 frame-based 的對數基頻做正規化動作，詳細步驟如下：

1. 將個別語者的對數基頻做排序，扣除所有對數基頻中最小 5%的對數基頻及最大 5%的對數基頻資料
2. 利用 1.所得到的對數基頻，求取個別語者的對數基頻之 mean 和 standard deviation
3. 利用 2.所求得的所有語者的對數基頻之 mean 和 standard deviation，將其分別再取平均值，

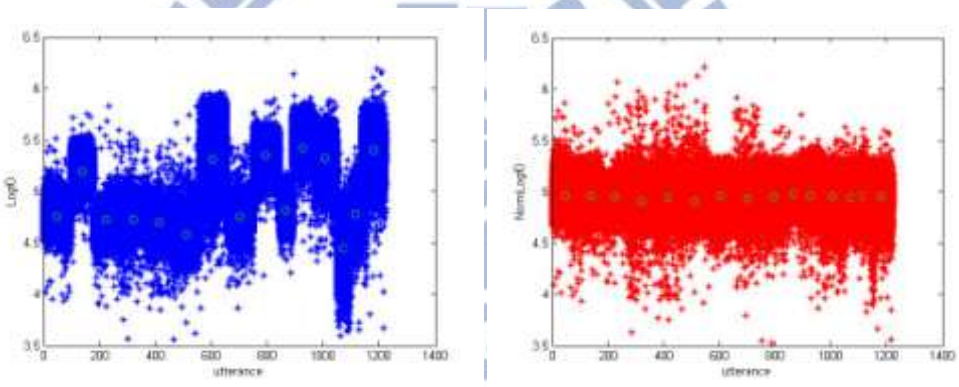
當作 average speaker 的 mean 和 standard deviation

4. 依據 2.和 3.求出的 mean、standard deviation、average speaker mean、average speaker standard deviation 做正規化，其正規化函數如下所示：

$$f = \frac{f' - \mu_s^f}{\sigma_s^f}\sigma_g^f + \mu_g^f \qquad (2\text{-}4)$$

其中 f' 及 f 是修正前後的音節長度，μ_s^f 和 σ_s^f 是語者音節長度的 mean 以及 standard deviation，μ_g^f 和 σ_g^f 是音節長度的 average speaker mean 和 average speaker standard deviation

透過 2-4 式可以將全部語者的對數基頻都對應到同樣的平均和範圍。圖 2.6(a)為所有語者原始對數基頻的分布圖，圖 2.6(b)為正規化後所有語者對數基頻的分布圖。



(a) (b)

圖 2.6：(a)未正規化前所有語者對數基頻分布，(b)正規化後所有語者對數基頻分布

2.3.4 音節能量正規化

與音節長度正規化類似，我們使用高斯正規化法對音節能量做正規化，詳細步驟如下：

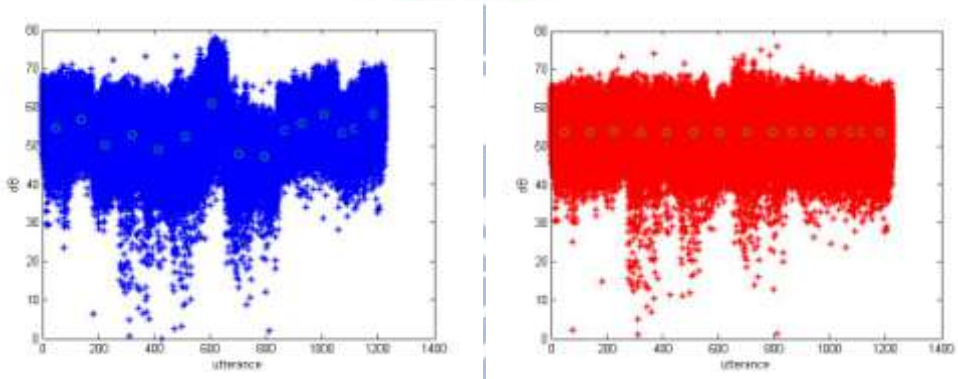
1. 將個別語者的音節能量做排序，扣除所有音節能量中最小 5%的音節能量及最大 5%的音節能量資料
2. 利用 1.所得到的音節能量，求取個別語者的音節能量之 mean 和 standard deviation
3. 利用 2.所求得的所有語者的音節能量之 mean 和 standard deviation，將其分別再取平均值，當作 average speaker 的 mean 和 standard deviation

4. 依據 2.和 3.求出的 mean、standard deviation、average speaker mean、average speaker standard deviation 做正規化，其正規化函數如下所示：

$$se = \frac{se' - \mu_s^{se}}{\sigma_s^{se}} \sigma_g^{se} + \mu_g^{se} \tag{2-5}$$

其中 se' 及 se 是修正前後的音節能量，μ_s^{se} 和 σ_s^{se} 是語者音節能量的 mean 以及 standard deviation，μ_g^{se} 和 σ_g^{se} 是音節能量的 average speaker mean 和 average speaker standard deviation

透過 2-5 式可以將全部語者的音節能量都對應到同樣的平均和範圍。圖 2.7(a)為未正規化前所有語者音節能量的分布圖，圖 2.7(b)為正規化後所有語者音節能量的分布圖。



(a) (b)

圖 2.7：(a)未正規化前所有語者音節能量分布，(b)正規化後所有語者音節能量分布

2.4 基於 MAP 調適之韻律模型調適

2.4.1 最大事後機率調適法則

MAP 的基本精神在於結合了事前機率以及調適語料來估測出新的模型參數，假設我們要依觀察到的資料 x 去評估一未知的母體參數(unobserved population parameter) θ，而資料 x 的抽樣分佈(sampling distribution)為 $f(\cdot)$ 且存在一 θ 的事前分佈(prior distribution) $g(\cdot)$，則 θ 的事後分佈(posterior distribution)可表示為：

$$f(\theta \mid x) = \frac{f(x \mid \theta) g(\theta)}{\int_{\theta' \in \Theta} f(x \mid \theta') g(\theta') d\theta'} \tag{2-6}$$

其中 Θ 為 $g(\cdot)$ 的定義域。而 MAP 法則可表示為：

$$\hat{\theta}_{MAP}(x) = \arg\max_{\theta} \frac{f(x|\theta)g(\theta)}{\int_{\theta' \in \Theta} f(x|\theta')g(\theta')d\theta'} = \arg\max_{\theta} f(x|\theta)g(\theta) \quad (2\text{-}7)$$

2.4.2 韻律模型調適與更新

PLM 演算法可視為一個韻律標記過程，並同時更新模型參數。在給定語料庫之韻律聲學特徵參數集合 **A**、相對應的語言參數集合 **L** 之下，找出一組最佳韻律標記集合 **T**，整個過程可以看成一參數最佳化問題，即

$$\mathbf{T}^* = \arg\max_{\mathbf{T}} P(\mathbf{T}|\mathbf{A},\mathbf{L}) = \arg\max_{\mathbf{T}} P(\mathbf{T},\mathbf{A}|\mathbf{L}) \quad (2\text{-}8)$$

韻律標記集合 **T={B, PS}**包含兩種重要的韻律訊息，第一種為音節邊界的停頓標記(Break Type)，用來表示階層式架構的韻律組成份子邊界，本論文定義韻律邊界停頓標記集合為 **B**={*B0, B1, B2-1, B2-2, B2-3, B3, B4*}；第二種韻律標記為音節韻律狀態分為 **PS**={**p, q, r**}，其所代表意義分別為音節基頻韻律狀態 **p**、音節長度韻律狀態 **q** 及音節能量韻律狀態 **r**。

本論文韻律聲學參數**A={X, Y, Z}**分為兩類，第一類為音節本身的聲學參數**X={sp, sd, se}**，分別為音節基頻軌跡 **sp**、音節長度 **sd** 及音節能量位階 **se**，本研究假設此類聲學參數與韻律狀態標記有很大相關性，與音節邊界停頓標記相關性非常小，本論文稱 **X** 為音節韻律參數(syllable prosodic feature)；第二類為音節邊界的聲學參數**{Y, Z}={pd, ed, pj, dl, df}**，分別為音節邊界的停頓時長(pause duration, *pd*)、能量低點位階(energy-dip level, *ed*)、正規化基頻差(normalized pitch jump, *pj*)及兩種正規化長度拉長因子(normalized duration lengthening factor, *dl* and *df*)等，假設此類型的聲學參數與停頓標記有很大相關性，與韻律狀態標記的相關性很小，本論文稱 **Y={pd, ed}**為音節邊界韻律參數(syllable-juncture prosodic feature)、**Z={pj, dl, df}**為音節間差分韻律參數(inter-syllable differential prosodic feature)。

在語言參數方面，用 **L** 表示所有的語言參數集合。其中特別將音節聲調、基本音節類型與韻母類型從 **L** 中獨立出來，用意在於這三種語言參數對音節基頻軌跡、音節長度及音節能量位階有顯著的影響，把剩餘的語言參數統一定義為 **l**(reduced linguistic feature set)。完整的符

號定義整理於表 2.1。

表 2.1：韻律標記、聲學參數及語言參數之表示符號 [12]

T : prosodic tag	**B** : break type={*B0*, *B1* ,*B2-1*, *B2-2*, *B2-3*, *B3*, *B4*}	
	PS : prosodic state	**p** : pitch prosodic state **q** : duration prosodic state **r** : energy prosodic state
A : prosodic feature	**X** : syllable prosodic feature	**sp** : syllable pitch contour **sd** : syllable duration **se** : syllable energy level
	Y :inter-syllabic prosodic feature	**pd** : pause duration **ed** : energy-dip level
	Z : differential prosodic features	**pj** : normalized pitch jump **dl** : normalized duration lengthening factor 1 **df** :normalized duration lengthening factor 2
L : linguistic feature	**l**: reduced linguistic feature set	
	t: syllable tone sequence **s**: base-syllable type sequence **f**: final type sequence	

綜合上述之討論，可將 $P(\mathbf{T},\mathbf{A}|\mathbf{L})$ 改寫成以下形式：

$$\begin{aligned} P(\mathbf{T},\mathbf{A}|\mathbf{L}) &= P(\mathbf{A}|\mathbf{T},\mathbf{L})P(\mathbf{T}|\mathbf{L}) = P(\mathbf{X},\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{PS},\mathbf{L})P(\mathbf{B},\mathbf{PS}|\mathbf{L}) \\ &\approx P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})P(\mathbf{PS}|\mathbf{B})P(\mathbf{B}|\mathbf{L}) \end{aligned} \tag{2-9}$$

其中 $P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})$ 稱為音節聲學模型，用來敘述音節韻律參數受到停頓標記 **B**、韻律狀態 **PS** 和語言參數 **L** 之間的影響而產生的變化；$P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$ 稱為停頓聲學模型，用以敘述在各個不同停頓標記 **B** 和語言參數 **L** 下，其韻律邊界的聲學特性；$P(\mathbf{PS}|\mathbf{B})$ 稱為韻律狀態模型，描述了韻律狀態在不同停頓標記 **B** 下的轉移變化；$P(\mathbf{B}|\mathbf{L})$ 稱為停頓語法模型，描述在不同的語言參數 **L** 下，各種停頓標記出現的頻率。

PLM 演算法是基於最大概似度法則(Maximum Likelihood, ML)，對所有語句找出最佳的韻律

標記，並估計模型參數。依據上述各子模型我們定義一目標函數(objective function)如下：

$$Q=\left(\prod_{n=1}^{N} p(\mathbf{sp}_n \mid p_n, B_{n-1}^{n}, t_{n-1}^{n+1})\, p(sd_n \mid q_n, t_n, s_n, u_n)\, p(se_n \mid r_n, t_n, f_n, u_n)\right)$$
$$\left(p(p_1)\,p(q)\,p(r_1)\prod_{n=2}^{N} p(p_n \mid p_{n-1}, B_{n-1})\, p(q_n \mid q_{n-1}, B_{n-1})\, p(r_n \mid r_{n-1}, B_{n-1})\right) \quad (2\text{-}10)$$
$$\left(\prod_{n=1}^{N-1}\left(p(pd_n, ed_n, pj_n, dl_n, df_n \mid B_n, \mathbf{l}_n)\, p(B_n \mid \mathbf{l}_n)\right)\right)$$

圖 2.8 為本研究所提出之基於 MAP 調適之韻律模型轉換流程圖，其中 $B_k^{(i)}$ 為第 k 位語者經過第 i 次遞迴以後的停頓標記，$k=1\sim K$；$P_k^{(i)}$ 為第 k 位語者經過第 i 次遞迴以後的韻律狀態標記，$k=1\sim K$；L_k 為第 k 位語者的語言資料，$k=1\sim K$；$\Lambda_{SA}^{(i)}$ 為第 i 次遞迴以後的音節聲學模型；$\Lambda_{BA}^{(i)}$ 為第 i 次遞迴以後的停頓聲學模型；$\Lambda_{PS}^{(i)}$ 為第 i 次遞迴以後的韻律狀態模型；$\Lambda_{BS}^{(i)}$ 為第 i 次遞迴以後的停頓語法模型。

我們將整個流程分成 **Part A** 和 **Part B**，以下針對這兩部分做詳細的說明。**Part A**，對每位語者，我們利用圖 2.3 訓練好的語者獨立韻律模型 Λ_{SI} 與每位語者個別的訓練音檔，利用 MAP 調適其音節聲學模型 Λ_{SA} 中各個 APs 與停頓聲學模型 Λ_{BA} 中的各個參數，使其成為該語者相關的音節聲學模型與停頓聲學模型；接著分別對每位語者利用語者獨立韻律模型、每位語者個別的訓練音檔、調適完的音節聲學模型中的各 APs 與停頓聲學模型中的各個參數和語言資料 L_k，更新該語者語音的停頓標記 B_k 和韻律狀態標記 P_k。

Part B，首先，我們結合 **Part A** 所有語者的新停頓標記 B_k 和新韻律狀態標記 P_k，並加上每位語者的語言資料 L_k，更新語者獨立韻律模型中的韻律狀態模型 Λ_{PS}；接著我們利用更新過後的停頓標記 B_k 和語言資料 L_k，更新停頓語法模型 Λ_{BS} 的樹狀結構和停頓聲學模型 Λ_{BA} 中每個子模型的樹狀結構；最後將語者獨立韻律模型 Λ_{SI} 中的韻律狀態模型 Λ_{PS}、停頓語法模型 Λ_{BS} 的樹狀結構和停頓聲學模型 Λ_{BA} 用 **Part A** 和 **Part B** 更新過後的模型替代。

最後，我們重複 **Part A** 和 **Part B** 的動作。

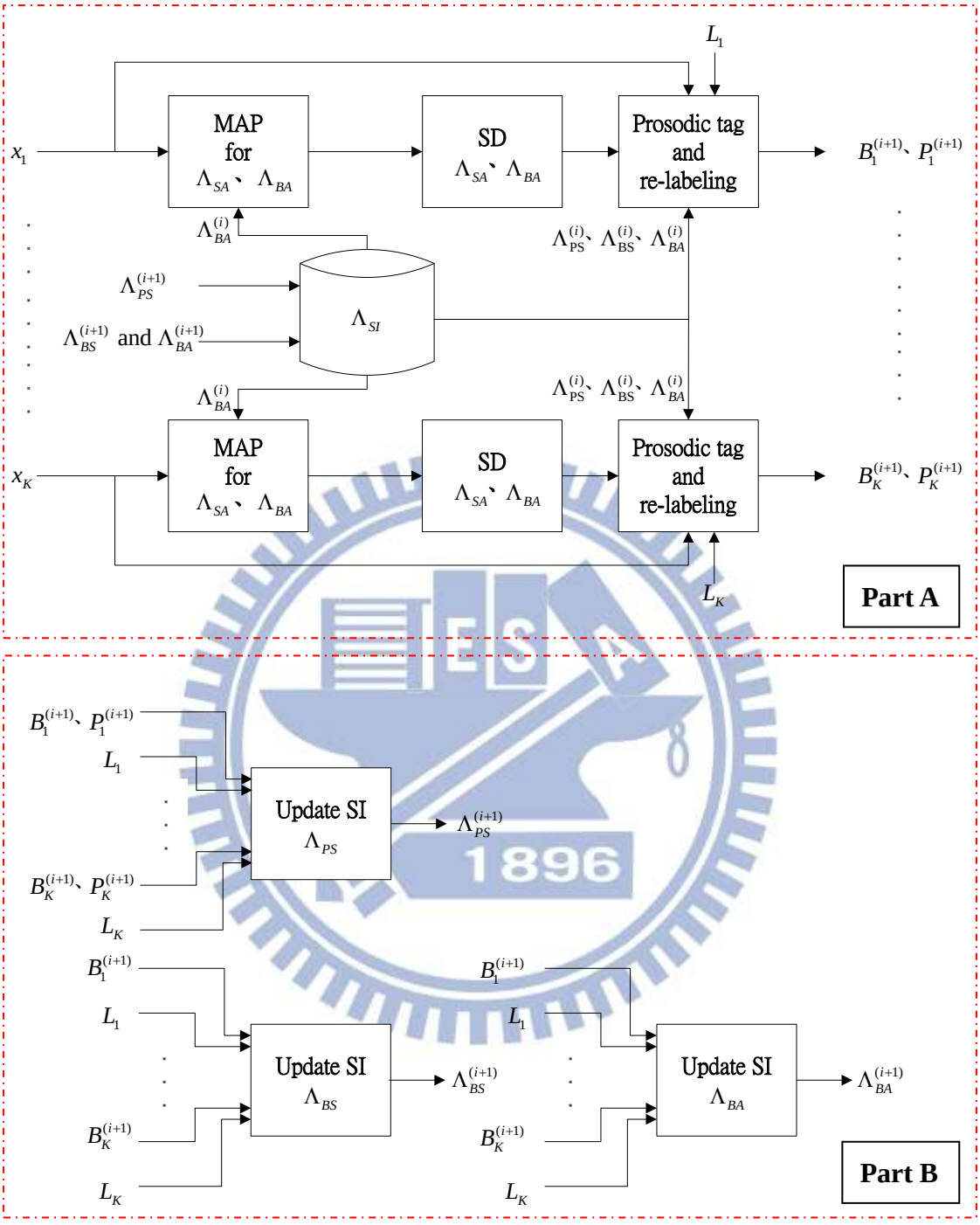


圖 2.8：本研究所提出之基於 MAP 之韻律模型調適流程圖

以下兩小節針對音節聲學模型與停頓聲學模型之 MAP 調適做更詳細的公式推導

2.4.3 音節聲學模型轉換

音節聲學模型 $P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})$ 可進一步分解成三個獨立子模型，分別用來模擬音節基頻軌跡、音節長度及音節能量位階，其數學式如下：

$$\begin{aligned} p(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L}) &\approx p(\mathbf{sp}|\mathbf{B},\mathbf{p},\mathbf{t})\, p(\mathbf{sd}|\mathbf{q},\mathbf{t},\mathbf{s},\mathbf{u})\, p(\mathbf{se}|\mathbf{r},\mathbf{t},\mathbf{f},\mathbf{u}) \\ &\approx \prod_{n=1}^{N} p(\mathbf{sp}_n|B_{n-1}^{n}, p_n, t_{n-1}^{n+1}) \prod_{n=1}^{N} p(sd_n|q_n,t_n,s_n,u_n) \prod_{n=1}^{N} p(se_n|r_n,t_n,f_n,u_n) \end{aligned} \tag{2-11}$$

我們針對三個子模型，使用 MAP 調適法則，將先前訓練好的語者獨立模型(Speaker Independent Model)當作先驗機率，調適出目標語者的模型 APs；APs 都調適完後，再更新共變異矩陣 $\mathbf{R}_{sp}$ 、R_{sd} 、R_{se} 。

2.4.3.1 音節基頻軌跡模型的 MAP 調適

音節基頻軌跡模型可表示成：

$$p(\mathbf{sp}_n \mid p_n, B_{n-1}^{n}, t_{t-1}^{t+1}) = N(\mathbf{sp}_n; \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\beta}^{f}_{B_{n-1},t_{n-1}^{n}} + \boldsymbol{\beta}^{b}_{B_n,t_n^{n+1}} + \boldsymbol{\mu}_{sp}, \mathbf{R}_{sp}) \tag{2-12}$$

其中 $p(\mathbf{sp}_n \mid p_n, B_{n-1}^{n}, t_{t-1}^{t+1})$ 用以模擬第 n 個音節基頻軌跡 $\mathbf{sp}_n$ ，在此假設所觀察到的 $\mathbf{sp}_n$ 受到的影響因素(Affecting Pattern, AP)為：目前的聲調 t_n 、目前的基頻韻律狀態 p_n 、以及在給定停頓標記 B_{n-1} 和 B_n 時，前後各一個音節聲調 t_{n-1} 和 t_n 所造成的連音影響，此處 $B_{n-1}^{n}=(B_{n-1},B_n)$ ，$t_{n-1}^{n+1}=(t_{n-1},t_n,t_{n+1})$ 。而 $\mathbf{sp}_n$ 是將音節基頻軌跡進行正交展開(orthogonal expansion)，投影到四個 Legendre 多項式基底所得到的四維正交參數[17]、tp_n 是 tone pair $t_n^{n+1}=(t_n,t_{n+1})$ ，$\boldsymbol{\beta}_{t_n}$ 及 $\boldsymbol{\beta}_{p_n}$ 則分別為目前音節音調 t_n 及目前音節韻律狀態 p_n 的 APs、$\boldsymbol{\beta}^{f}_{B_{n-1},t_{n-1}^{n}}$ 及 $\boldsymbol{\beta}^{b}_{B_n,t_n^{n+1}}$ 分別是第 n-1 個和第 n+1 個音節所貢獻的前後連音效應 APs、$\mathbf{sp}_n^r$ 為正規化後的 $\mathbf{sp}_n$ ，即 $\mathbf{sp}_n$ 扣除 $\boldsymbol{\beta}_{t_n}$ 、$\boldsymbol{\beta}_{p_n}$ 、$\boldsymbol{\beta}^{f}_{B_{n-1},t_{n-1}^{n}}$ 、$\boldsymbol{\beta}^{b}_{B_n,t_n^{n+1}}$ 和 $\boldsymbol{\mu}_{sp}$ 的殘餘值(residual)、$\mathbf{R}_{sp}$ 定義為 $\mathbf{sp}_n^r$ 的共變數矩陣(covariance matrix)。其調適步驟如下：

Step 1. 調適 $\boldsymbol{\mu}_{sp}$

在 MAP 調適法則下，可表示調適後的模型參數如下：

$$\hat{\boldsymbol{\mu}}_{sp} = (N_{sp}\mathbf{R}_{sp})(\mathbf{R}_{sp\_adap} + N_{sp}\mathbf{R}_{sp})^{-1}\boldsymbol{\mu}_{sp\_adap} + \mathbf{R}_{sp\_adap}(\mathbf{R}_{sp\_adap} + N_{sp}\mathbf{R}_{sp})^{-1}\boldsymbol{\mu}_{sp} \tag{2-13}$$

其中 N_{sp} 為調適語料的觀測資料數，$\boldsymbol{\mu}_{sp\_adap}$ 為調適語料的 $\boldsymbol{\mu}_{sp}$，其分佈為 $N(\boldsymbol{\mu}_{sp\_adap}, \mathbf{R}_{sp\_adap})$，事前機率則假設為 $N(\boldsymbol{\mu}_{sp}, \mathbf{R}_{sp})$ 之機率分佈。

Step 2. 調適 $\boldsymbol{\beta}_t$

中文語音可以分成 5 個 tone，故我們在調適 $\boldsymbol{\beta}_t$ 時，也將其分成 5 個 tone 分別調適，$\boldsymbol{\beta}_t, t = 1 \sim 5$，在 MAP 調適法則下，可表示調適後的模型參數如下：

$$\hat{\boldsymbol{\beta}}_t = (N_t\mathbf{R}_t)(\mathbf{R}_{t\_adap} + N_t\mathbf{R}_t)^{-1}\boldsymbol{\beta}_{t\_adap} + \mathbf{R}_{t\_adap}(\mathbf{R}_{t\_adap} + N_t\mathbf{R}_t)^{-1}\boldsymbol{\beta}_t \quad t = 1 \sim 5 \tag{2-14}$$

其中 N_t 為調適語料中屬於第 t 個 tone 的觀測資料數，$\boldsymbol{\beta}_{t\_adap}$ 為調適語料中屬於第 t 個 tone 的 $\boldsymbol{\beta}_t$，其分佈為 $N(\boldsymbol{\beta}_{t\_adap}, \mathbf{R}_{t\_adap})$，事前機率則假設為 $N(\boldsymbol{\beta}_t, \mathbf{R}_t)$ 之機率分佈。

Step 3. 調適 $\boldsymbol{\beta}_p$

本研究中將資料分成 16 個 prosodic state，故我們在調適 $\boldsymbol{\beta}_p$ 時，也將其分成 16 個 prosodic state 分別調適，$\boldsymbol{\beta}_p, p = 1 \sim 16$，在 MAP 調適法則下，可表示調適後的模型參數如下：

$$\hat{\boldsymbol{\beta}}_p = (N_p\mathbf{R}_p)(\mathbf{R}_{p\_adap} + N_p\mathbf{R}_p)^{-1}\boldsymbol{\beta}_{p\_adap} + \mathbf{R}_{p\_adap}(\mathbf{R}_{p\_adap} + N_p\mathbf{R}_p)^{-1}\boldsymbol{\beta}_p, p = 1 \sim 16 \tag{2-15}$$

其中 N_p 為調適語料中屬於第 p 個 prosodic state 的觀測資料數，$\boldsymbol{\beta}_{p\_adap}$ 為調適語料中屬於第 p 個 prosodic state 的 $\boldsymbol{\beta}_p$，其分佈為 $N(\boldsymbol{\beta}_{p\_adap}, \mathbf{R}_{p\_adap})$，事前機率則假設為 $N(\boldsymbol{\beta}_p, \mathbf{R}_p)$ 之機率分佈。

Step 4. 調適 $\boldsymbol{\beta}^f_{B_{n-1}, t^n_{n-1}}$、$\boldsymbol{\beta}^b_{B_n, t^{n+1}_n}$

由於$\boldsymbol{\beta}^{f}_{B_{n-1},t^{n}_{n-1}}$、$\boldsymbol{\beta}^{b}_{B_n,t^{n+1}_{n}}$分別是第$n-1$個和第$n+1$個音節所貢獻的前後連音效應 APs，考慮 5 個 tone 及 8 個 break label 的影響，在調適$\boldsymbol{\beta}^{f}_{B_{n-1},t^{n}_{n-1}}$、$\boldsymbol{\beta}^{b}_{B_n,t^{n+1}_{n}}$時，我們將其各分成 200 個類別分別調適，$\boldsymbol{\beta}^{f}_{B_{n-1},t^{n}_{n-1}},t=1\sim5,B=1\sim8$、$\boldsymbol{\beta}^{b}_{B_n,t^{n+1}_{n}},t=1\sim5,B=1\sim8$，在 MAP 調適法則下，可表示調適後的模型參數如：

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{f}_{B_{n-1},t^{n}_{n-1}} &= (N^{f}_{B_{n-1},t^{n}_{n-1}}\mathbf{R}^{f}_{B_{n-1},t^{n}_{n-1}})((\mathbf{R}^{f}_{B_{n-1},t^{n}_{n-1}})_{adap}+N^{f}_{B_{n-1},t^{n}_{n-1}}\mathbf{R}^{f}_{B_{n-1},t^{n}_{n-1}})^{-1}(\boldsymbol{\beta}^{f}_{B_{n-1},t^{n}_{n-1}})_{adap}\\ &\quad+(\mathbf{R}^{f}_{B_{n-1},t^{n}_{n-1}})_{adap}((\mathbf{R}^{f}_{B_{n-1},t^{n}_{n-1}})_{adap}+N^{f}_{B_{n-1},t^{n}_{n-1}}\mathbf{R}^{f}_{B_{n-1},t^{n}_{n-1}})^{-1}\boldsymbol{\beta}^{f}_{B_{n-1},t^{n}_{n-1}}\\ &\quad t=1\sim5,B=1\sim8\end{aligned} \tag{2-16}$$

其中$N^{f}_{B_{n-1},t^{n}_{n-1}}$為調適語料中第$n-1$個音節屬於第$t-1$個 tone、第$B-1$個 break label，第$n$個音節屬於第$t$個 tone 的觀測資料數，$(\boldsymbol{\beta}^{f}_{B_{n-1},t^{n}_{n-1}})_{adap}$為調適語料中第$n-1$個音節屬於第$t-1$個 tone、第$n$個音節屬於第$t$個 tone、第$B$個 break label 的$\boldsymbol{\beta}^{f}_{B_{n-1},t^{n}_{n-1}}$，其分佈為$N((\boldsymbol{\beta}^{f}_{B_{n-1},t^{n}_{n-1}})_{adap},(\mathbf{R}^{f}_{B_{n-1},t^{n}_{n-1}})_{adap})$，事前機率則假設為$N(\boldsymbol{\beta}^{f}_{B_{n-1},t^{n}_{n-1}},\mathbf{R}^{f}_{B_{n-1},t^{n}_{n-1}})$之機率分佈。

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{b}_{B_n,t^{n+1}_{n}} &= (N^{b}_{B_n,t^{n+1}_{n}}\mathbf{R}^{b}_{B_n,t^{n+1}_{n}}))((\mathbf{R}^{b}_{B_n,t^{n+1}_{n}})_{adap}+N^{b}_{B_n,t^{n+1}_{n}}\mathbf{R}^{b}_{B_n,t^{n+1}_{n}})^{-1}(\boldsymbol{\beta}^{b}_{B_n,t^{n+1}_{n}})_{adap}\\ &\quad+(\mathbf{R}^{b}_{B_n,t^{n+1}_{n}})_{adap}((\mathbf{R}^{b}_{B_n,t^{n+1}_{n}})_{adap}+N^{b}_{B_n,t^{n+1}_{n}}\mathbf{R}^{b}_{B_n,t^{n+1}_{n}})^{-1}\boldsymbol{\beta}^{b}_{B_n,t^{n+1}_{n}}\\ &\quad t=1\sim5,B=1\sim8\end{aligned} \tag{2-17}$$

其中$N^{b}_{B_n,t^{n+1}_{n}}$為調適語料中第n個音節屬於第t個 tone、第B個 break label，第$n+1$個音節屬於第$t+1$個 tone 的觀測資料數，$(\boldsymbol{\beta}^{b}_{B_n,t^{n+1}_{n}})_{adap}$為調適語料中第$n$個音節屬於第$t$個 tone、第$B$個 break label，第$n+1$個音節屬於第$t+1$個 tone 的$\boldsymbol{\beta}^{b}_{B_n,t^{n+1}_{n}}$，其分佈為$N((\boldsymbol{\beta}^{b}_{B_n,t^{n+1}_{n}})_{adap},(\mathbf{R}^{b}_{B_n,t^{n+1}_{n}})^{s}_{adap})$，事前機率則假設為$N(\boldsymbol{\beta}^{b}_{B_n,t^{n+1}_{n}},\mathbf{R}^{b}_{B_n,t^{n+1}_{n}})$之機率分佈。

2.4.3.2 音節長度模型的 MAP 調適

音節長度模型可表示成：

$$p(sd_n \mid q_n,s_n,t_n,u_n)=N(sd_n;\gamma_{t_n}+\gamma_{s_n}+\gamma_{q_n}+\gamma_{u_n}+\mu_{sd},R_{sd}) \tag{2-18}$$

其中 γ_{t_n}、γ_{s_n}、γ_{q_n} 和 γ_{u_n} 分別為聲調、基本音節類型、韻律狀態和句子對 sd_n 的 APs，μ_{sd} 和 R_{sd} 分別為 sd_n 總體平均及其殘餘值之變異數其調適步驟如下：

Step 1. γ_u 不調

Step 2. 調適 μ_{sd}

在 MAP 調適法則下，可表示調適後的模型參數如：

$$\hat{\mu}_{sd} = \frac{N_{sd}(\tau_{sd})^2}{(\sigma_{sd})^2 + N_{sd}(\tau_{sd})^2}\mu_{sd\_adap} + \frac{(\sigma_{sd})^2}{(\sigma_{sd})^2 + N_{sd}(\tau_{sd})^2}\mu_{sd} \tag{2-19}$$

其中 N_{sd} 為調適語料的觀測資料數，μ_{sd_adap} 為調適語料的 μ_{sd}，其分佈為 $N(\mu_{sd\_adap},(\sigma_{sd})^2)$，事前機率則假設為 $N(\mu_{sd},(\tau_{sd})^2)$ 之機率分佈。

Step 3. 調適 γ_t

中文語音可以分成5個tone，故我們在調適 γ_t 時，也將其分成5個tone分別調適，$\gamma_t, t=1\sim5$，在 MAP 調適法則下，可表示調適後的模型參數如下：

$$\hat{\gamma}_t = \frac{N_t(\tau_t)^2}{(\sigma_t)^2 + N_t(\tau_t)^2}\gamma_{t\_adap} + \frac{(\sigma_t)^2}{(\sigma_t)^2 + N_t(\tau_t)^2}\gamma_t, t=1\sim5 \tag{2-20}$$

其中 N_t 為調適語料中屬於第 t 個 tone 的觀測資料數，γ_{t_adap} 為調適語料中屬於第 t 個 tone 的 γ_t，其分佈為 $N(\gamma_{t\_adap},(\sigma_t)^2)$，事前機率則假設為 $N(\gamma_t,(\tau_t)^2)$ 之機率分佈。

Step 4. 調適 γ_s

本研究中將中文語音的 411 音節合併為 82 類 base-syllable type，故我們在調適 γ_s 時，也將其分成 82 類 base-syllable type 分別調適，$\gamma_s, s=1\sim82$，在 MAP 調適法則下，可表示調適後的模型參數如：

$$\hat{\gamma}_s = \frac{N_s(\tau_s)^2}{(\sigma_s)^2 + N_s(\tau_s)^2}\gamma_{s\_adap} + \frac{(\sigma_s)^2}{(\sigma_s)^2 + N_s(\tau_s)^2}\gamma_s, s=1\sim82 \tag{2-21}$$

其中 N_s 為調適語料中屬於第 s 類 base-syllable type 的觀測資料數，γ_{s_adap} 為調適語料中屬於第 s 類 base-syllable type 的 γ_s ，其分佈為 $N(\gamma_{s\_adap},(\sigma_s)^2)$，事前機率則假設為 $N(\gamma_s,(\tau_s)^2)$ 之機率分佈。

Step 5. 調適 γ_q

本研究中將資料分成 16 個 prosodic state，故我們在調適 γ_q 時，也將其分成 16 個 prosodic state 分別調適，$\gamma_q, q=1\sim16$，在 MAP 調適法則下，可表示調適後的模型參數如：

$$\hat{\gamma}_q=\frac{N_q(\tau_q)^2}{(\sigma_q)^2+N_q(\tau_q)^2}\gamma_{q\_adap}+\frac{(\sigma_q)^2}{(\sigma_q)^2+N_q(\tau_q)^2}\gamma_q, q=1\sim16 \tag{2-22}$$

其中 N_q 為調適語料中屬於第 q 個 prosodic state 的觀測資料數，γ_{q_adap} 為調適語料中屬於第 q 個 prosodic state 的 γ_q ，其分佈為 $N(\gamma_{q\_adap},(\sigma_q)^2)$，事前機率則假設為 $N(\gamma_q,(\tau_q)^2)$ 之機率分佈。

2.4.3.3 音節能量位階的 MAP 調適

我們知道音節能量位階模型可表示成：

$$p(se_n\mid r_n,f_n,t_n,u_n)=N(se_n;\alpha_{t_n}+\alpha_{f_n}+\alpha_{r_n}+\alpha_{u_n}+\mu_{se},R_{se}) \tag{2-23}$$

其中 α_{t_n} 、α_{f_n} 、α_{r_n} 和 α_{u_n} 分別為聲調、聲母類型、韻律狀態和句子對 se_n 的 APs，μ_{se} 和 R_{se} 則分別為 se_n 總體平均及其殘餘值之變異數，其調適步驟如下：

Step 1. α_u 不調

Step 2. 調適 μ_{se}

在 MAP 調適法則下，可表示調適後的模型參數如：

$$\hat{\mu}_{se}=\frac{N_{se}(\tau_{se})^2}{(\sigma_{se})^2+N_{se}(\tau_{se})^2}\mu_{se\_adap}+\frac{(\sigma_{se})^2}{(\sigma_{se})^2+N_{se}(\tau_{se})^2}\mu_{se} \tag{2-24}$$

其中 N_{se} 為調適語料的觀測資料數，μ_{se_adap} 為調適語料的 μ_{se} ，其分佈為 $N(\mu_{se\_adap},(\sigma_{se})^2)$，事前機率則假設為 $N(\mu_{se},(\tau_{se})^2)$ 之機率分佈。

Step 3. 調適α_t

中文語音可以分成 5 個 tone，故我們在調適α_t時，也將其分成 5 個 tone 分別調適，$\alpha_t, t=1\sim 5$，在 MAP 調適法則下，可表示調適後的模型參數如：

$$\hat{\alpha}_t = \frac{N_t(\tau_t)^2}{(\sigma_t)^2 + N_t(\tau_t)^2}\alpha_{t\_adap} + \frac{(\sigma_t)^2}{(\sigma_t)^2 + N_t(\tau_t)^2}\alpha_t, t=1\sim 5 \quad (2\text{-}25)$$

其中N_t為調適語料中屬於第t個tone的觀測資料數，α_{t_adap}為調適語料中屬於第t個tone的α_t，其分佈為$N(\alpha_{t_adap},(\sigma_t)^2)$，事前機率則假設為$N(\alpha_t,(\tau_t)^2)$之機率分佈。

Step 4. 調適α_f

本研究中將中文語音的 411 音節合併為 40 類 final type，故我們在調適α_f時，也將其分成 40 類 final type 分別調適，$\alpha_f, f=1\sim 40$，在 MAP 調適法則下，可表示調適後的模型參數如：

$$\hat{\alpha}_f = \frac{N_f(\tau_f)^2}{(\sigma_f)^2 + N_f(\tau_f)^2}\alpha_{f\_adap} + \frac{(\sigma_f)^2}{(\sigma_f)^2 + N_f(\tau_f)^2}\alpha_f, f=1\sim 40 \quad (2\text{-}26)$$

其中N_f為調適語料中屬於第f類 final type 的觀測資料數，α_{f_adap}為調適語料中屬於第f類 final type 的α_f，其分佈為$N(\alpha_{f_adap},(\sigma_f)^2)$，事前機率則假設為$N(\alpha_f,(\tau_f)^2)$之機率分佈。

Step 5. 調適α_r

本研究中將資料分成 16 個 prosodic state，故我們在調適α_r時，也將其分成 16 個 prosodic state 分別調適，$\alpha_r, r=1\sim 16$，在 MAP 調適法則下，可表示調適後的模型參數如：

$$\hat{\alpha}_r = \frac{N_r(\tau_r)^2}{(\sigma_r)^2 + N_r(\tau_r)^2}\alpha_{r\_adap} + \frac{(\sigma_r)^2}{(\sigma_r)^2 + N_r(\tau_r)^2}\alpha_r, r=1\sim 16 \quad (2\text{-}27)$$

其中N_r為調適語料中屬於第r個 prosodic state 的觀測資料數，α_{r_adap}為調適語料中屬於第r個 prosodic state 的α_r，其分佈為$N(\alpha_{r_adap},(\sigma_r)^2)$，事前機率則假設為$N(\alpha_r,(\tau_r)^2)$之機率分佈。

2.4.4 停頓聲學模型轉換

將停頓聲學模型 $P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$ 做進一步分解

$$P(\mathbf{Y},\mathbf{Z}\,|\,\mathbf{B},\mathbf{L}) \approx P(\mathbf{Y},\mathbf{Z}\,|\,\mathbf{B},\mathbf{l}) \approx \prod_{n=1}^{N} p(pd_n, ed_n, pj_n, dl_n, df_n \,|\, B_n, \mathbf{l}_n) \tag{2-28}$$

我們維持 SI 決策樹的架構不變，僅使用 MAP 調適法則調整每個 leaf node 的 5 個子模型。本研究使用 PLM 訓練 SI model，其 leaf node 共 40 個，故我們在調適停頓長度(pd)、能量下降程度(ed)、正規化基頻差(pj)、正規化長度拉長因子(dl 、 df)時，也將其各分成 49 個 leaf node 分別調適，在此我們只調適各參數的平均值，$\mu_{pd}(i), i=1\sim 49$、$\mu_{ed}(i), i=1\sim 49$、$\mu_{pj}(i), i=1\sim 49$ 、$\mu_{dl}(i), i=1\sim 49$ 、 $\mu_{df}(i), i=1\sim 49$ ，在 MAP 調適法則下，可表示調適後的模型參數如下：

$$\hat{\mu}_{pd}(i) = \frac{N_{pd}^{i}(\tau_{pd}(i))^2}{(\sigma_{pd}(i))^2 + N_{pd}^{i}(\tau_{pd}(i))^2}\mu_{pd\_adap}(i) + \frac{(\sigma_{pd}(i))^2}{(\sigma_{pd}(i))^2 + N_{pd}^{i}(\tau_{pd}(i))^2}\mu_{pd}, i=1\sim 49 \tag{2-29}$$

$$\hat{\mu}_{ed}(i) = \frac{N_{ed}^{i}(\tau_{ed}(i))^2}{(\sigma_{ed}(i))^2 + N_{ed}^{i}(\tau_{ed}(i))^2}\mu_{ed\_adap}(i) + \frac{(\sigma_{ed}(i))^2}{(\sigma_{ed}(i))^2 + N_{ed}^{i}(\tau_{ed}(i))^2}\mu_{ed}, i=1\sim 49 \tag{2-30}$$

$$\hat{\mu}_{pj}(i) = \frac{N_{pj}^{i}(\tau_{pj}(i))^2}{(\sigma_{pj}(i))^2 + N_{pj}^{i}(\tau_{pj}(i))^2}\mu_{pj\_adap}(i) + \frac{(\sigma_{pj}(i))^2}{(\sigma_{pj}(i))^2 + N_{pj}^{i}(\tau_{pj}(i))^2}\mu_{pj}, i=1\sim 49 \tag{2-31}$$

$$\hat{\mu}_{dl}(i) = \frac{N_{dl}^{i}(\tau_{dl}(i))^2}{(\sigma_{dl}(i))^2 + N_{dl}^{i}(\tau_{dl}(i))^2}\mu_{dl\_adap}(i) + \frac{(\sigma_{dl}(i))^2}{(\sigma_{dl}(i))^2 + N_{dl}^{i}(\tau_{dl}(i))^2}\mu_{dl}, i=1\sim 49 \tag{2-32}$$

$$\hat{\mu}_{df}(i)=\frac{N_{df}^{i}(\tau_{df}(i))^{2}}{(\sigma_{df}(i))^{2}+N_{df}^{i}(\tau_{df}(i))^{2}}\mu_{df\_adap}(i)+\frac{(\sigma_{df}(i))^{2}}{(\sigma_{df}(i))^{2}+N_{df}^{i}(\tau_{df}(i))^{2}}\mu_{df},i=1\sim 49 \quad (2\text{-}33)$$

其中 N_{pd}^{i} 、N_{ed}^{i} 、N_{pj}^{i} 、N_{dl}^{i} 、N_{df}^{i} 為調適語料中屬於第 i 個 leaf node 的 *pd* 、*ed* 、*pj* 、*dl* 、*df* 觀測資料數，$\mu_{pd\_adap}(i)$ 、$\mu_{ed\_adap}(i)$ 、$\mu_{pj\_adap}(i)$ 、$\mu_{dl\_adap}(i)$ 、$\mu_{df\_adap}(i)$ 為調適語料中屬於第 i 個 leaf node 的參數 *pd* 、*ed* 、*pj* 、*dl* 、*df* 的平均值，其分佈分別為 $N(\mu_{pd\_adap}(i),(\sigma_{pd}(i))^{2})$ 、$N(\mu_{ed\_adap}(i),(\sigma_{ed}(i))^{2})$ 、$N(\mu_{pj\_adap}(i),(\sigma_{pj}(i))^{2})$ 、$N(\mu_{dl\_adap}(i),(\sigma_{dl}(i))^{2})$ 、$N(\mu_{df\_adap}(i),(\sigma_{df}(i))^{2})$ ，事前機率則假設分別為 $N(\mu_{pd}(i),(\tau_{pd}(i))^{2})$ 、$N(\mu_{ed}(i),(\tau_{ed}(i))^{2})$ 、$N(\mu_{pj}(i),(\tau_{pj}(i))^{2})$ 、$N(\mu_{dl}(i),(\tau_{dl}(i))^{2})$ 、$N(\mu_{df}(i),(\tau_{df}(i))^{2})$ 之機率分佈。

第三章　漢語語者韻律轉換

本論文將會對音節之基頻軌跡、長度及能量三種韻律參數做語者韻律轉換。在本章中分別介紹高斯正規化轉換方法及本論文所提出利用多語者漢語韻律模型及 MAP 調適法則的轉換方法。

3.1 高斯正規化轉換方法

高斯正規化(Gaussian Normalization)的方式，其意義為對平均值與變異數做一線性轉換，此方法亦稱為平均值/變異數轉換(Mean/Variance Transformation)。此方法的優點為簡單實作，且訓練語料可以是非平行語料，常做為韻律轉換的基本方法與比較的對象。令 $\mathbf{x}_n$ 與 $\mathbf{y}_n$ 分別表示來源(Source)語者與目標(Target)語者在第 n 個音節的韻律參數；接著假設來源與目標語者每個音節的韻律參數分別服從高斯分佈如下：

$$P(\mathbf{x}_n) = N(\mathbf{x}_n; \boldsymbol{\mu}_x, \Sigma_{xx}) \quad \text{and} \quad P(\mathbf{y}_n) = N(\mathbf{y}_n; \boldsymbol{\mu}_y, \Sigma_{yy}) \tag{3-1}$$

其中，$\boldsymbol{\mu}_x$ 與 $\boldsymbol{\mu}_y$ 分別為來源和目標語者的期望值向量；Σ_{xx} 與 Σ_{yy} 分別為來源與目標語者的共變異數矩陣，此共變異數矩陣通常假設為對角化矩陣。因此，以高斯正規化的方式對 $\mathbf{x}_n$ 轉換，轉換函式如下：

$$\hat{\mathbf{y}}_n = (\Sigma_{yy})^{\frac{1}{2}} (\Sigma_{xx})^{\frac{-1}{2}} (\mathbf{x}_n - \boldsymbol{\mu}_x) + \boldsymbol{\mu}_y \tag{3-2}$$

其中，$\hat{\mathbf{y}}_n$ 為轉換後第 n 個音節的韻律參數。

3.2 MAP 調適法則轉換方法

圖 3.1 為本研究所提出的 MAP 調適法則韻律轉換方法流程圖，其中 Λ_{SD} 為語者相關的韻律模型；Λ_{SA} 為音節聲學模型；Λ_{BA} 為停頓聲學模型；B_s、P_s 為 Source 語者的停頓標記和韻律狀態標記。

首先，我們對 Source 語者的測試語料抽取韻律參數，並進行韻律參數正規化，再利用 MAP 調適出的 Source 語者的韻律模型對其做停頓及韻律狀態的標記；其次，我們利用這些停頓及韻律狀態標記和 MAP 調適出的 Target 語者的音節聲學模型及停頓聲學模型中的各聲學參數重新合成音節基頻軌跡、音節長度和音節能量後，將其做反正規化；最後，我們將 Target 語者的測試語料也抽取韻律參數，搭配 Source 語者的韻律參數，去計算我們轉換出來的音節基頻軌跡、音節長度和音節能量的 NMSE。

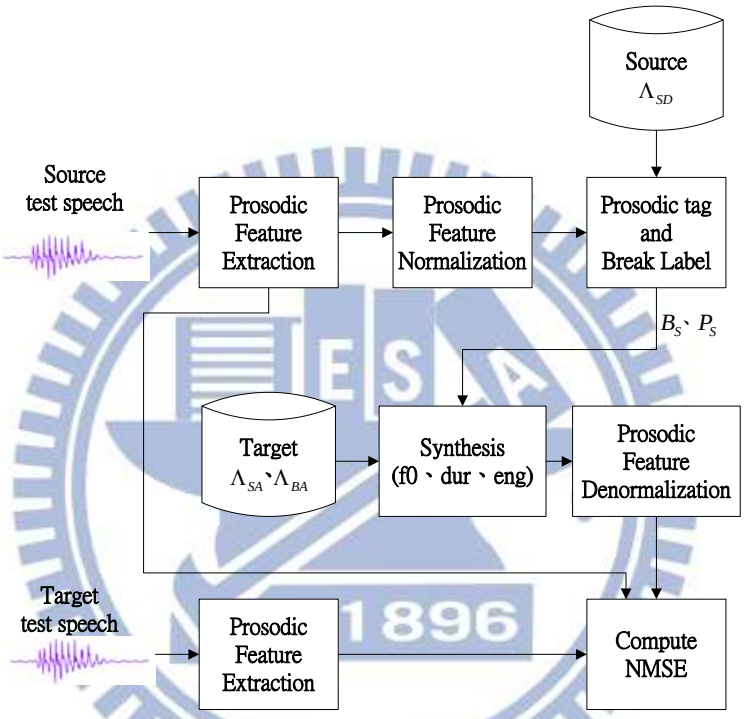


圖 3.1：MAP 調適法則韻律轉換方法流程圖

第四章 實驗結果與分析

本實驗所使用的語料庫為國立交通大學語音處理實驗室自行錄製的部分韻律平行語料，其中包含 9 男 6 女(男生包含：Arron、daniel、Jimmy、kook、Merry、Paul、pulu、sung、tu，女生包含：byetwo、normal、ppp、rebecca、shanli、ysu)共 15 位語者。音檔皆為取樣率 20kHz 錄製並降頻至 16kHz 之 pcm 格式，其解析度為 16bits 的單聲道音檔；切割資訊是由 Hidden Markov Model Tool Kit(HTK)切割並經由人工手動校正，共 1219 句，總音節數為 157887 個，每位語者的語料統計資訊如表 4.1、表 4.2 所示：

表 4.1：15 位語者訓練語料使用的音檔數及音節數

Speaker	**Arron**	**byetwo**	**daniel**	**Jimmy**	**kook**	**Merry**	**normal**	**Paul**
音檔數	95	82	95	95	95	95	95	95
音節數	13220	11314	13220	10576	11715	10672	12870	11090
Speaker	**ppp**	**pulu**	**rebecca**	**shanli**	**sung**	**tu**	**ysu**	**加總**
音檔數	95	46	77	83	46	43	82	**1219**
音節數	12413	5907	10691	11469	5910	5527	11293	**157887**

表 4.2：15 位語者測試語料使用的音檔數及音節數

Speaker	**Arron**	**byetwo**	**daniel**	**Jimmy**	**kook**	**Merry**	**normal**	**Paul**
音檔數	5	5	5	5	5	5	5	5
音節數	649	649	649	649	649	649	649	649
Speaker	**ppp**	**pulu**	**rebecca**	**shanli**	**sung**	**tu**	**ysu**	**加總**
音檔數	5	5	5	5	5	5	5	**75**
音節數	649	649	649	649	649	649	649	**9735**

依據圖 2.8 的系統架構及 2-10 式，我們分別觀察不同階段的個別語者目標總概似度(total likelihood of objective function)及其總和如表 4.3 所示；其中 **Original** 代表所有參數皆未經過 MAP 調適直接計算的目標總概似度；**MAP** 代表音節聲學模型中 APs 和停頓聲學模型中各參數經過 MAP 調適，但還未更新停頓及韻律狀態標記的目標總概似度；**B_PS_Updated** 代表經過圖 2.8 Part A 部分後的目標總概似度；**Iteration 1** 代表經過圖 2.8 Part A + Part B 兩部份一次遞迴後的目標總概似度；**Iteration 2** 代表經過圖 2.8 Part A + Part B 兩部份兩次遞迴後的目標總概似度。

圖 4.1 顯示所有語者各階段的目標總概似度總和，可以發現經過 MAP 調適參數及更新其它子模型後，由於韻律狀態模型、停頓語法模型及停頓聲學模型的樹狀結構是所有語者共用的，故對於個別語者其目標總概似度有可能增加或減少，但對於所有語者的目標總概似度總和會呈現增加的現象，並且逐漸接近收斂；此結果符合我們原先的預期。

表 4.3：每位語者在不同階段的個別目標總概似度及其總和

Speaker / Total_Likelihood	Arron	byetwo	daniel	Jimmy	kook	Merry	normal	Paul
Original	116340	130210	149540	95040	108740	106420	126800	100680
MAP	127870	130160	149480	96164	109480	106330	130500	100350
B_PS_Updated	128080	131180	150310	96401	109650	106780	131590	100500
Iteration 1	128610	131390	150730	96435	109810	106880	131590	100720
Iteration 2	128670	131420	150790	96486	109820	106880	131590	100750
Speaker / Total_Likelihood	**ppp**	**pulu**	**rebecca**	**shanli**	**sung**	**tu**	**ysu**	**加總**
Original	128540	69350	117330	110730	57963	56558	98909	**1573150**
MAP	130050	70982	115320	110930	57644	59511	99927	**1594698**
B_PS_Updated	131470	71258	116230	111290	58206	59669	101020	**1603634**
Iteration 1	131690	71313	116210	111300	58360	59836	101020	**1605894**
Iteration 2	131750	71317	116220	111310	58381	59857	101040	**1606281**

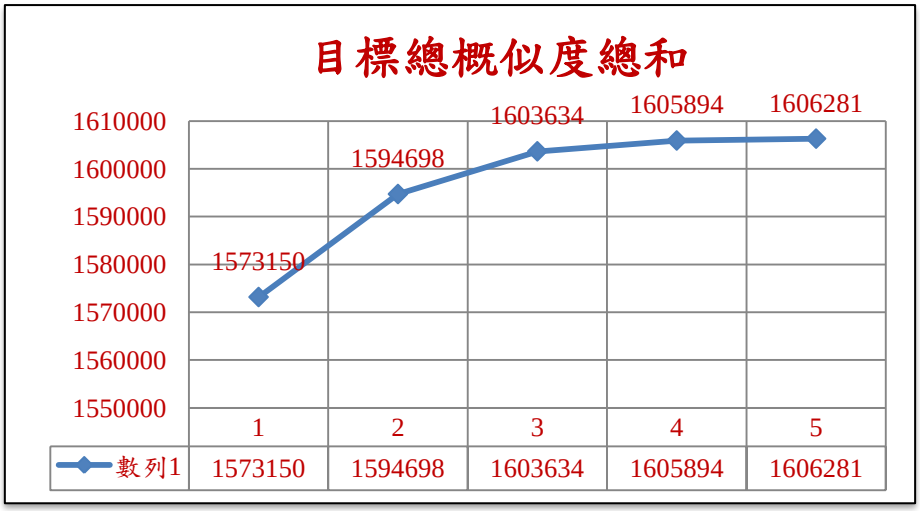


圖 4.1：所有語者不同階段的目標總概似度總和

4.1 韻律標記結果之分析

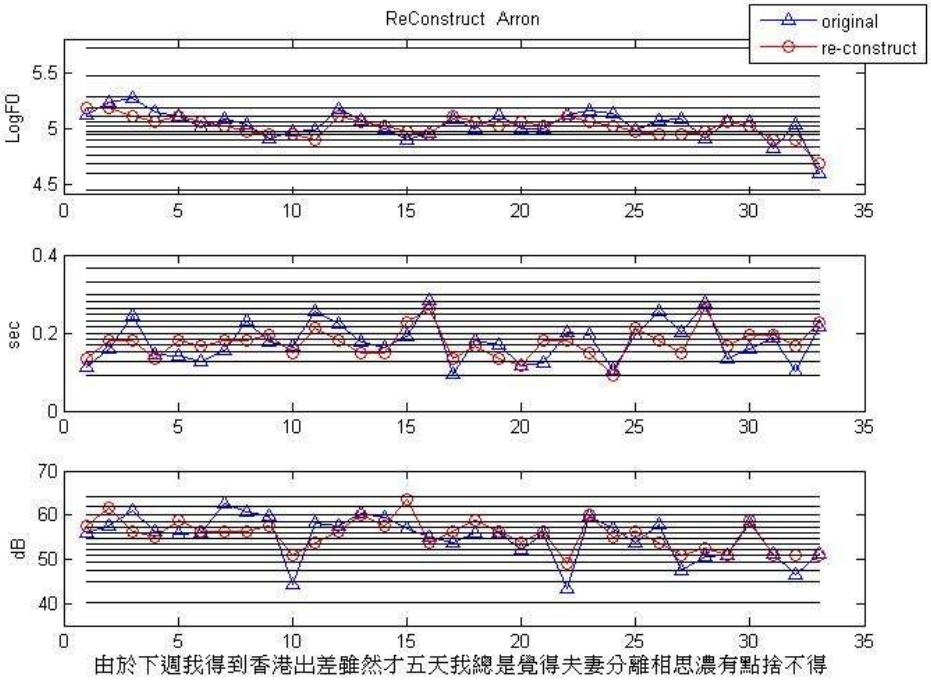
圖 4.2 代表各語者個別的韻律狀態標記；由每張圖中可以觀察到(mean + prosodic state)有相對於 original 較為平滑的曲線，我們可以將其視為韻律上的趨勢變化；而圖中的水平線，代表 mean 加 16 個 prosodic state 後所包含的音高、音節長度及音節能量範圍。

首先，我們從圖 4.2(a)~(o)上方的圖(Logf0)，可以觀察到不同語者在念同一個句子時，其結尾的音高變化可以分為上升及下降兩種，其中上升的有 Jimmy、Merry、ysu，其餘的是下降；由此我們可以看出不同語者在句子結尾時，會有自己特有的音高變化，可以視為個別語者的說話特性。

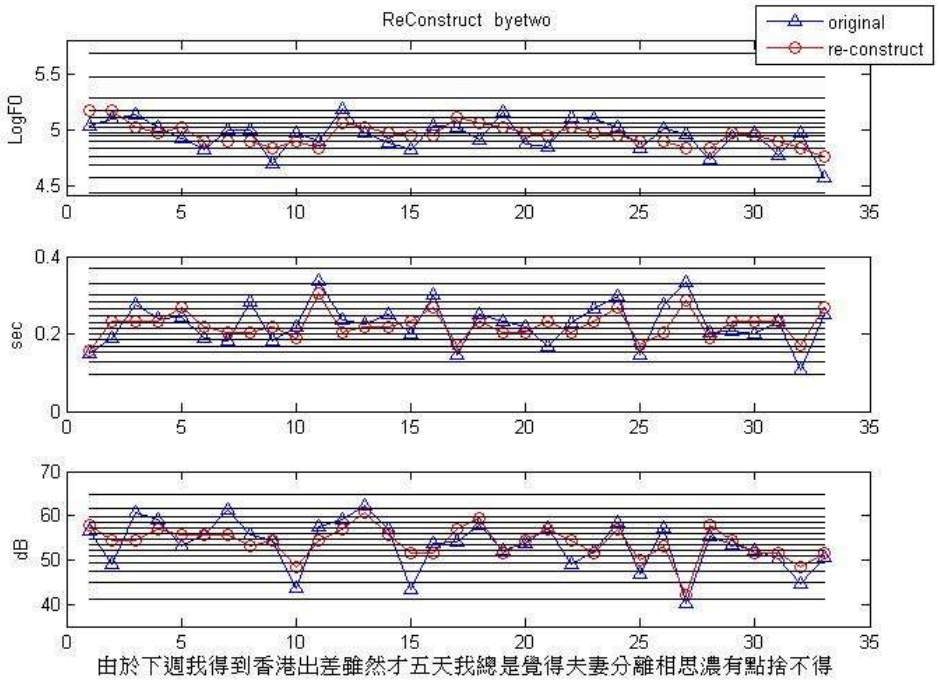
另外，我們從(a)~(o)中間的圖(duration)搭配上方的圖(Logf0)，可以發現在音節長度長的地方，會伴隨著基頻重置的現象；而單看中間的圖，可以發現在句子結尾的時候，通常也會有較長的音節長度，此為 final lengthening 的現象(其中(e)、(h)句子結尾時，長度反而下降，重新檢查原始音檔後，發現其原因為(e)的最後一個音念的不完全就結束，而(h)是在句子結束後，錄到了物品敲擊桌子的聲音)。

最後，我們從(a)~(o)下方的圖(energy)，可以觀察到不同語者說話時的能量變化範圍大致可分為兩種，一種是能量範圍變化較劇烈的語者，例如： Jimmy、Merry，而另一種是能量變化較為平緩的語者，例如：byetwo、daniel，由此可看出念同一句話時，有的人語氣比較有輕

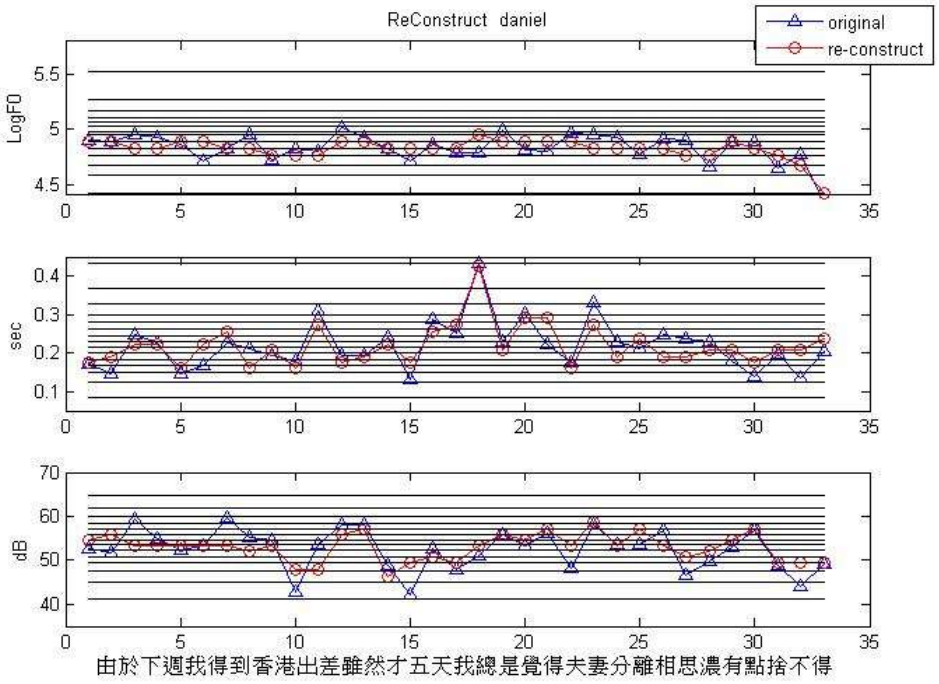
重強弱的變化，有的人則較為平淡，可將其視為個別語者的說話特性；其次，我們發現大部分語者的第 10 個音節(出)和第 15 個音節(五)都有能量突然下降的趨勢，並且我們的(mean + prosodic state)並不能很好的描述出這樣的趨勢，經過觀察發現其原因為第 10 個音節(出)和第 15 個音節(五)屬於”wu”類的韻母，此類韻母音節能量位階最小，所以才導致我們僅利用(mean + prosodic state)的韻律趨勢無法很好的描述這類型的韻母，此韻母類型對應到 411 音節類型如”su”、 ”tu”等。



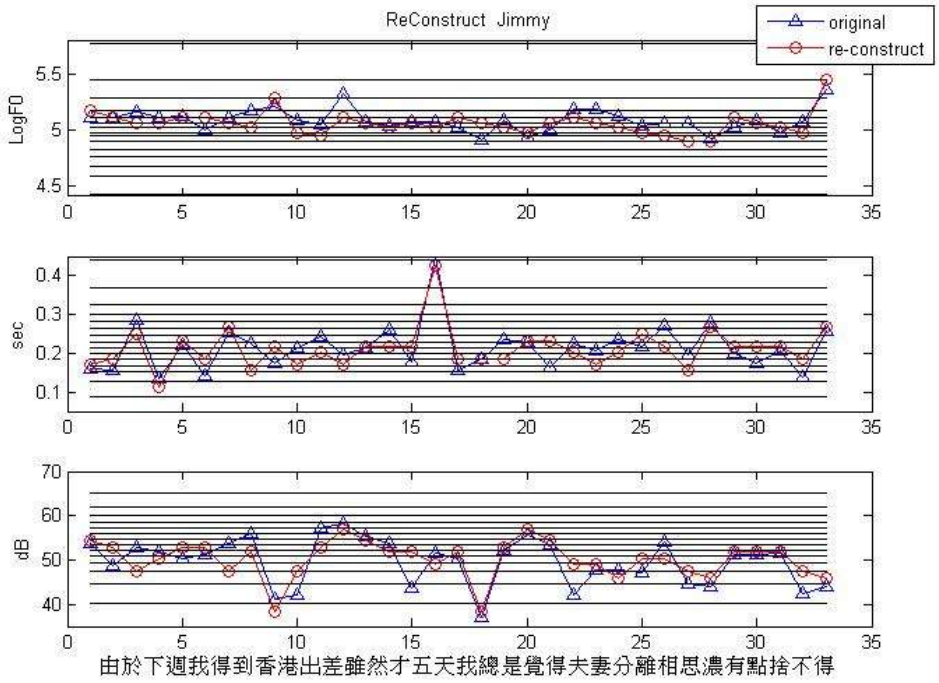
(a) Arron 的韻律狀態標記



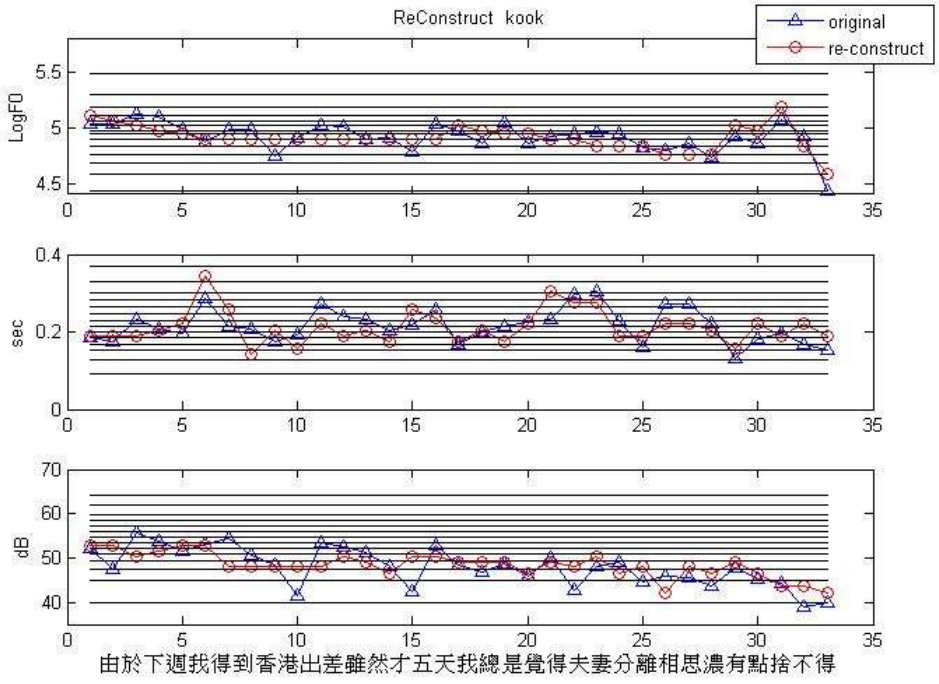
(b) byetwo 的韻律狀態標記



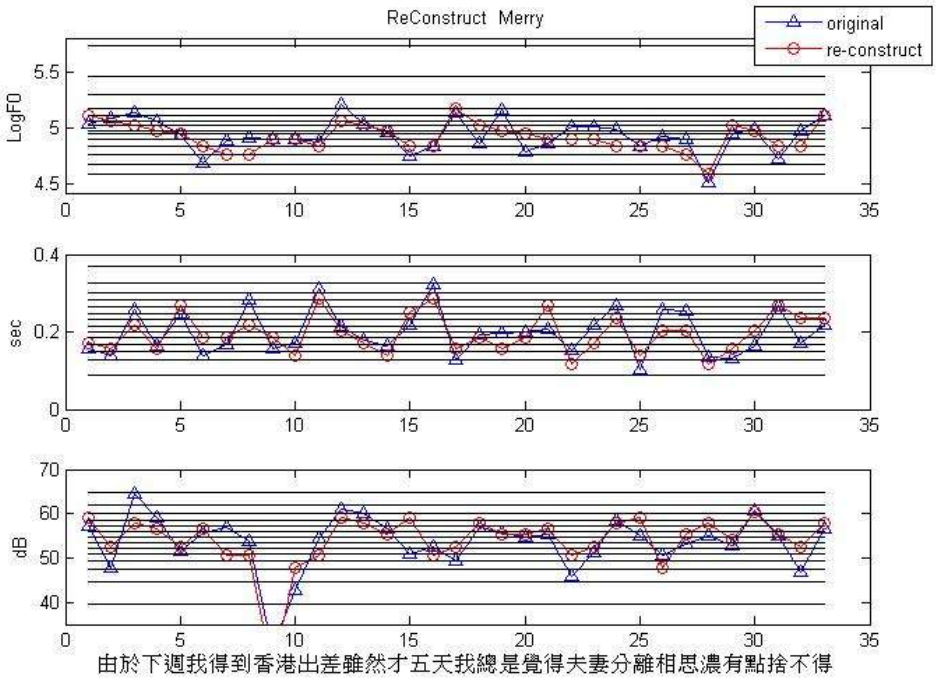
(c) daniel 的韻律狀態標記



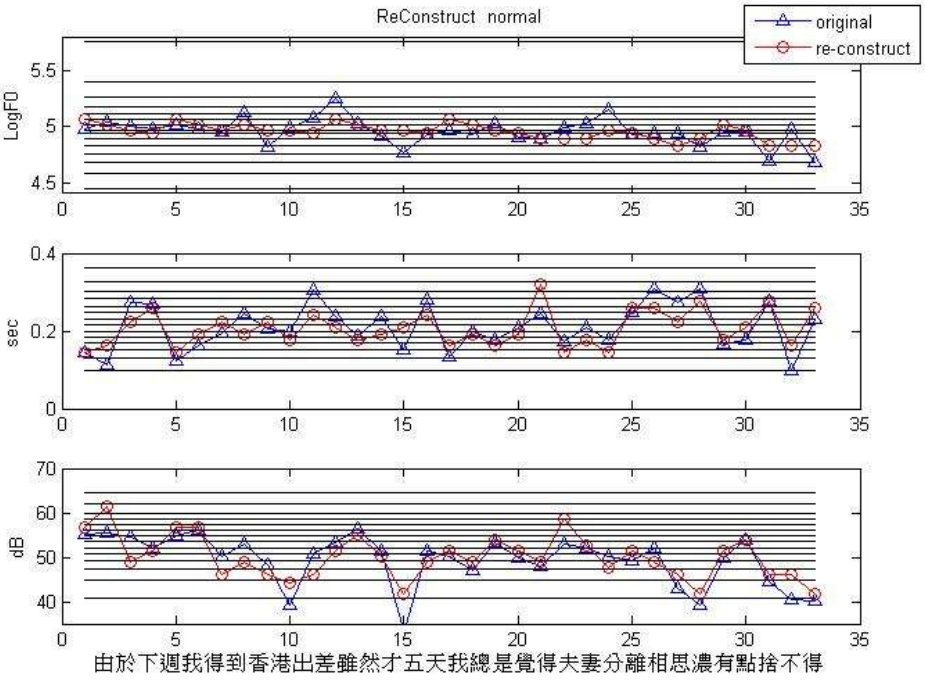
(d) Jimmy 的韻律狀態標記



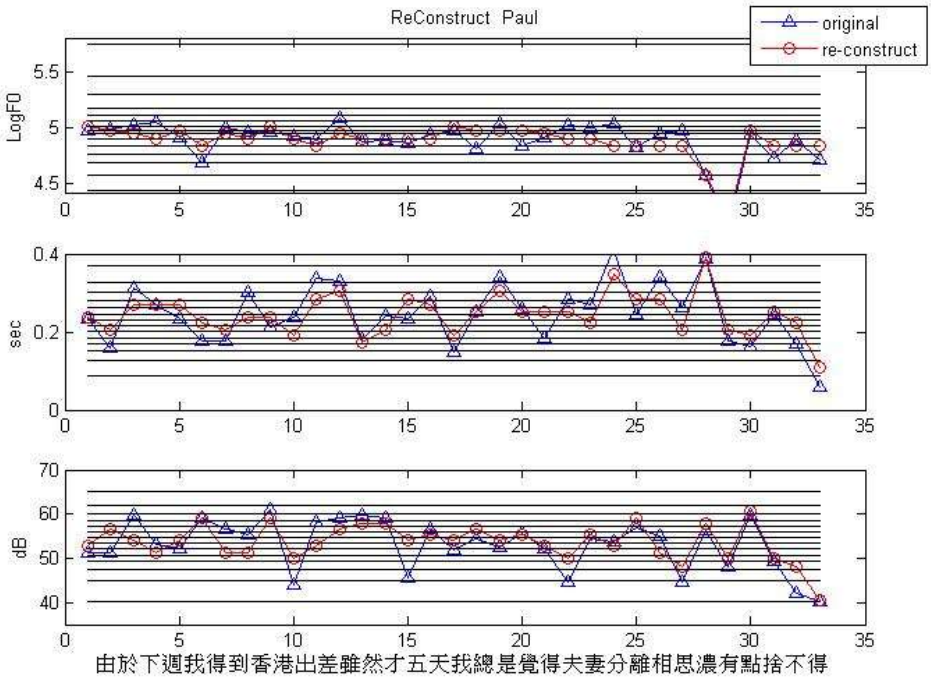
(e) kook 的韻律狀態標記



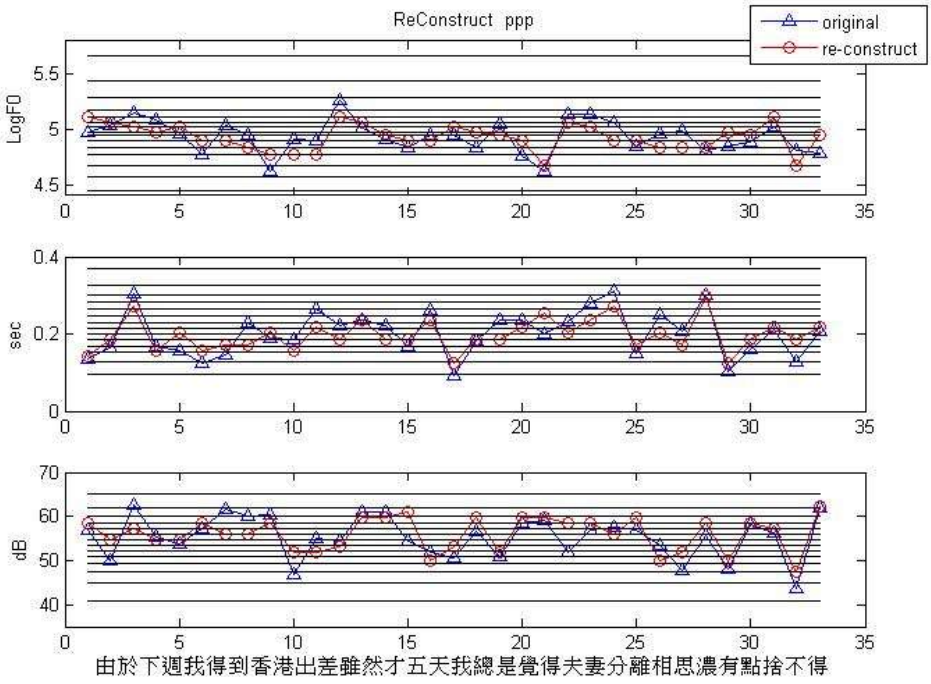
(f) Merry 的韻律狀態標記



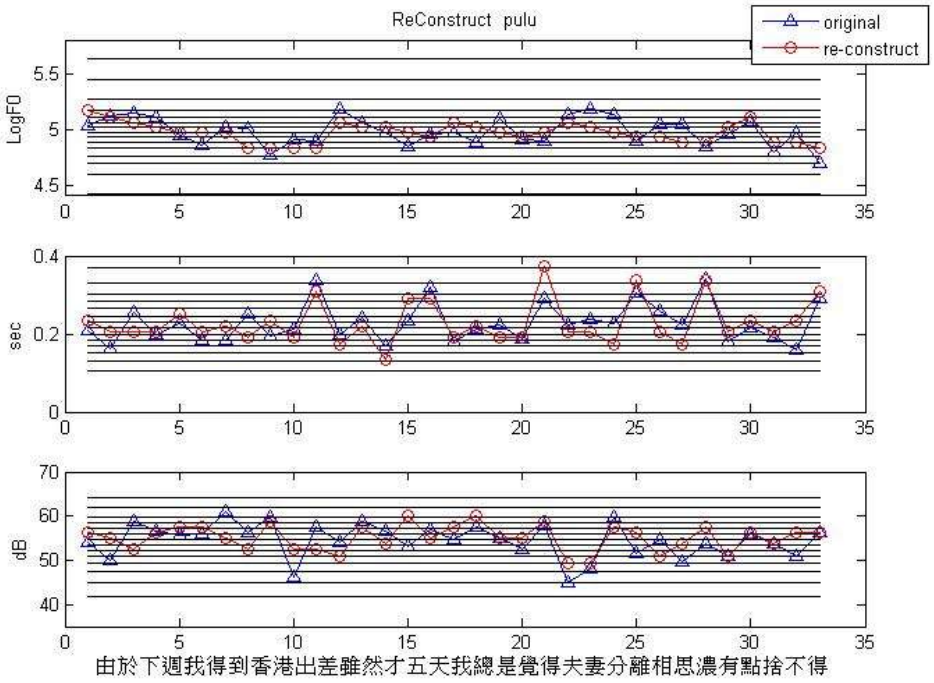
(g) normal 的韻律狀態標記



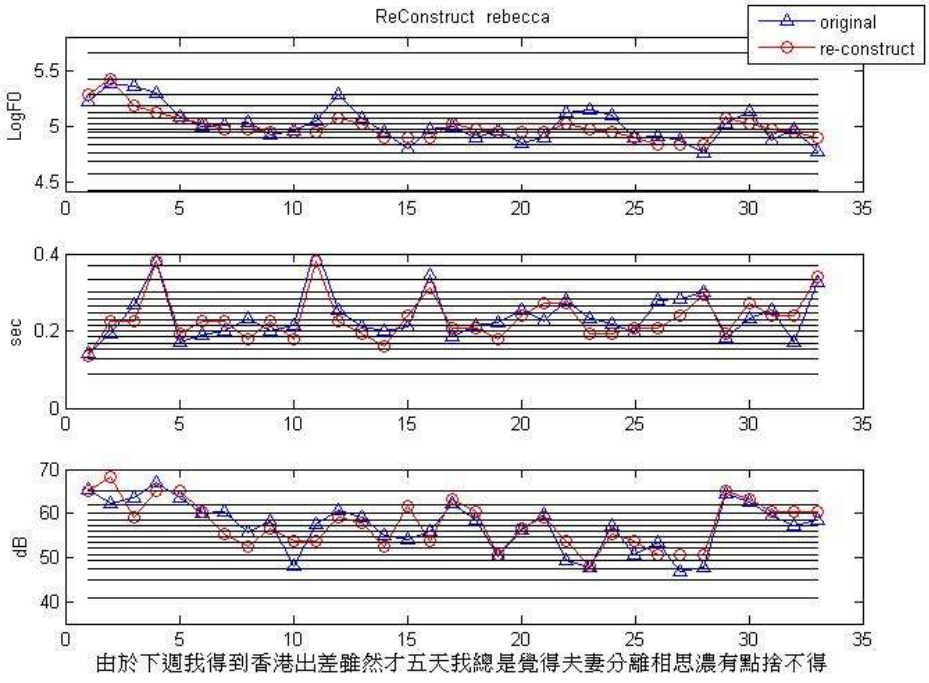
(h) Paul 的韻律狀態標記



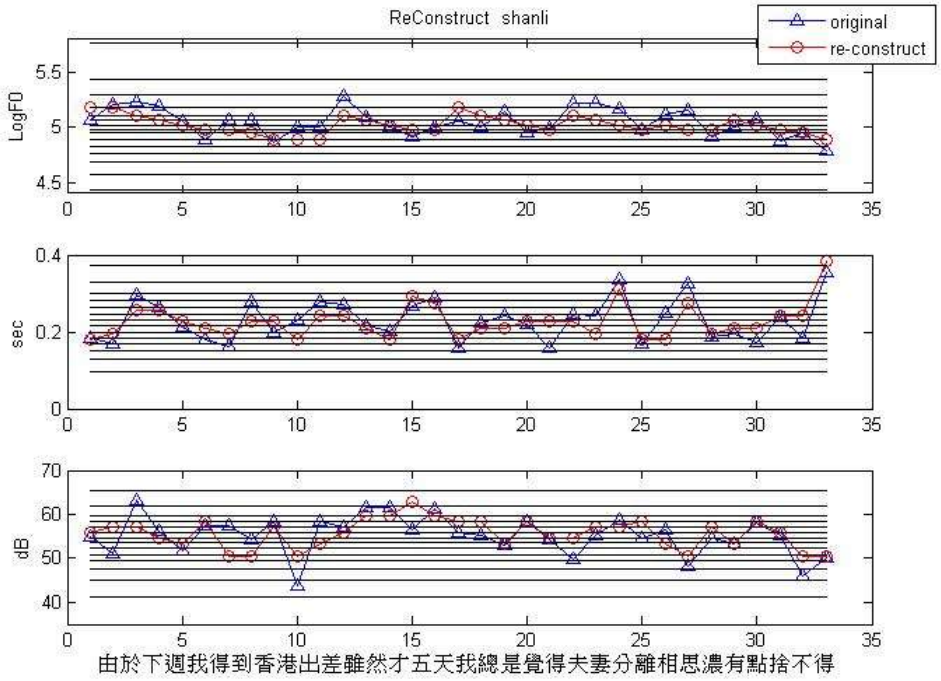
(i) ppp 的韻律狀態標記



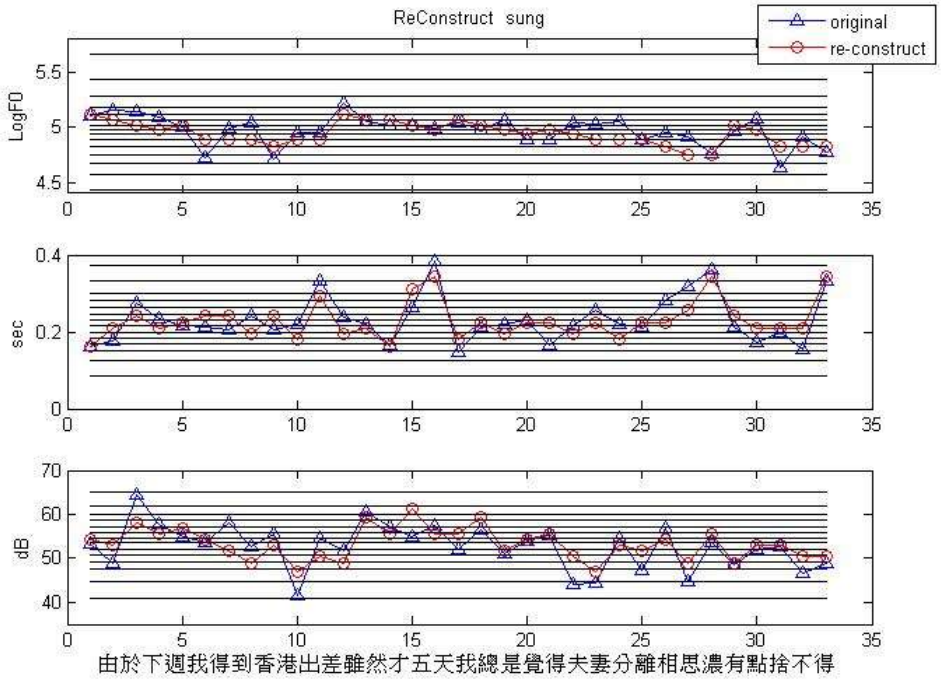
(j) pulu 的韻律狀態標記



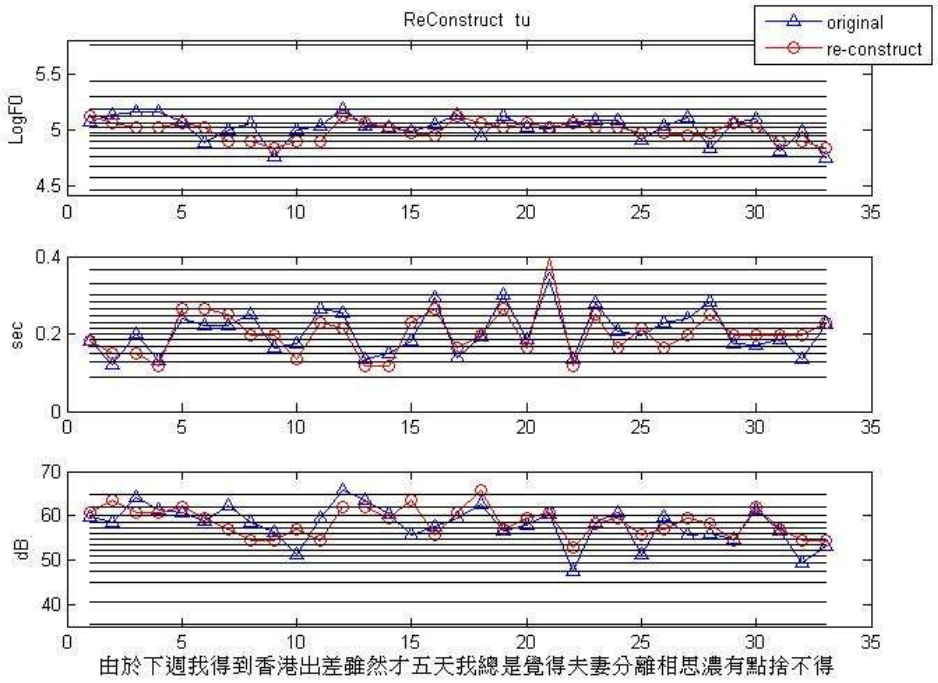
(k) rebecca 的韻律狀態標記



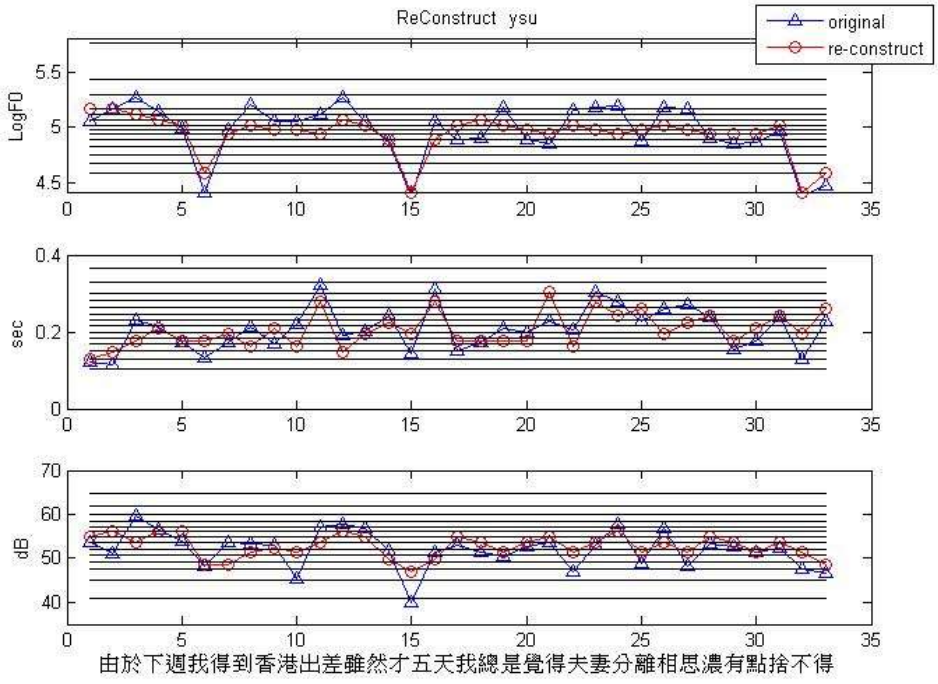
(l) shanli 的韻律狀態標記



(m)sung 的韻律狀態標記



(n) tu 的韻律狀態標記



(o) ysu 的韻律狀態標記

圖 4.2：(a)~(o)分別代表個別語者韻律狀態標記，每張圖由上至下分別為對數音節基頻平均值、音節長度(sec)和音節能量位階(dB)的(mean+ prosodic state)和 original 對照比較圖

圖 4.3 為語者獨立韻律模型及個別語者的語者相關韻律模型中音高韻律狀態標記的統計長條圖，由圖中我們可以發現，全部語者及個別語者的標記分佈都集中在中間的韻律狀態，而兩旁的韻律標記分佈數量都有急遽的下降，此現象是由於語者在說話時會有少許音高特別高或特別低，使得我們在標記韻律狀態時，必須照顧這些較為極端的音高分佈，而這些較為極端的音高在數量上相較於語者的總音節數來的少很多，才導致這個現象。

另外，我們由圖 2.6(b)可以看出各語者正規化後音高分佈的範圍不盡相同，例如：Jimmy 在相對高音及低音部分皆有比較多的資料量、normal 在相對高音及低音部分幾乎沒有什麼資料量、shanli 只有在相對低音的部分有比較多資料量，相對高音部分則沒什麼資料量；由圖 4.3 我們可以觀察到不同語者，例如：Jimmy 的音高韻律狀態標記(第二列第一張圖)從 1 到 16 都有資料量分佈、normal 的音高韻律狀態標記(第二列第四張圖)缺少 1 及 16、shanli 的音高韻律狀態標記(第四列第一張圖)缺少 16。綜合以上兩張圖，我們可以看到每位語者的音高分布範圍，確實有效的被調適成屬於自己的音高韻律狀態標記，亦可反向藉由音高韻律標記，得知每位語者其說話時的音高分布範圍的大小。

同音高韻律狀態標記長條圖的觀察，我們從圖 4.4、圖 2.4(b)及圖 4.5、圖 2.7(b)中也可以發現類似的情形，例如：Arron、normal、tu、ysu 的音節長度韻律狀態標記皆缺少 16，從圖 2.4(b)中確實可以觀察到這幾位語者比較沒有較長的音節長度分佈；byetwo、pulu 音節能量韻律狀態標記皆缺少 1 及 2，從圖 2.7(b)中確實可以觀察到這幾位語者比較沒有較小的音節能量分佈。

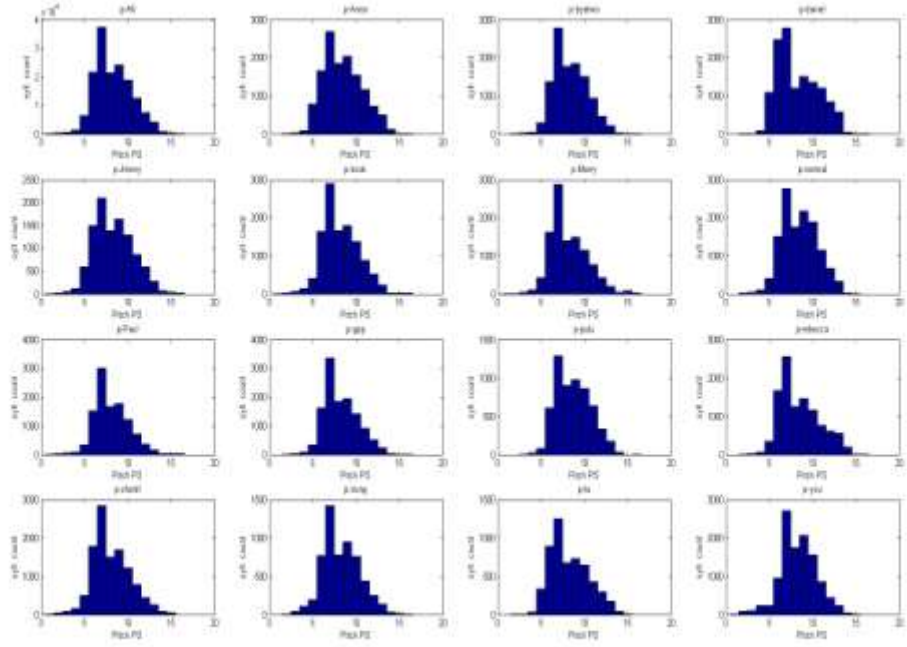


圖 4.3：全部語者(左上)及個別語者的音高韻律狀態標記分佈長條圖

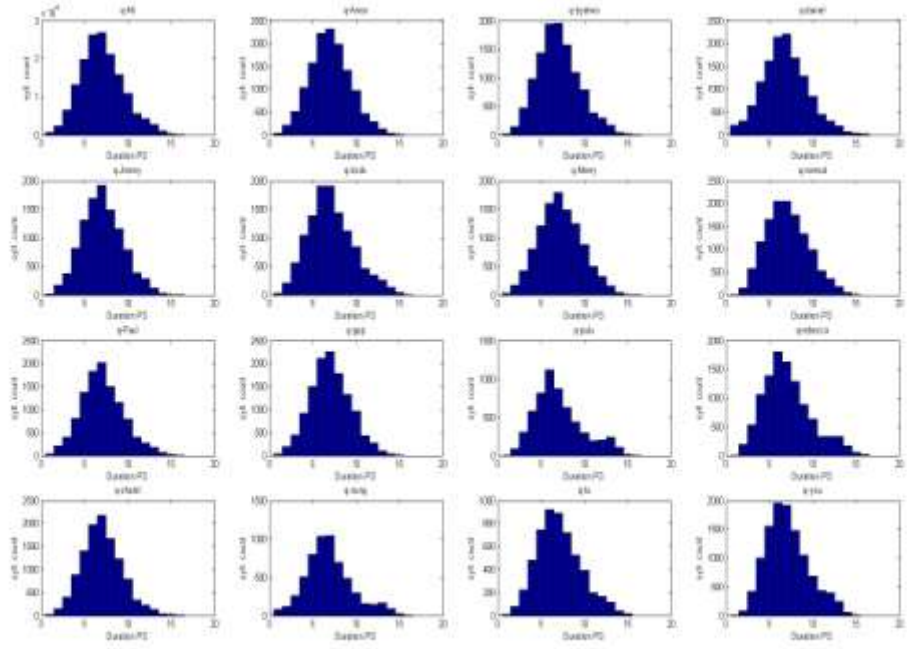


圖 4.4：全部語者(左上)及個別語者的音節長度韻律狀態標記分佈長條圖

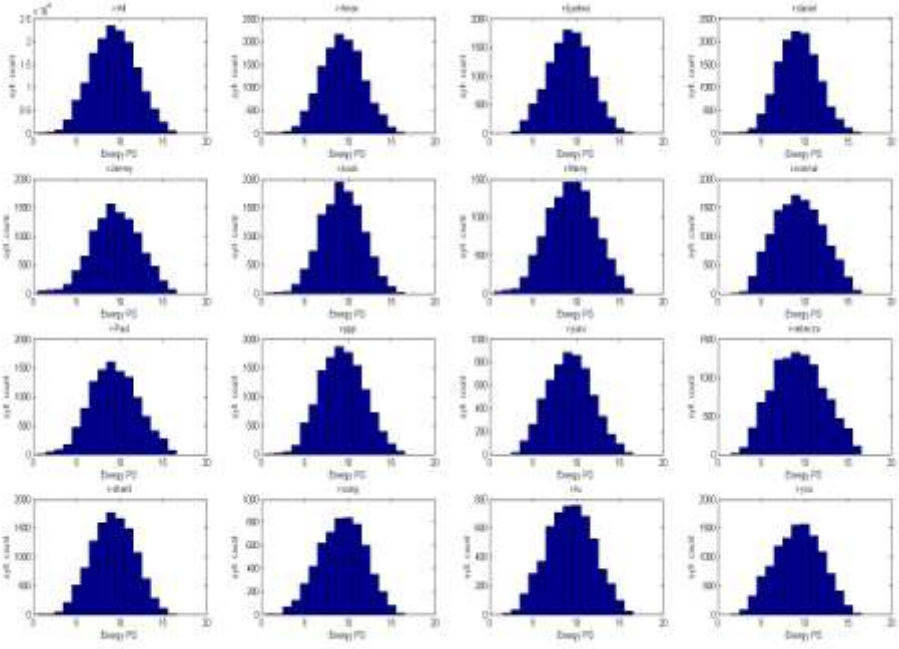


圖 4.5：全部語者(左上)及個別語者的音節能量韻律狀態標記分佈長條圖

4.2 韻律轉換實驗結果

此章節我們利用表 4.2 所述的測試語料進行每位語者的韻律轉換，並利用客觀性評估標準，比較高斯正規化轉換方法及本研究所提出的 MAP 調適法則轉換方法。

4.2.1 實驗客觀性評估標準

為了評估轉換是否成功的將 Source 語者貼近於 Target 語者的分布，我們使用 NMSE (Normalized Mean Square Error)評估轉換後的結果；若 NMSE 小於 1 代表轉換有效果，值越小轉換結果越好。

4.2.1.1 基頻轉換之客觀性評估：

以 NMSE 的評估標準如下：

$$NMSE = \frac{\frac{1}{N}\sum_{1}^{N}\|y_n - \hat{y}_n\|^2}{\frac{1}{N}\sum_{1}^{N}\|y_n - x_n\|^2} \quad (4\text{-}1)$$

其中 N 為測試語料中所有的音節數，x_n、y_n 與 $\hat{y}_n$ 分別為來源、目標以及轉換後的基頻軌跡參數。

4.2.1.2 音節長度與能量轉換之客觀性評估：

以 NMSE 的評估標準如下：

$$NMSE = \frac{\frac{1}{N}\sum_{1}^{N}(y_n - \hat{y}_n)^2}{\frac{1}{N}\sum_{1}^{N}(y_n - x_n)^2} \quad (4\text{-}2)$$

其中 N 為測試語料中所有的音節數，x_n、y_n 與 $\hat{y}_n$ 分別為來源、目標以及轉換後的音長與能量。

4.2.1 韻律轉換實驗結果與分析

表 4.4 到表 4.9 分別列出了高斯正規化法及 MAP 調適法則韻律轉換方法對 15 位語者音節基頻軌跡、音節長度和音節能量互相轉換之 NMSE，表 4.10 則是分別就 15 位語者音節基頻軌跡、音節長度和音節能量互相轉換之 NMSE 的加總。由表 4.10 中可以看出，本研究所提出的 MAP 調適法則韻律轉換方法較傳統的高斯正規化法在音節基頻軌跡、音節長度和音節能量的轉換上皆較高斯正規化法有較好的效果。

首先，我們從表 4.4 和表 4.5 中可以看到，當男生和女生互相轉換時，無論是高斯正規化法或 MAP 調適法則韻律轉換方法，都可以得到很好的轉換效果，其原因為男女生在先天上講話音高就有所區別，如圖 4.6(a)、圖 4.6(b)所示，所以經過轉換後，因為我們利用 Target 語者的統計資訊做反正規化動作，所以可以得到相較於同性別互相轉換時更好的結果；其次，我們觀察圖 4.7、圖 4.8，這兩張圖中的 Target 語者分別是 Jimmy 及 Merry，這兩位語者從 4.1 節中韻律標記的結果分析可以發現，他們在句子結尾時會有習慣性音高拉高的說話特性，我們在這

兩張圖中也可以看到一樣的特性，而這樣的特性在使用高斯正規化法做轉換時，由於高斯正規化法只是單純的移動平均值及利用變異數放大縮小音節基頻軌跡，所以當 Source 語者的句子結尾沒有音高拉高的特性時，不可能轉換成 Target 語者句子結尾音高拉高的特性；而利用 MAP 調適法則韻律轉換方法，我們在部份轉換例子時可以達成讓本來沒有尾音拉高的 Source 語者，轉換成 Target 語者有尾音拉高的說話特性，其原因為有部分語者在句子結尾時沒有特定的音高降低或拉高的特性，而 MAP 調適法則韻律轉換方法將音節基頻軌跡拆解成許多因子的影響，如 2-12 式所示，所以當我們將上述的語者當成 Source 語者時，我們有機會利用韻律狀態標記呈現出 Target 說語者句子結尾音高拉高的特性，如圖 4.9 所示。

表 4.4：利用高斯正規化法做音節基頻軌跡轉換之 NMSE

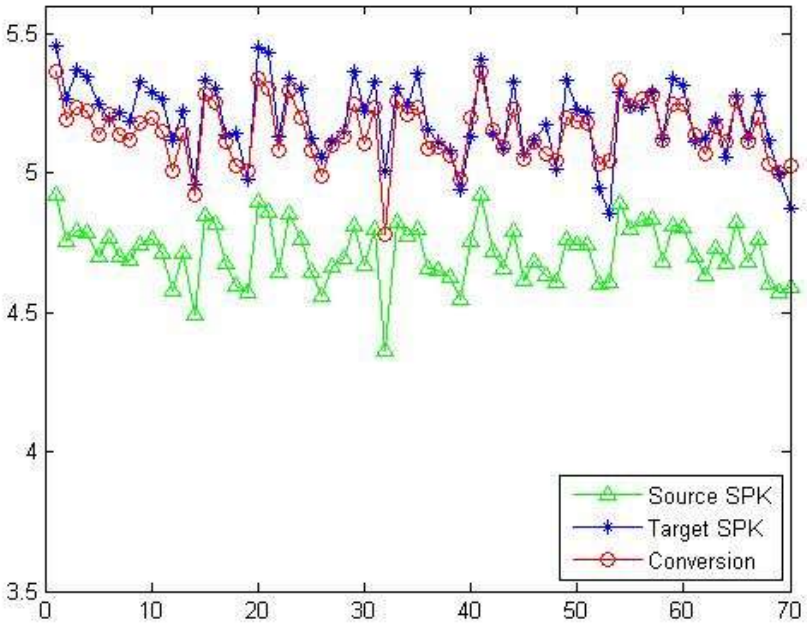
Target / Source	Arron	daniel	Jimmy	kook	Merry	Paul	pulu	sung	tu
Arron	X	0.95	1.01	0.87	0.67	0.93	0.85	0.55	0.94
daniel	0.62	X	0.75	0.70	1.13	0.69	0.60	1.19	0.75
Jimmy	1.17	1.32	X	0.76	1.13	0.94	1.58	1.23	1.53
kook	1.33	1.65	1.02	X	1.33	0.97	1.50	1.30	1.45
Merry	0.50	1.29	0.73	0.64	X	0.57	0.58	1.35	0.76
Paul	1.19	1.39	1.06	0.82	1.01	X	1.69	1.27	1.53
pulu	0.59	0.65	0.95	0.67	0.55	0.92	X	0.53	1.02
sung	0.16	0.51	0.30	0.24	0.51	0.28	0.22	X	0.26
tu	0.66	0.81	0.93	0.66	0.72	0.82	1.05	0.64	X
byetwo	0.14	0.15	0.26	0.17	0.16	0.27	0.24	0.33	0.21
normal	0.04	0.07	0.11	0.06	0.09	0.12	0.05	0.15	0.08
ppp	0.04	0.07	0.15	0.07	0.08	0.13	0.07	0.13	0.09
rebecca	0.04	0.10	0.11	0.05	0.07	0.09	0.07	0.12	0.07
shanli	0.10	0.12	0.15	0.09	0.11	0.18	0.16	0.22	0.18
ysu	0.13	0.11	0.15	0.12	0.17	0.18	0.18	0.23	0.20

Target / Source	byetwo	normal	ppp	rebecca	shanli	ysu
Arron	0.15	0.18	0.04	0.06	0.15	0.32
daniel	0.10	0.22	0.05	0.09	0.12	0.18
Jimmy	0.32	0.59	0.18	0.18	0.25	0.45
kook	0.29	0.45	0.12	0.11	0.21	0.49
Merry	0.12	0.29	0.06	0.07	0.12	0.33
Paul	0.38	0.69	0.17	0.17	0.34	0.59
pulu	0.18	0.17	0.05	0.07	0.17	0.32
sung	0.10	0.19	0.04	0.05	0.09	0.17
tu	0.16	0.26	0.06	0.07	0.18	0.35
byetwo	X	1.33	0.50	0.51	0.82	1.17
normal	0.30	X	0.28	0.36	0.71	1.05
ppp	0.54	1.27	X	0.88	1.04	1.04
rebecca	0.40	1.19	0.67	X	0.86	0.96
shanli	0.58	2.20	0.71	0.78	X	1.29
ysu	0.49	1.88	0.41	0.51	0.75	X

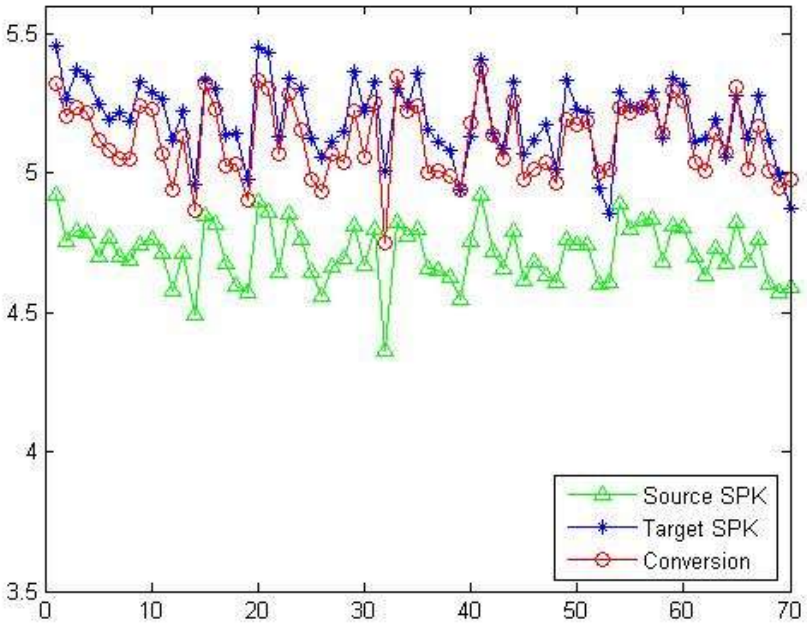
表 4.5：利用 MAP 調適法則轉換方法做音節基頻軌跡轉換之 NMSE

Target / Source	Arron	daniel	Jimmy	kook	Merry	Paul	pulu	sung	tu
Arron	X	0.92	0.99	0.87	0.70	0.94	0.96	0.58	1.03
daniel	0.63	X	0.79	0.73	1.18	0.68	0.67	1.29	0.77
Jimmy	0.69	1.06	X	0.65	0.86	0.73	0.88	0.77	0.98
kook	1.03	1.59	0.94	X	1.11	0.80	1.07	0.96	1.14
Merry	0.43	1.32	0.66	0.62	X	0.53	0.45	1.10	0.62
Paul	1.06	1.50	1.05	0.86	0.94	X	1.28	0.99	1.48
pulu	0.67	0.62	0.94	0.69	0.56	0.94	X	0.59	1.06
sung	0.23	0.81	0.31	0.27	0.65	0.29	0.28	X	0.37
tu	0.64	0.61	0.91	0.63	0.74	0.83	0.85	0.56	X
byetwo	0.12	0.11	0.23	0.14	0.14	0.25	0.20	0.25	0.20
normal	0.04	0.07	0.12	0.07	0.10	0.12	0.07	0.16	0.08
ppp	0.04	0.08	0.15	0.08	0.09	0.13	0.08	0.13	0.09
rebecca	0.05	0.10	0.11	0.05	0.07	0.09	0.07	0.12	0.08
shanli	0.10	0.10	0.14	0.09	0.11	0.15	0.16	0.20	0.17
ysu	0.09	0.08	0.13	0.12	0.15	0.15	0.14	0.18	0.15

Target / Source	byetwo	normal	ppp	rebecca	shanli	ysu
Arron	0.14	0.20	0.04	0.06	0.15	0.29
daniel	0.10	0.26	0.05	0.10	0.11	0.21
Jimmy	0.23	0.38	0.11	0.11	0.18	0.38
kook	0.25	0.35	0.08	0.09	0.19	0.46
Merry	0.11	0.22	0.05	0.06	0.10	0.30
Paul	0.33	0.66	0.13	0.14	0.34	0.62
pulu	0.17	0.22	0.06	0.07	0.17	0.33
sung	0.13	0.28	0.05	0.06	0.12	0.20
tu	0.14	0.23	0.06	0.07	0.17	0.30
byetwo	X	1.16	0.44	0.40	0.72	1.13
normal	0.26	X	0.30	0.40	0.71	1.08
ppp	0.57	1.36	X	1.00	1.03	1.08
rebecca	0.39	1.19	0.72	X	0.83	0.90
shanli	0.53	2.26	0.64	0.74	X	1.12
ysu	0.39	1.41	0.31	0.38	0.59	X

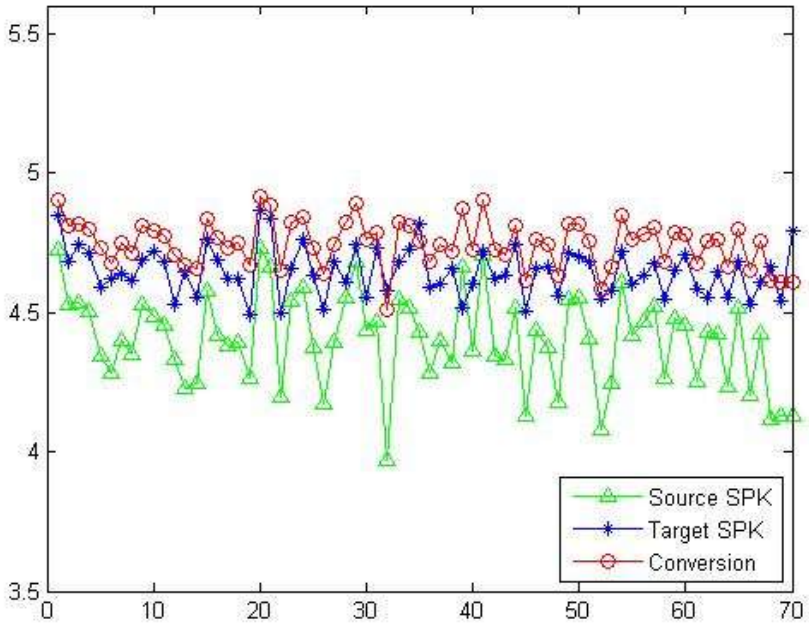


(a)

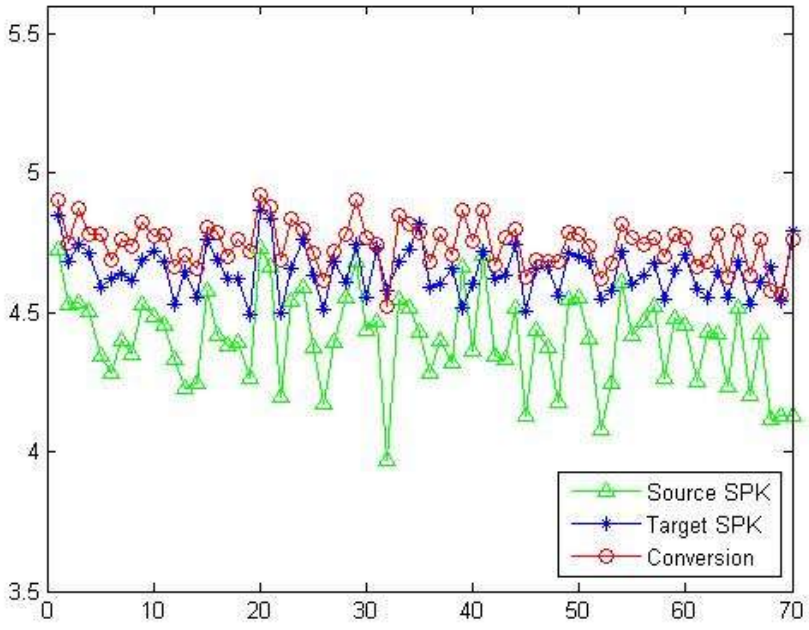


(b)

圖 4.6：Source 語者(Arron)與 Target 語者(byetwo)利用(a)高斯正規化法，(b) MAP 調適法則韻律轉換方法，做音節基頻軌跡轉換結果

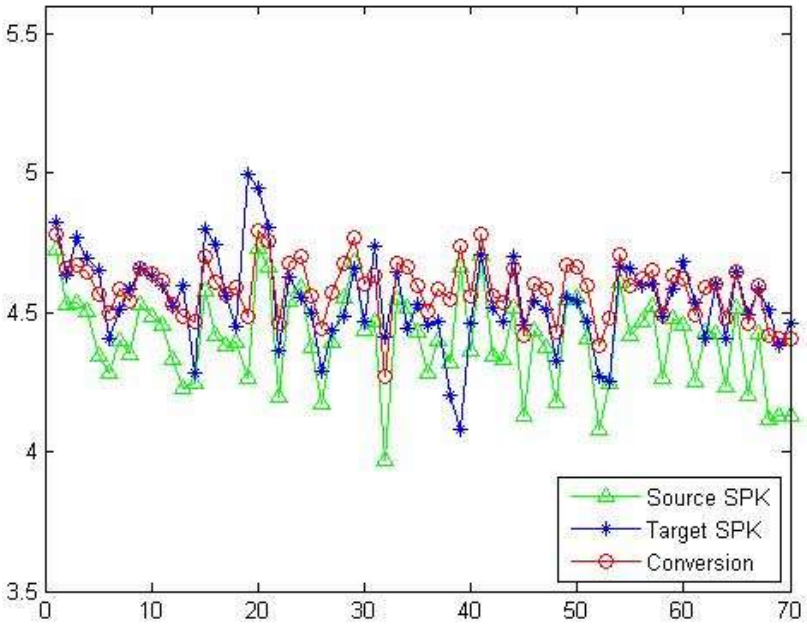


(a)

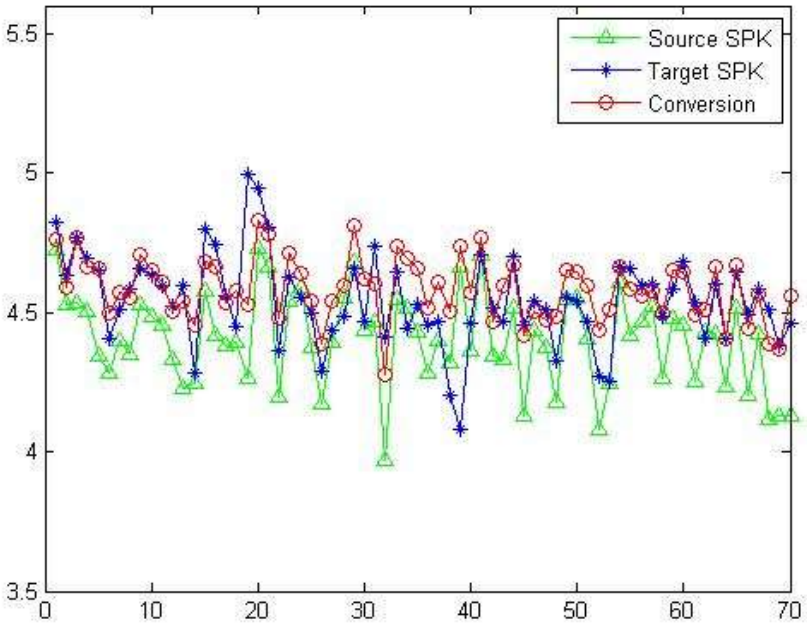


(b)

圖 4.7：Source 語者(sung)與 Target 語者(Jimmy)利用(a)高斯正規化法，(b) MAP 調適法則韻律轉換方法，做音節基頻軌跡轉換結果

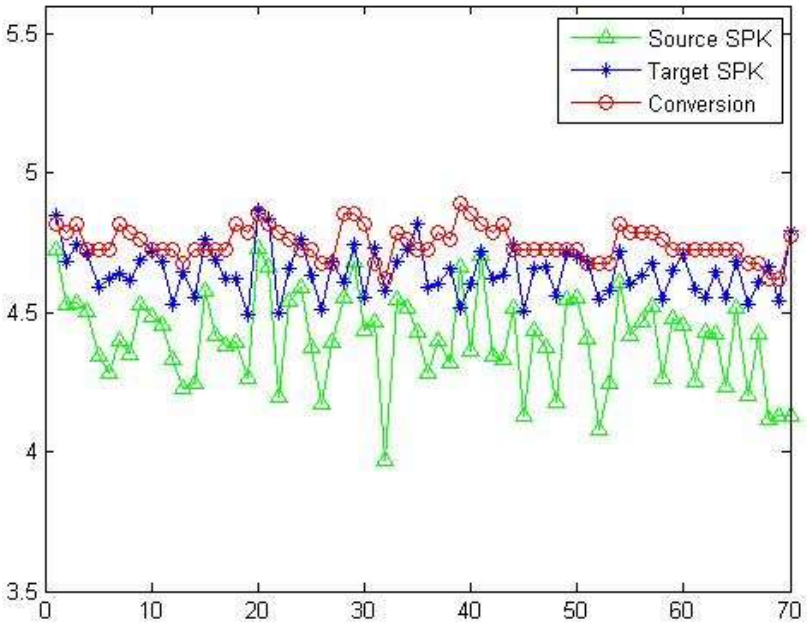


(a)

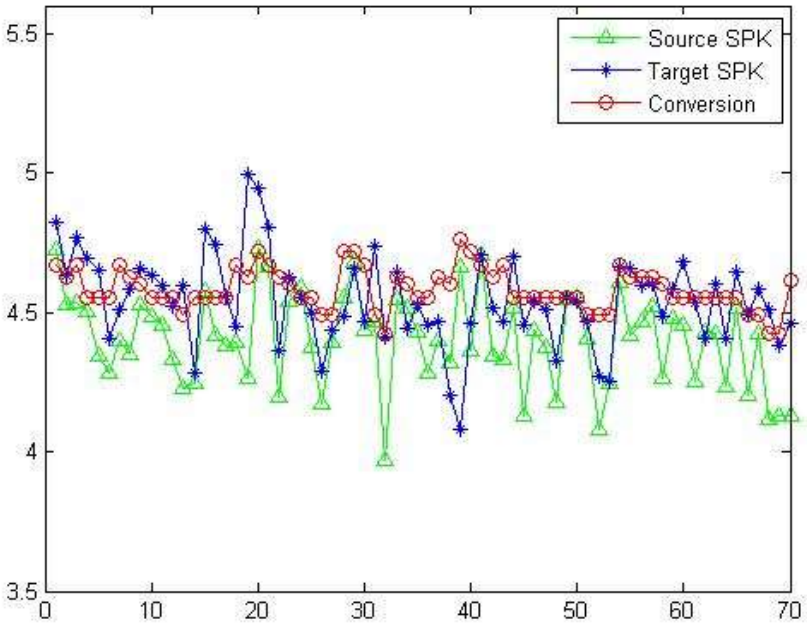


(b)

圖 4.8：Source 語者(sung)與 Target 語者(Merry)利用(a)高斯正規化法，(b) MAP 調適法則韻律轉換方法，做音節基頻軌跡轉換結果



(a)



(b)

圖 4.9：Source 語者(sung) (a) Target 語者(Jimmy)，(b) Target 語者(Merry)使用 (mean + prosodic state)轉換的結果

表 4.6 和表 4.7 分別列出了高斯正規化法及 MAP 調適法則韻律轉換方法對 15 位語者音節長度互相轉換之 NMSE，從中我們可以看出 pulu、rebecca、ysu 這三位 Target 語者的轉換效果較差，其原因為這三位語者的音節長度變化較為劇烈，如圖 4.10 到圖 4.12 所示，從 Target 的線段都可以看到有特別長的音節長度；其次，我們從 4.1 小節中圖 4.4 可以看到，上述所講的三位語者其音節長度韻律狀態標記分佈相較於其他語者在 11 到 14 的韻律狀態標記其分佈的比例來的高一些，而圖 4.13 為 pulu、rebecca、ysu 三位語者使用(mean + prosodic state)轉換的結果，可以看到轉換的結果受到 Source 語者的韻律狀態標記所主導，所以無法很好的轉換成上述三位語者較為劇烈的音節長度變化。

在 4.1 節中我們看到 Jimmy、Merry 兩位語者在能量分佈上的變動範圍較其他語者來的劇烈，如圖 4.14 所示，而從表 4.8 和表 4.9 中可以看到這兩位語者的轉換效果不佳，其原因為轉換的結果會被 Source 語者的韻律狀態標記所主導，如圖 4.15 所示，所以若 Source 語者與 Target 語者的韻律狀態標記相差太多，我們將無法有效的利用 MAP 調適法則韻律轉換方法進行轉換。

表 4.6：利用高斯正規化法做音節長度轉換之 NMSE

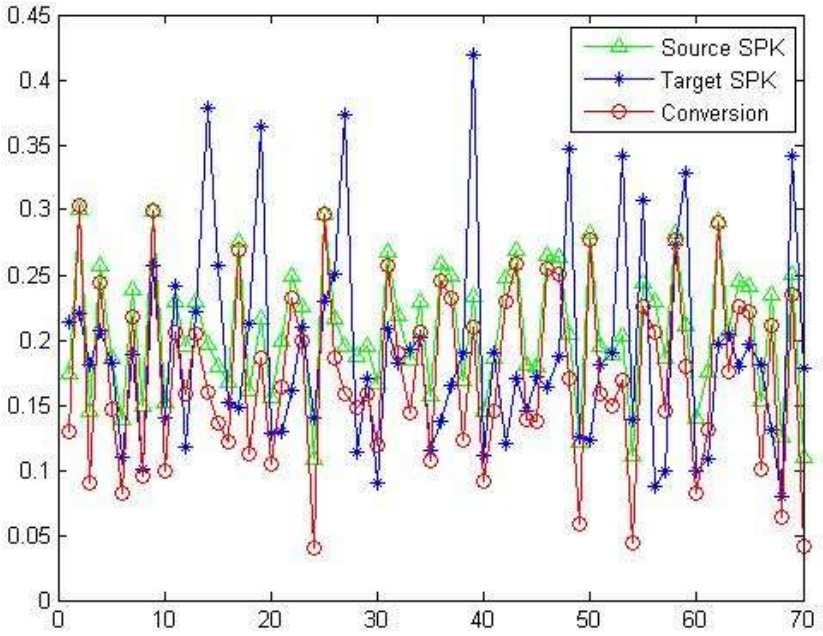
Target / Source	Arron	daniel	Jimmy	kook	Merry	Paul	pulu	sung	tu
Arron	X	1.17	0.56	1.14	0.92	0.92	1.38	1.40	0.69
daniel	0.88	X	0.53	1.00	0.79	0.86	1.16	1.10	0.75
Jimmy	0.44	0.55	X	0.54	0.41	1.01	1.06	0.64	0.98
kook	0.90	1.06	0.56	X	0.87	0.89	1.33	1.40	0.65
Merry	1.16	1.32	0.67	1.37	X	1.00	1.47	1.46	0.83
Paul	0.62	0.77	0.88	0.76	0.54	X	1.12	0.79	0.96
pulu	0.72	0.81	0.72	0.87	0.61	0.87	X	0.79	0.80
sung	0.97	1.02	0.58	1.22	0.80	0.81	1.04	X	0.71
tu	0.52	0.75	0.95	0.61	0.49	1.06	1.15	0.77	X
byetwo	0.80	1.00	1.31	0.98	0.59	1.25	1.31	0.69	1.41
normal	0.76	0.75	0.44	0.90	0.69	0.66	0.89	0.97	0.51
ppp	1.00	1.23	0.67	1.17	0.85	0.99	1.41	1.25	0.89
rebecca	0.84	0.83	0.55	1.06	0.69	0.72	0.87	0.87	0.61
shanli	0.79	1.05	0.85	0.93	0.68	1.10	1.35	0.99	0.99
ysu	0.64	0.72	0.47	0.73	0.63	0.74	1.03	1.04	0.52

Target / Source	byetwo	normal	ppp	rebecca	shanli	ysu
Arron	0.78	1.27	0.98	1.54	0.85	1.39
daniel	0.73	0.95	0.91	1.14	0.85	1.16
Jimmy	0.99	0.57	0.51	0.77	0.72	0.79
kook	0.76	1.20	0.92	1.54	0.80	1.25
Merry	0.73	1.45	1.06	1.60	0.93	1.71
Paul	0.82	0.75	0.66	0.90	0.80	1.09
pulu	0.67	0.78	0.73	0.83	0.77	1.16
sung	0.47	1.12	0.85	1.10	0.74	1.56
tu	1.03	0.64	0.65	0.83	0.80	0.85
byetwo	X	0.80	0.78	0.80	1.07	1.34
normal	0.47	X	0.68	1.15	0.57	1.18
ppp	0.77	1.15	X	1.32	0.97	1.46
rebecca	0.43	1.05	0.71	X	0.62	1.55
shanli	0.96	0.88	0.88	1.05	X	1.20
ysu	0.60	0.92	0.66	1.31	0.60	X

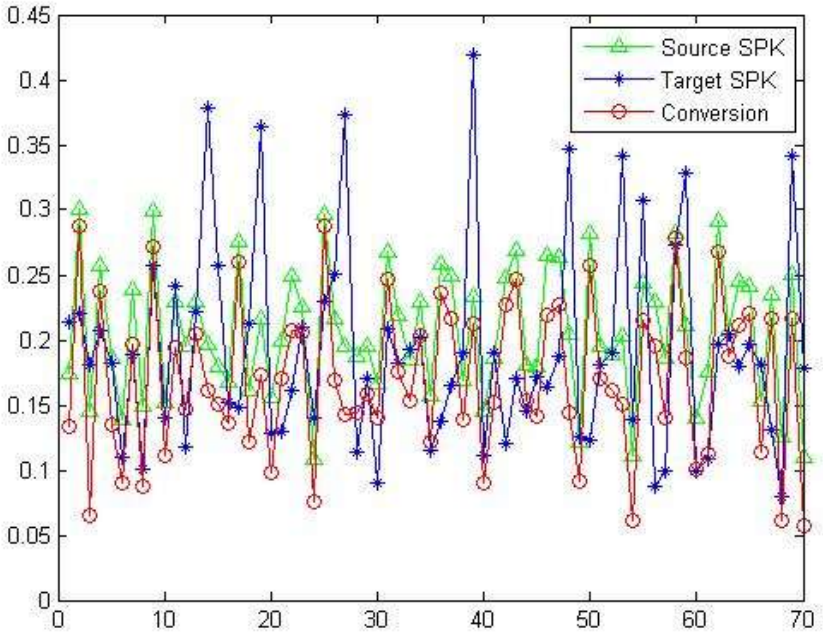
表 4.7：利用 MAP 調適法則轉換方法做音節長度轉換之 NMSE

Target / Source	Arron	daniel	Jimmy	kook	Merry	Paul	pulu	sung	tu
Arron	X	1.03	0.53	1.06	0.90	0.86	1.32	1.29	0.67
daniel	0.96	X	0.58	1.05	0.83	0.90	1.24	1.09	0.80
Jimmy	0.44	0.49	X	0.52	0.41	0.99	1.05	0.59	0.97
kook	0.93	0.99	0.57	X	0.86	0.89	1.35	1.31	0.65
Merry	1.09	1.08	0.61	1.17	X	0.90	1.25	1.24	0.76
Paul	0.66	0.73	0.93	0.76	0.55	X	1.15	0.74	1.01
pulu	0.75	0.73	0.72	0.88	0.61	0.83	X	0.74	0.83
sung	1.05	1.00	0.62	1.26	0.84	0.82	1.07	X	0.76
tu	0.52	0.66	0.94	0.59	0.49	1.05	1.11	0.70	X
byetwo	0.80	0.85	1.31	0.91	0.58	1.16	1.24	0.61	1.37
normal	0.73	0.63	0.43	0.81	0.67	0.61	0.83	0.86	0.49
ppp	0.98	1.06	0.67	1.10	0.85	0.94	1.35	1.15	0.86
rebecca	0.92	0.82	0.61	1.11	0.74	0.75	0.94	0.84	0.67
shanli	0.77	0.93	0.86	0.88	0.67	1.06	1.31	0.92	0.96
ysu	0.65	0.65	0.48	0.70	0.61	0.72	1.02	0.97	0.52

Target / Source	byetwo	normal	ppp	rebecca	shanli	ysu
Arron	0.76	1.27	0.96	1.49	0.83	1.36
daniel	0.74	1.06	0.97	1.18	0.93	1.28
Jimmy	0.99	0.60	0.52	0.75	0.71	0.79
kook	0.76	1.23	0.93	1.54	0.81	1.30
Merry	0.66	1.45	1.00	1.41	0.85	1.58
Paul	0.83	0.86	0.70	0.89	0.85	1.15
pulu	0.64	0.83	0.71	0.83	0.78	1.18
sung	0.47	1.28	0.92	1.10	0.80	1.72
tu	1.01	0.65	0.65	0.81	0.81	0.83
byetwo	X	0.83	0.79	0.76	1.06	1.34
normal	0.45	X	0.65	1.09	0.55	1.14
ppp	0.77	1.23	X	1.28	0.97	1.43
rebecca	0.45	1.37	0.79	X	0.73	1.78
shanli	0.97	0.89	0.88	1.03	X	1.17
ysu	0.59	0.96	0.66	1.35	0.61	X

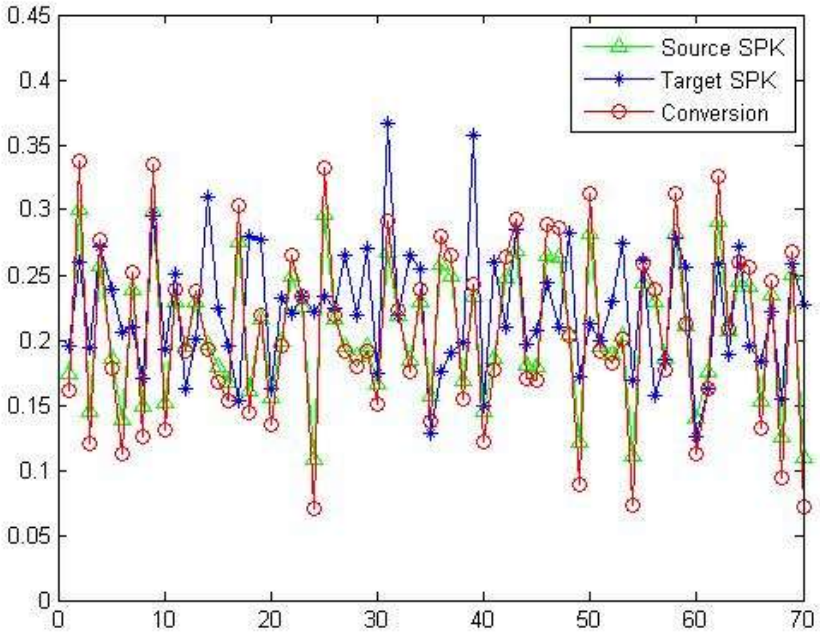


(a)

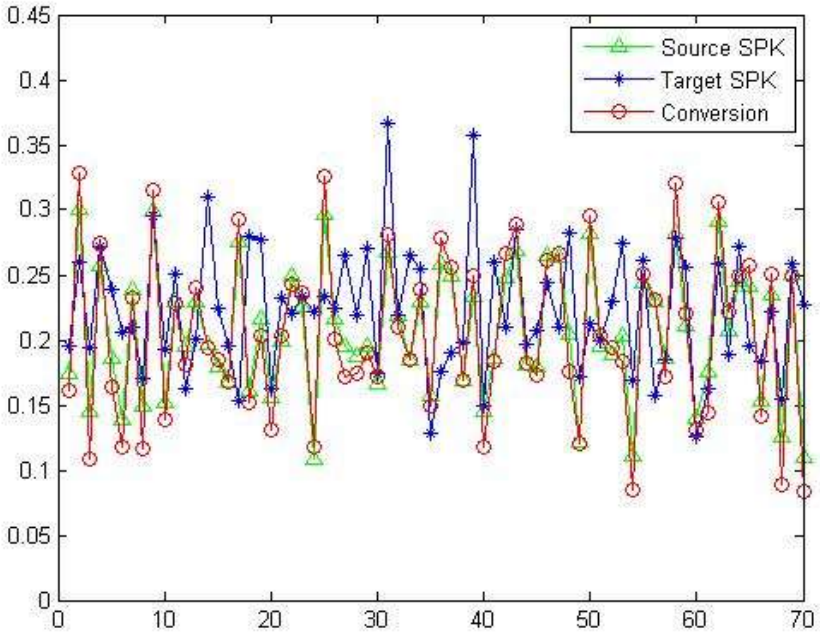


(b)

圖 4.10：Source 語者(ppp)與 Target 語者(rebecca)利用(a)高斯正規化法，(b) MAP 調適法則韻律轉換方法，做音節長度轉換結果

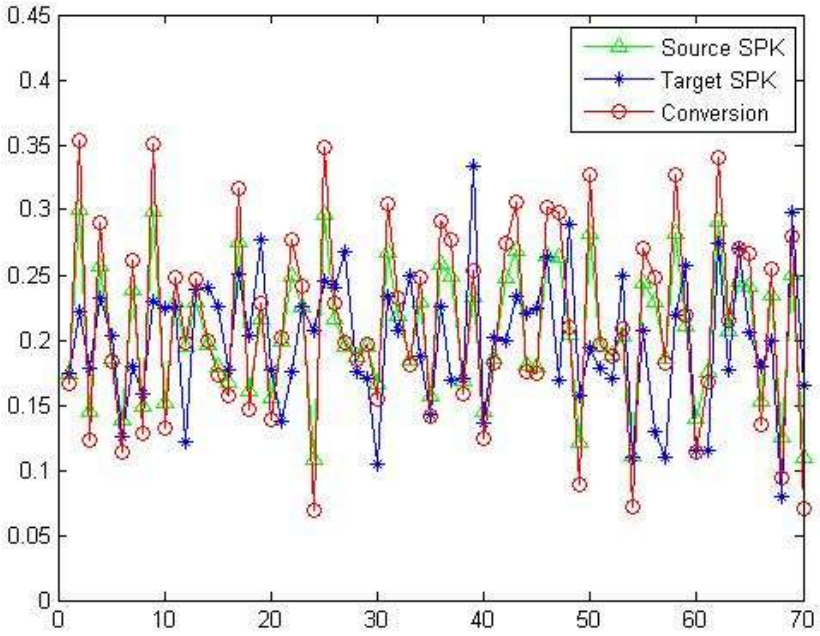


(a)

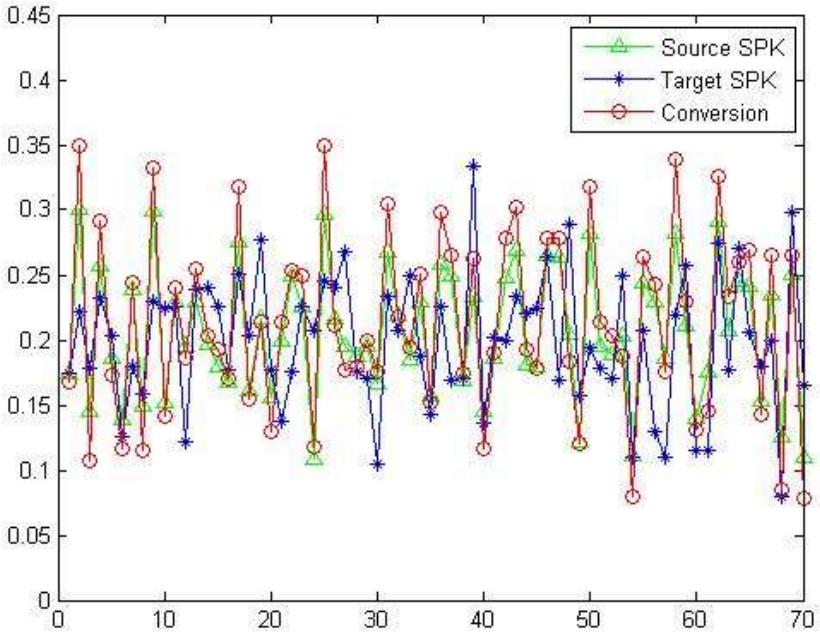


(b)

圖 4.11：Source 語者(ppp)與 Target 語者(pulu)利用(a)高斯正規化法，(b) MAP 調適法則韻律轉換方法，做音節長度轉換結果

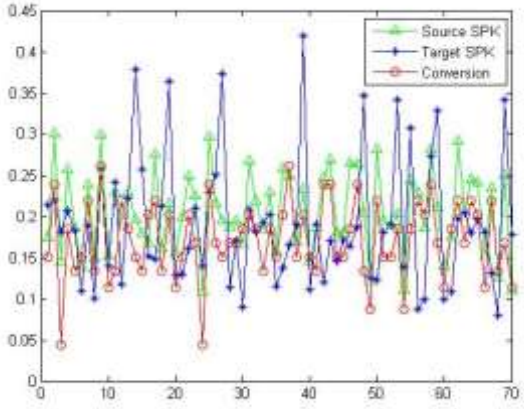


(a)

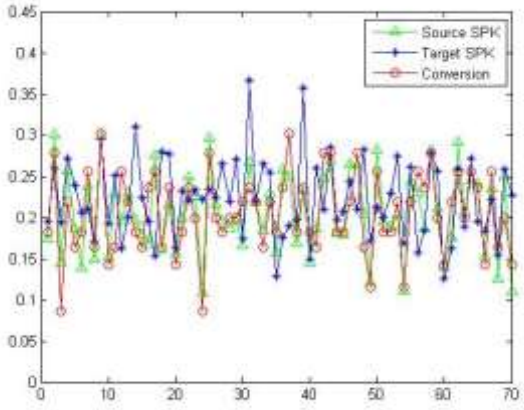


(b)

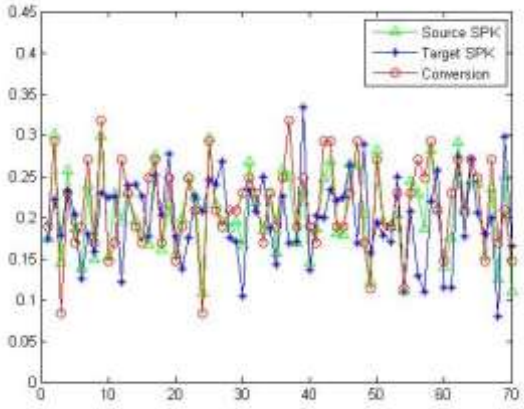
圖 4.12：Source 語者(ppp)與 Target 語者(ysu)利用(a)高斯正規化法，(b) MAP 調適法則韻律轉換方法，做音節長度轉換結果



(a)



(b)



(c)

圖 4.13：Source 語者(ppp)與(a) Target 語者(rebecca)，(b) Target 語者(pulu)，(c) Target 語者(ysu)使用(mean + prosodic state)轉換的結果

表 4.8：利用高斯正規化法做音節能量轉換之 NMSE

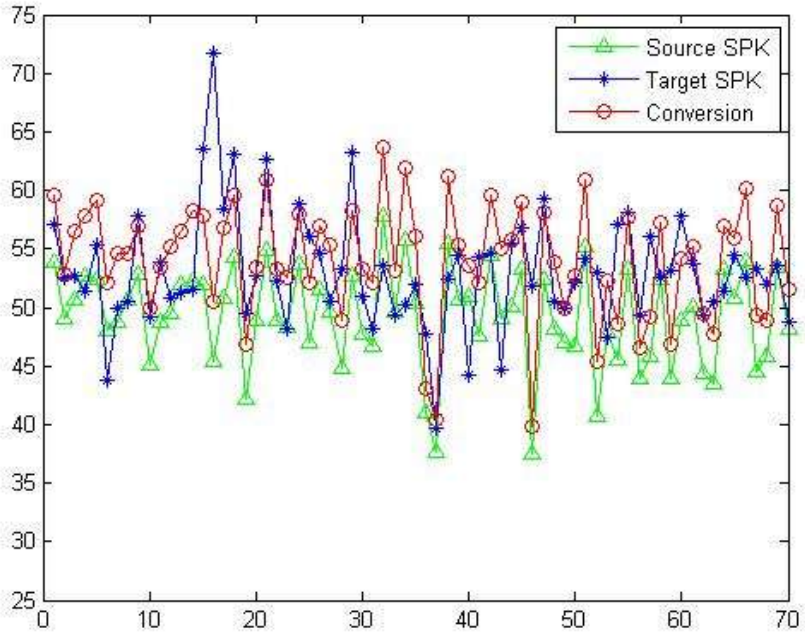
Target / Source	Arron	daniel	Jimmy	kook	Merry	Paul	pulu	sung	tu
Arron	X	0.82	0.97	0.38	0.85	0.53	0.80	0.63	0.86
daniel	0.80	X	0.92	1.08	1.16	1.02	0.58	0.55	0.58
Jimmy	0.91	0.90	X	0.66	1.13	0.78	0.70	0.66	0.71
kook	0.33	0.95	0.60	X	0.63	1.05	0.26	0.27	0.25
Merry	0.65	0.92	0.92	0.56	X	0.75	0.60	0.62	0.55
Paul	0.54	1.06	0.83	1.23	0.98	X	0.46	0.50	0.41
pulu	1.21	0.90	1.13	0.47	1.18	0.69	X	0.84	0.97
sung	1.07	0.96	1.18	0.52	1.37	0.84	0.94	X	0.86
tu	1.28	0.89	1.12	0.44	1.07	0.60	0.96	0.76	X
byetwo	0.93	0.75	0.85	0.37	0.72	0.40	0.46	0.36	0.64
normal	0.54	0.56	0.51	0.18	0.42	0.29	0.36	0.25	0.44
ppp	0.67	1.21	1.03	1.82	1.67	1.28	0.57	0.84	0.51
rebecca	1.63	0.91	1.35	0.38	1.47	0.88	1.38	1.07	1.36
shanli	0.64	0.66	0.65	0.26	0.49	0.30	0.36	0.25	0.47
ysu	1.29	0.76	1.02	0.22	0.89	0.59	0.92	0.59	1.10

Target / Source	byetwo	normal	ppp	rebecca	shanli	ysu
Arron	0.59	0.84	0.44	0.88	0.47	1.03
daniel	0.46	0.85	0.77	0.48	0.47	0.59
Jimmy	0.50	0.76	0.64	0.69	0.45	0.77
kook	0.20	0.24	1.02	0.17	0.16	0.15
Merry	0.35	0.50	0.84	0.61	0.27	0.55
Paul	0.26	0.46	0.84	0.48	0.23	0.48
pulu	0.45	0.87	0.56	1.13	0.40	1.11
sung	0.38	0.66	0.92	0.97	0.31	0.79
tu	0.61	1.04	0.50	1.10	0.52	1.31
byetwo	X	1.69	0.18	0.75	1.06	1.51
normal	0.68	X	0.23	0.33	0.67	0.53
ppp	0.17	0.55	X	0.82	0.18	0.72
rebecca	0.87	0.95	0.99	X	0.70	1.03
shanli	0.92	1.44	0.16	0.52	X	1.04
ysu	1.20	1.04	0.59	0.70	0.96	X

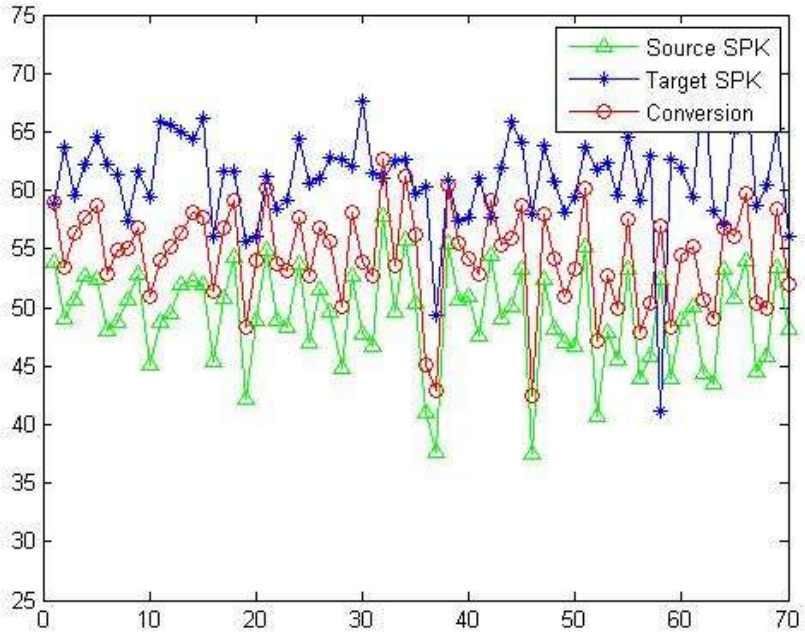
表 4.9：利用 MAP 調適法則轉換方法做音節能量轉換之 NMSE

Target / Source	Arron	daniel	Jimmy	kook	Merry	Paul	pulu	sung	tu
Arron	X	0.81	0.89	0.37	0.81	0.49	0.80	0.62	0.82
daniel	0.68	X	0.59	0.72	0.71	0.58	0.40	0.35	0.38
Jimmy	1.03	0.91	X	0.67	1.13	0.74	0.74	0.68	0.73
kook	0.33	0.94	0.53	X	0.57	0.95	0.25	0.25	0.24
Merry	0.65	0.91	0.89	0.57	X	0.71	0.60	0.62	0.54
Paul	0.61	1.07	0.79	1.28	0.95	X	0.47	0.51	0.41
pulu	1.14	0.89	1.03	0.45	1.06	0.61	X	0.83	0.94
sung	1.06	0.95	1.10	0.52	1.30	0.77	0.95	X	0.84
tu	1.26	0.88	1.04	0.44	1.01	0.55	1.00	0.75	X
byetwo	0.92	0.75	0.77	0.36	0.66	0.36	0.48	0.36	0.63
normal	0.52	0.56	0.48	0.17	0.39	0.27	0.35	0.24	0.43
ppp	0.68	1.21	0.97	1.78	1.57	1.16	0.57	0.83	0.50
rebecca	1.76	0.90	1.22	0.38	1.40	0.80	1.41	1.06	1.35
shanli	0.66	0.66	0.62	0.26	0.47	0.29	0.37	0.25	0.48
ysu	1.23	0.75	0.92	0.22	0.82	0.53	0.84	0.55	1.02

Target / Source	byetwo	normal	ppp	rebecca	shanli	ysu
Arron	0.57	0.84	0.43	0.86	0.45	1.01
daniel	0.31	0.57	0.52	0.33	0.27	0.41
Jimmy	0.53	0.82	0.66	0.71	0.45	0.82
kook	0.19	0.23	0.95	0.17	0.15	0.15
Merry	0.34	0.51	0.81	0.61	0.27	0.55
Paul	0.26	0.50	0.82	0.50	0.22	0.50
pulu	0.43	0.83	0.52	1.11	0.38	1.09
sung	0.38	0.66	0.89	0.98	0.29	0.80
tu	0.60	1.05	0.47	1.09	0.50	1.34
byetwo	X	1.69	0.17	1.14	1.04	1.53
normal	0.65	X	0.22	0.32	0.64	0.52
ppp	0.17	0.57	X	0.83	0.19	0.74
rebecca	0.89	0.95	0.97	X	0.66	1.04
shanli	0.93	1.47	0.16	0.51	X	1.05
ysu	1.11	1.00	0.54	0.68	0.84	X

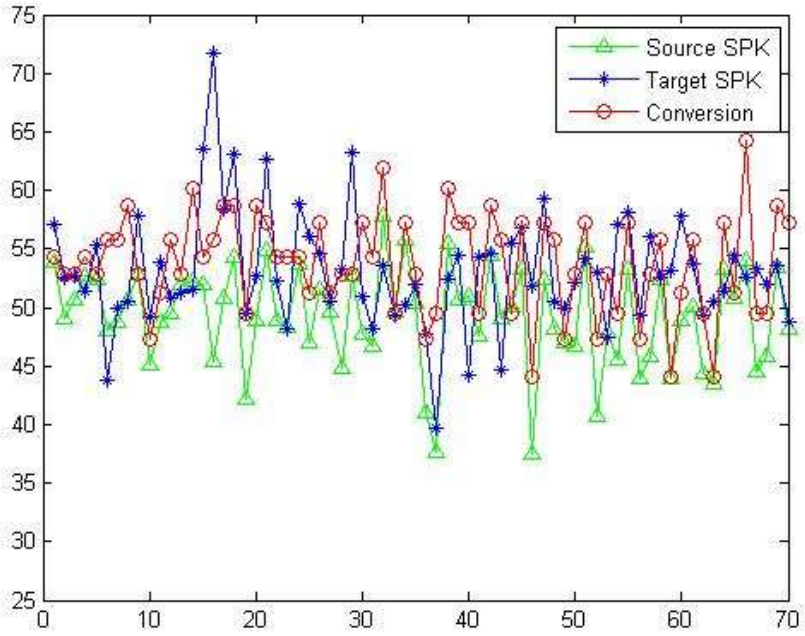


(a)

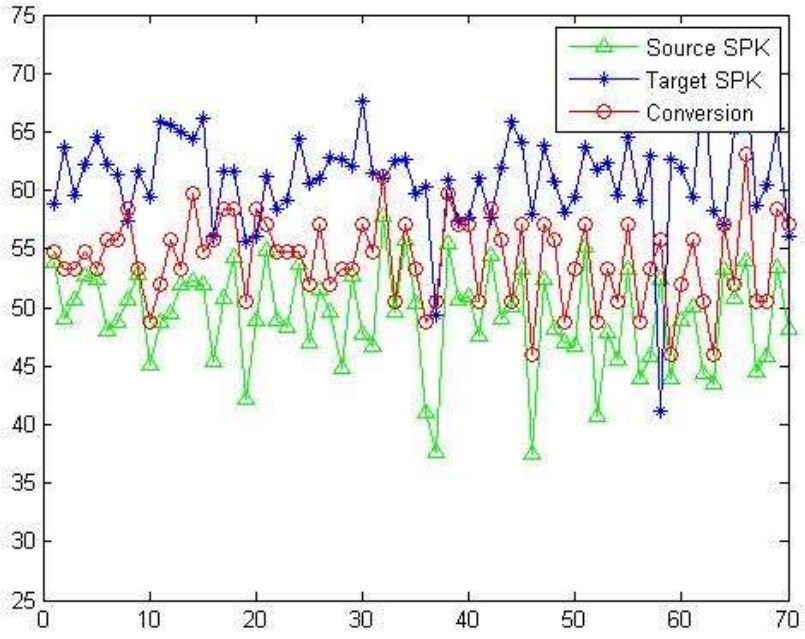


(b)

圖 4.14：Source 語者(Paul)與(a) Target 語者(Merry)，(b) Target 語者(Jimmy)利用 MAP 調適法則韻律轉換方法，做音節能量轉換結果



(a)



(b)

圖 4.15：Source 語者(Paul)與(a) Target 語者(Merry)，(b) Target 語者(Jimmy)利用 (mean + prosodic state)轉換的結果

表 4.10：高斯正規化法及 MAP 調適法則韻律轉換方法在音節基頻軌跡、音節長度及音節能量的 NMSE 總和

	音節基頻軌跡	音節長度	音節能量
高斯正規化法	108.25	190.96	153.94
MAP 調適法則韻律轉換方法	99.87	188.80	147.78

由以上音節基頻軌跡、音節長度及音節能量的討論，我們推測在轉換時，若 Source 與 Target 語者說話特性不同，我們將無法直接利用 MAP 調適法則韻律轉換方法達到有效的轉換，故我們使用誤差矩陣(confusion matrix)來檢驗特定 Source 語者對不同 Target 語者的韻律狀態標記差異程度，並從先前各 NMSE 表格中選出轉換較好及較差的例子，並藉由韻律狀態標記的誤差矩陣統計其狀態差異機率分佈。

首先，我們檢驗音節基頻軌跡轉換，先前討論過在音節基頻軌跡轉換時，由於先天音高的不同，不同性別與同性別的轉換效果有所差異，故在此我們選取轉換效果較好的一男一女及轉換效果較差的一男一女當成 Target 語者來檢視其誤差矩陣。圖 4.16 到圖 4.19 為 Source 語者 Jimmy 分別對 kook(男)、daniel(男)、ppp(女)及 ysu(女)四位 Target 語者利用各自誤差矩陣統計音節基頻軌跡韻律狀態標記差異之機率分佈圖，其中虛線框框代表 Source 及 Target 語者韻律狀態標記相同，而實線框框代表兩語者標記差異較大之處；從表 4.5 中我們看到，在男生方面，Jimmy→daniel 的 NMSE 大於 Jimmy→kook，對照圖 4.16 和圖 4.17 的虛線框框，我們可以看到 Jimmy→kook 比 Jimmy→daniel 相同的韻律標記比例較高，而實線框框部分，則是 Jimmy→kook 比 Jimmy→daniel 來的比例小；而女生方面由先前的討論知道其轉換都有一定的效果，但我們從圖 4.18 和圖 4.19 中，也可以看到 NMSE 較小的 ppp 比 ysu 在需線框框部分比例來的高、實線框框部分比例來的低；故此情況無論在同性別轉換或不同性別轉換，都可以看到相同的結果。

圖 4.20 和圖 4.21 為 Source 語者 Arron 對 Jimmy 及 rebecca 兩位 Target 語者利用各自誤差矩陣統計音節長度韻律狀態標記差異之機率分佈圖，4.22 和圖 4.23 為 Source 語者 byetwo 對 sung 及 normal 兩位 Target 語者利用各自誤差矩陣統計音節能量韻律狀態標記差異之機率分佈

圖，同音節基頻軌跡轉換的討論，在音節長度及音節能量轉換中也可以看到相似的情況。

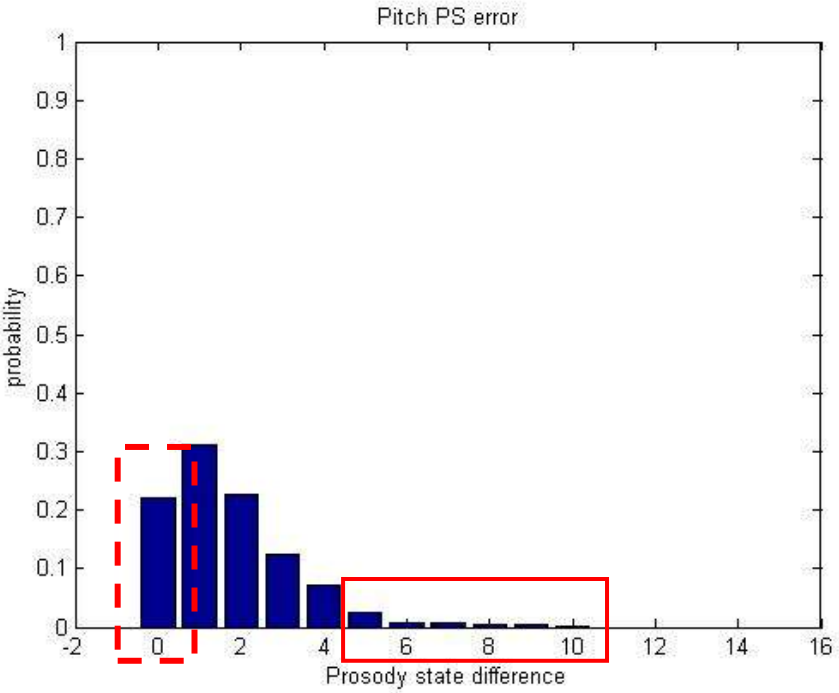


圖 4.16：利用誤差矩陣統計音節基頻軌跡韻律狀態標記差異之機率分佈圖(Jimmy→kook)

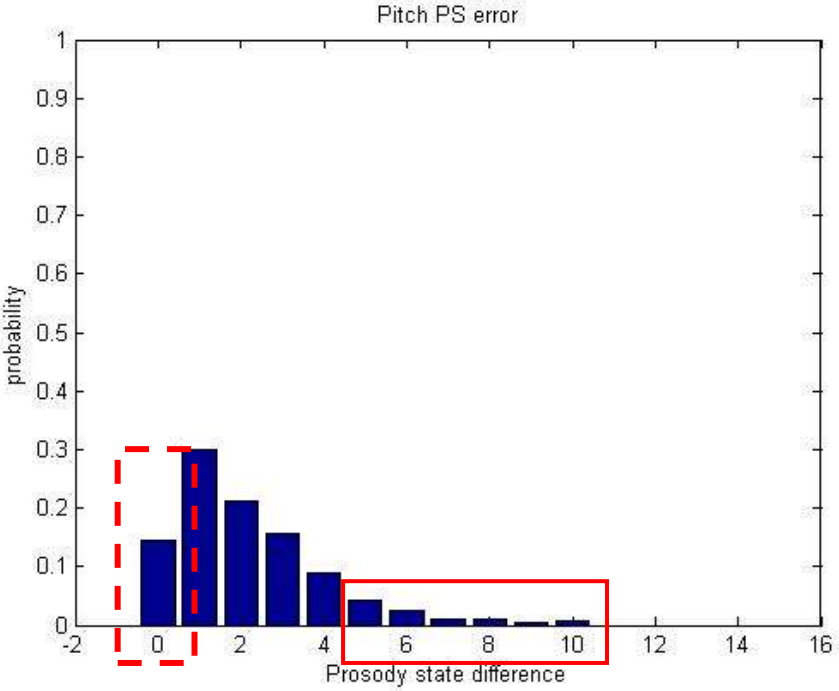


圖 4.17：利用誤差矩陣統計音節基頻軌跡韻律狀態標記差異之機率分佈圖(Jimmy→daniel)

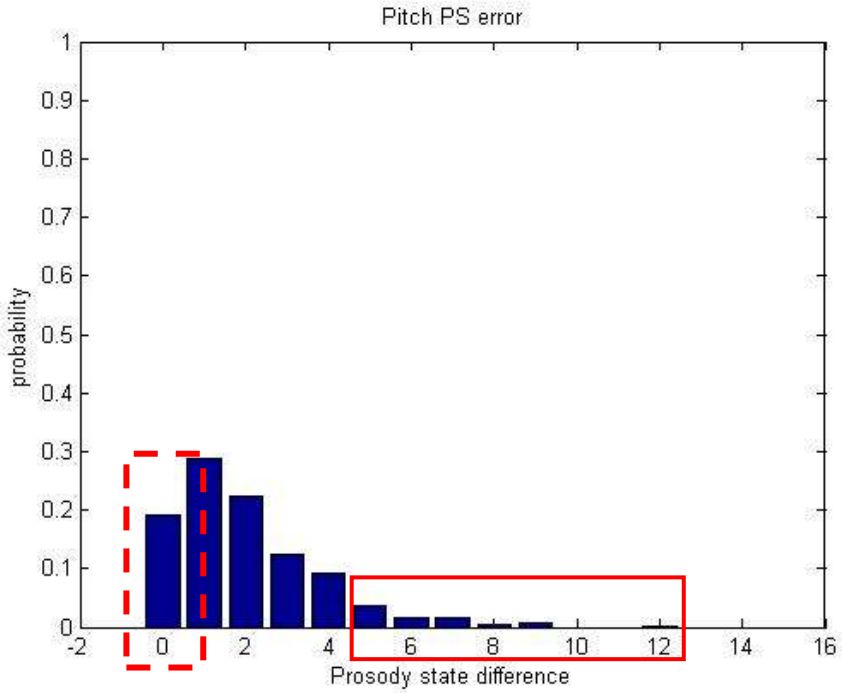


圖 4.18：利用誤差矩陣統計音節基頻軌跡韻律狀態標記差異之機率分佈圖(Jimmy→ppp)

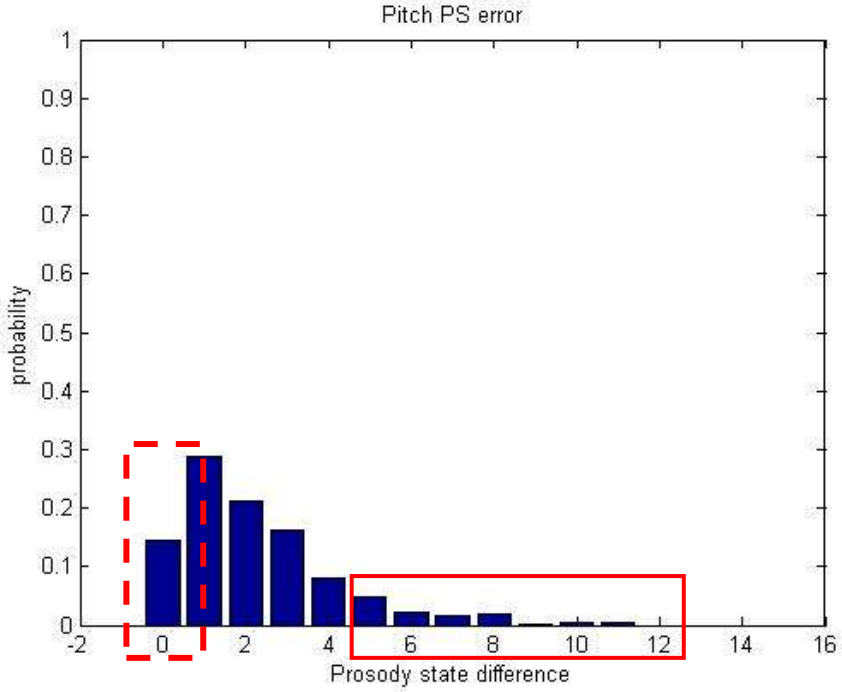


圖 4.19：利用誤差矩陣統計音節基頻軌跡韻律狀態標記差異之機率分佈圖(Jimmy→ysu)

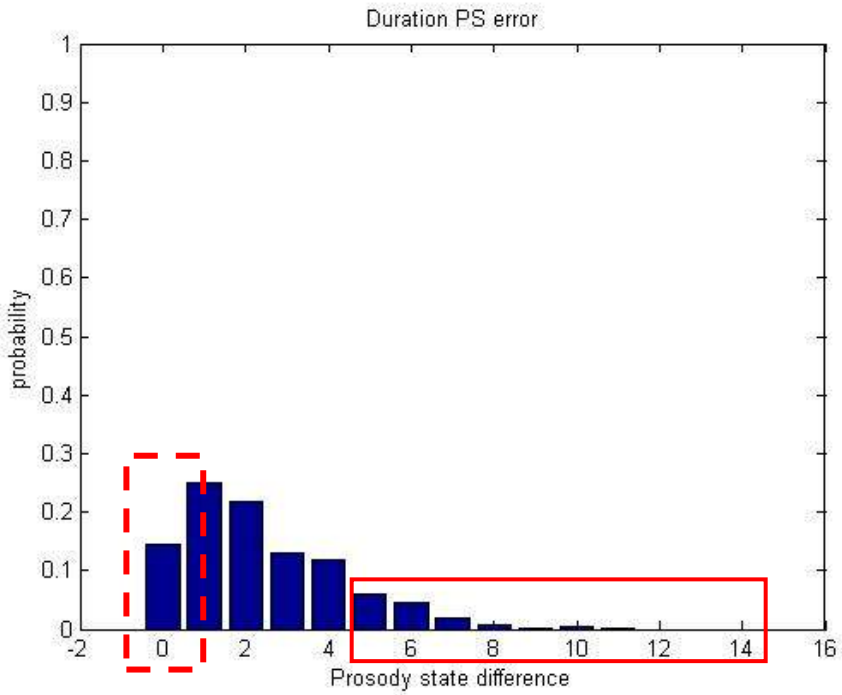


圖 4.20：利用誤差矩陣統計音節長度韻律狀態標記差異之機率分佈圖(Arron→Jimmy)

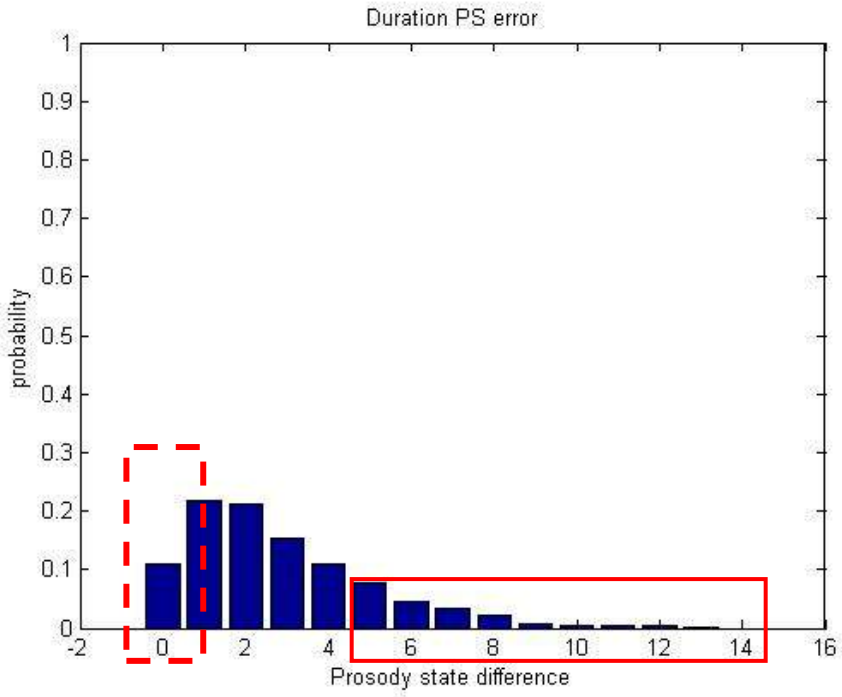


圖 4.21：利用誤差矩陣統計音節長度韻律狀態標記差異之機率分佈圖(Arron→rebecca)

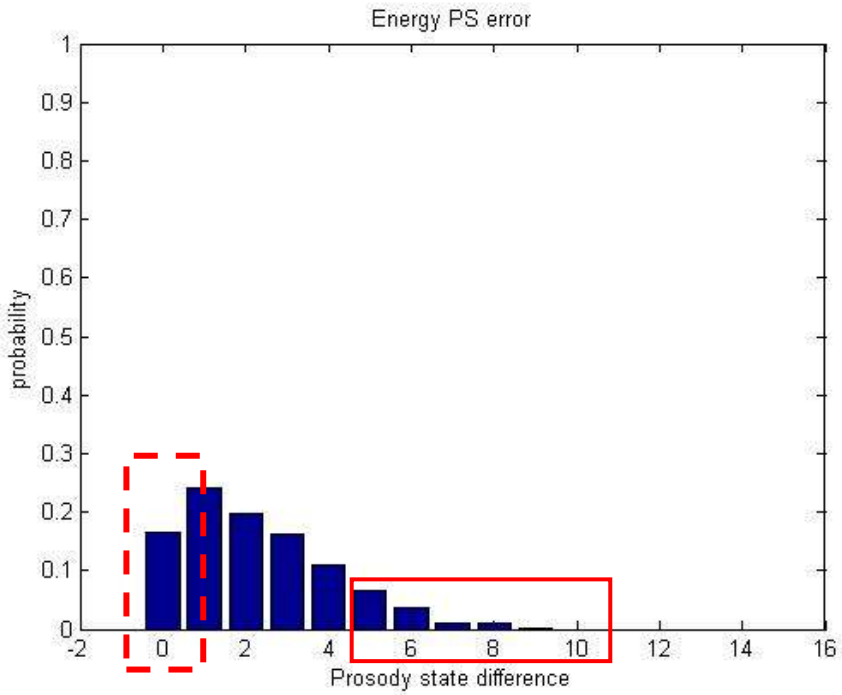


圖 4.22：利用誤差矩陣統計音節能量韻律狀態標記差異之機率分佈圖(byetwo→sung)

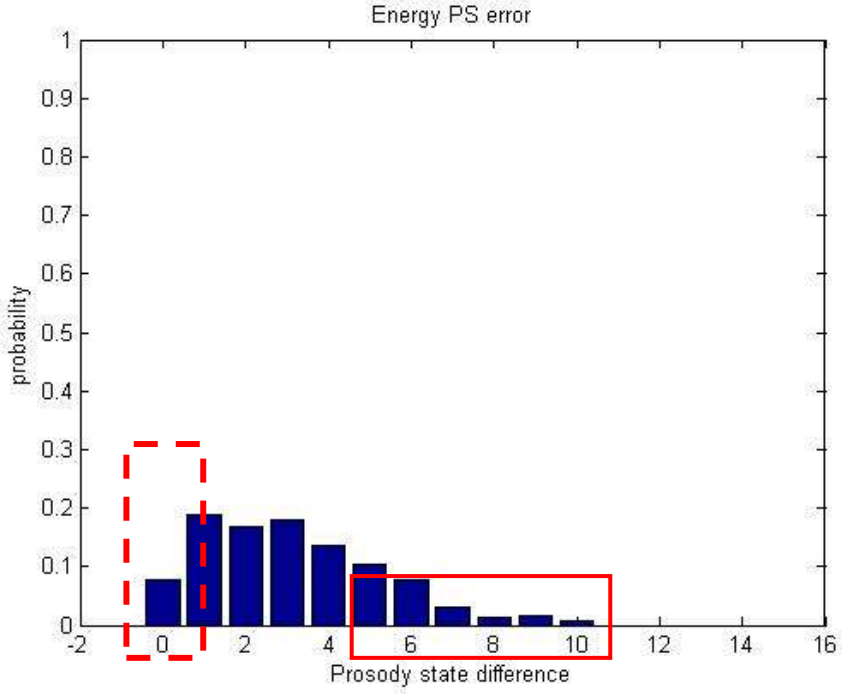


圖 4.23：利用誤差矩陣統計音節能量韻律狀態標記差異之機率分佈圖(byetwo→normal)

上述討論顯示，Source 語者與 Target 語者韻律狀態標記不一致時，確實影響我們轉換的效果；但是，若 Source 語者與 Target 語者的韻律狀態 AP 有大的差異時，如表 4.11 的第一個韻律狀態 AP，再從圖 4.24 的實線框框當中，我們可以看到轉換結果有被拉回的現象，此效果若使用高斯正規化轉換方法則無法收到效益。

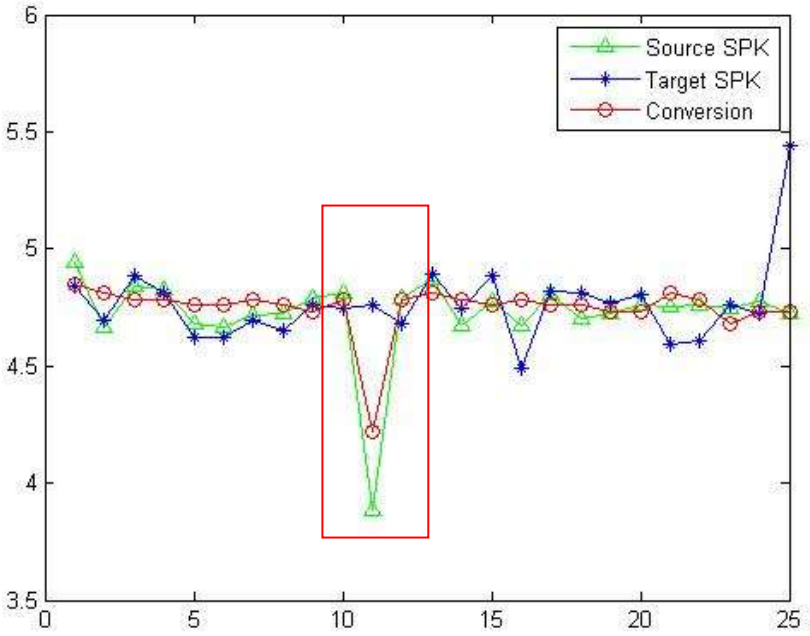


圖 4.24：Source 語者(Jimmy)對 Target 語者(Paul)使用(mean + prosodic state)做音節基頻軌跡轉換的結果

表 4.11：音節基頻軌跡的韻律狀態 AP

PS	Jimmy	Paul
1	-0.9049	-0.7761
2	-0.5448	-0.5293
3	-0.3834	-0.3883
4	-0.2855	-0.2833
5	-0.2037	-0.2028
6	-0.1328	-0.1291

最後，我們探討 major PM 發生時及 major PM 發生前後音節基頻軌跡、音節長度及音節能量韻律狀態標記的差異。

在音節基頻軌跡部分，圖 4.25 及圖 4.26 為 Jimmy→kook 及 Jimmy→daniel 利用誤差矩陣統計音節基頻軌跡在 major PM 下韻律狀態標記差異之機率分佈圖，並且我們同時觀察 major PM 發生時及 major PM 發生前後各一個音節的韻律狀態標記差異，其中 Jimmy→kook 的 NMSE 相較 Jimmy→daniel 來的低；從圖中我們發現在 major PM 發生時及 major PM 發生前後各一個音節，Jimmy→kook 的韻律狀態標記差異皆比 Jimmy→daniel 來的小，並且 Jimmy 和 kook 在 major PM 發生後的音高重置現象較 daniel 來的一致。

在音節長度部分，圖 4.27 及圖 4.28 為 Arron→Jimmy 及 Arron→rebecca 利用誤差矩陣統計音節長度在 major PM 下韻律狀態標記差異之機率分佈圖，其中 Arron→Jimmy 的 NMSE 相較 Arron→rebecca 來的低；從圖中我們發現在 major PM 發生時及 major PM 發生前後各一個音節，Arron→Jimmy 的韻律狀態標記差異皆比 Arron→rebecca 來的小，並且 Arron 和 Jimmy 在 major PM 發生前的尾音拉長現象較 rebecca 來的一致。

在音節能量部分，圖 4.29 及圖 4.30 為 byetwo→sung 及 byetwo→normal 利用誤差矩陣統計音節能量在 major PM 下韻律狀態標記差異之機率分佈圖，其中 byetwo→sung 的 NMSE 相較 byetwo→normal 來的低；從圖中我們發現在 major PM 發生時及 major PM 發生前一個音節，byetwo→sung 的韻律狀態標記差異皆比 byetwo→normal 來的小，但在 major PM 發生後一個音節，則是 byetwo→normal 的韻律狀態標記差異比 byetwo→sung 來的小，代表 byetwo 和 normal 的能量重置現象較 sung 來的一致。

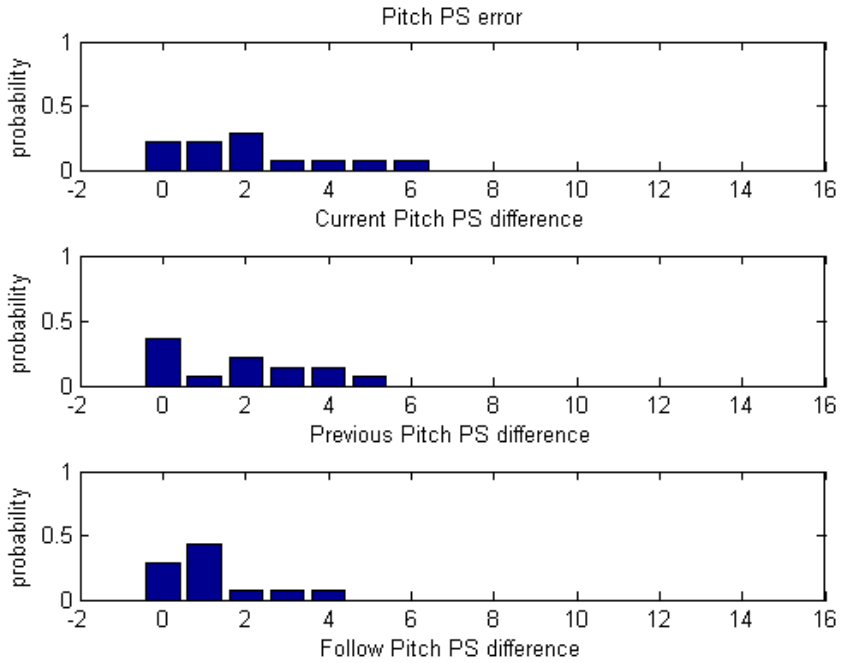


圖 4.25：利用誤差矩陣統計音節基頻軌跡在 major PM 下韻律狀態標記差異之機率分佈圖 (Jimmy→kook)(由上至下分別是 major PM 發生時、major PM 發生前及 major PM 發生後一個音節)

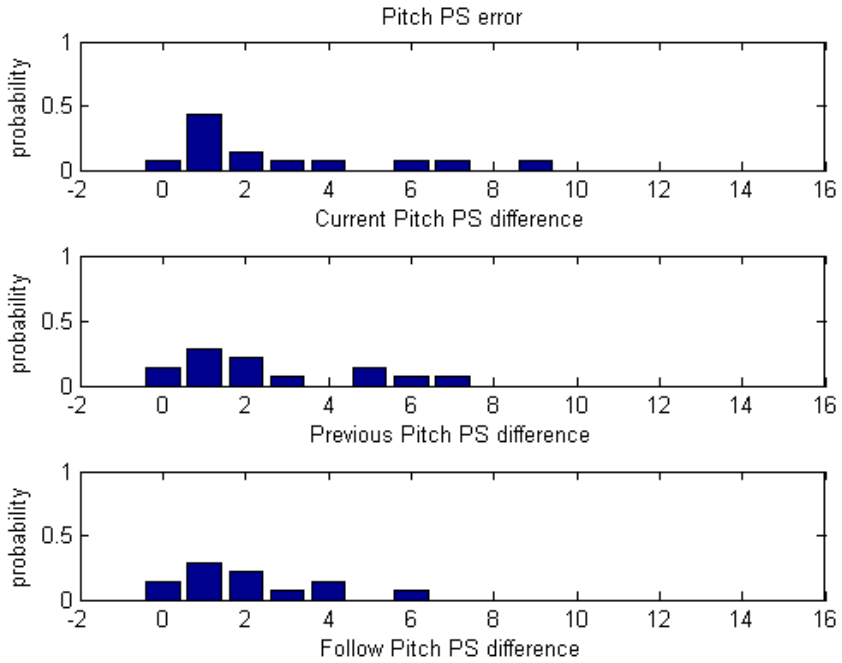


圖 4.26：利用誤差矩陣統計音節基頻軌跡在 major PM 下韻律狀態標記差異之機率分佈圖 (Jimmy→daniel)(由上至下分別是 major PM 發生時、major PM 發生前及 major PM 發生後一個音節)

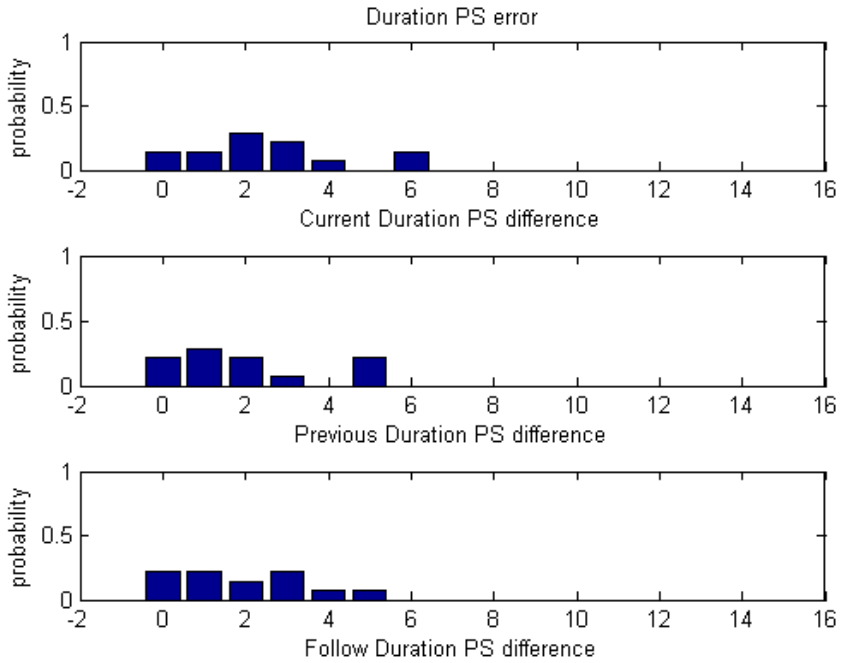


圖 4.27：利用誤差矩陣統計音節長度在 major PM 下韻律狀態標記差異之機率分佈圖 (Arron→Jimmy)(由上至下分別是 major PM 發生時、major PM 發生前及 major PM 發生後一個音節)

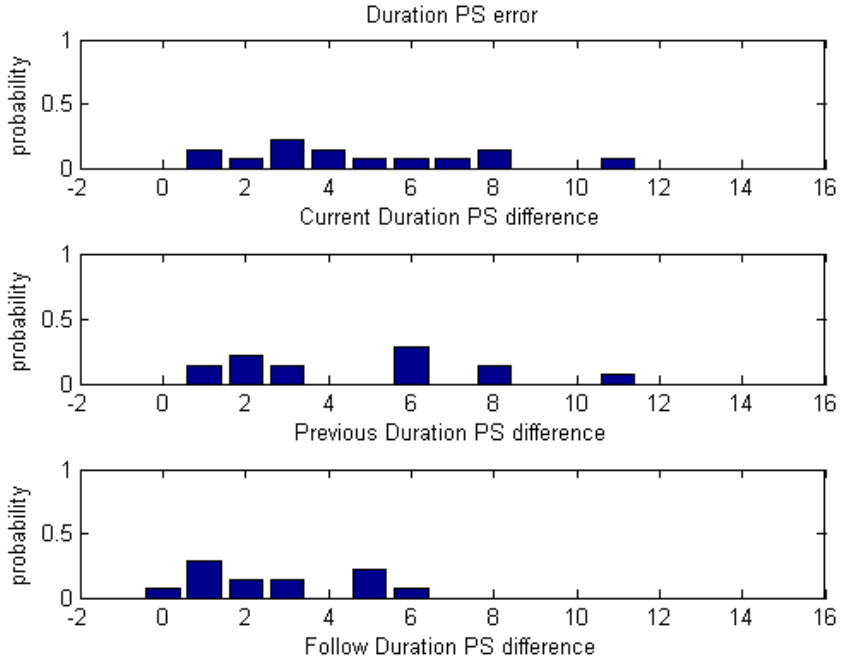


圖 4.28：利用誤差矩陣統計音節長度在 major PM 下韻律狀態標記差異之機率分佈圖 (Arron→rebecca)(由上至下分別是 major PM 發生時、major PM 發生前及 major PM 發生後一個音節)

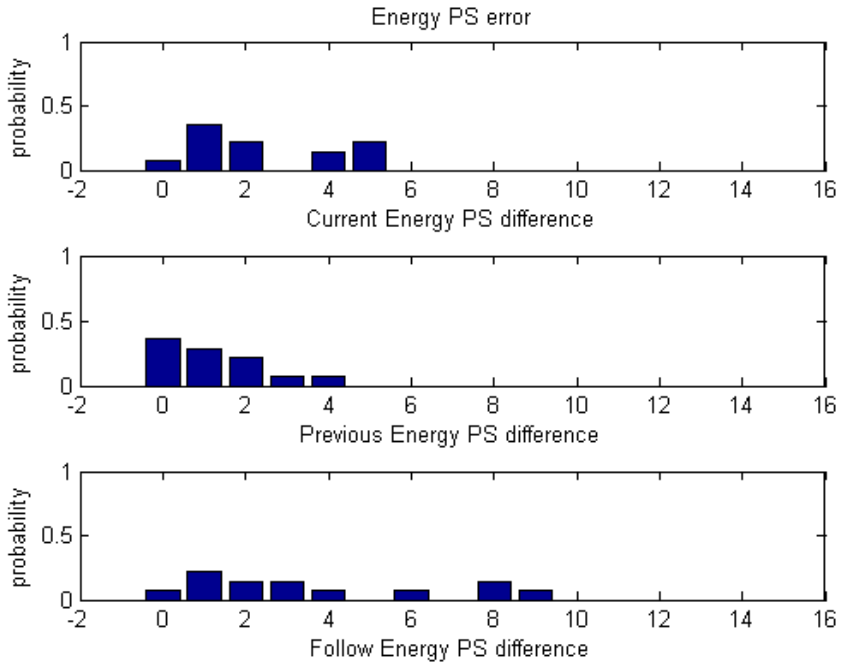


圖 4.29：利用誤差矩陣統計音節能量在 major PM 下韻律狀態標記差異之機率分佈圖 (byetwo→sung)(由上至下分別是 major PM 發生時、major PM 發生前及 major PM 發生後一個音節)

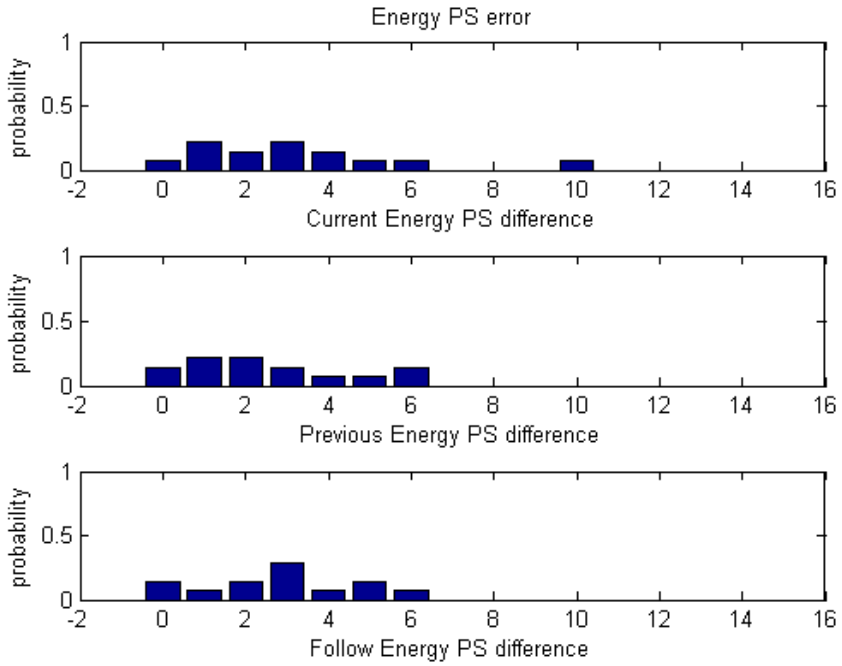


圖 4.30：利用誤差矩陣統計音節能量在 major PM 下韻律狀態標記差異之機率分佈圖 (byetwo→normal)(由上至下分別是 major PM 發生時、major PM 發生前及 major PM 發生後一個音節)

第五章 結論與未來展望

5.1 結論

本論文利用 PLM 演算法建構多語者漢語韻律模型及自動完成停頓及韻律標記，並藉此模型提出利用 MAP 調適法則的語者轉換方法；有別於傳統的韻律轉換方法，我們使用韻律模型將韻律參數拆解成各個影響因素，並利用 MAP 調適法則調適影響因素，建立個別語者的語者相關韻律模型。

從客觀性評估結果可以得到我們所提出的方法在音節基頻軌跡、音節長度及音節能量的轉換都較傳統高斯正規化轉換方法來的有效，並且在 Source 語者與 Target 語者的韻律狀態 AP 有大的差異時，使用 MAP 調適法則轉換方法可以得到有效的轉換結果，而高斯正規化法只能對整體平均值及標準差做轉換，無法對單一音節的誤差做補償。

5.2 未來展望

由於不同語者先天上皆具有獨特的說話特性，例如：語速、抑揚頓挫變化程度、唸同一詞組其音高、音長、能量皆有可能有所差距……等等；本論文提出的利用韻律模型及 MAP 調適法則做語者轉換方式，可以反映出部分語者說話特性的不同，但仍有不足之處。未來可以針對語者說話特性做進一步的研究探討，例如：使用 SR-HPM[18]建立帶有語速的韻律模型，使其轉換更為有效。

參考文獻

[1] Y. Stylianou, O. Cappe, and E. Moulines, “Continuous Probabilistic Transform for Voice Conversion,” IEEE Trans. On Speech and Audio Processing, vol. 6, no.2, pp.131-142, Mar. 1998.

[2] C. C. Hsia, C. H. Wu and J. Q. Wu, “Conversion Function Clustering and Selection Using Linguistic and Spectral Information for Emotional Voice Conversion,” IEEE Trans. Computers, 56(9):1225-1254, 2007.

[3] H. Duxans, A. Bonafonte, A. Kain and J. van Santen, “Including Dynamic and Phonetic Information in Voice Conversion Systems,” in Proc. Of ICSLP 2004, pp.5-8, Jeju Island, South Korea, 2004.

[4] O. Türk, O.Büyük, A. Haznedaroglu and L. M. Arslan, “Application of Voice Conversion for Cross-Language Rap Singing Transformation,” in Proc. Of ICASSP, pp.3597-3600, Taipei, Taiwan, April 2009.

[5] K. Y. Park and H. S. Kim, “Narrowband to wideband conversion of speech using GMM based transformation,” in Proc. ICASSP, Istanbul, Turkey, Jun. 2000, pp.1847-1850.

[6] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari and K. Shikano, “GMM-based Voice Conversion Applied to Emotional Speech Synthesis,” in Proc. Of EUROSPEECH’03, pp.2401-2404, Geneva, Switzerland, 2003.

[7] J. Tao, Y. Kang and A. Li., “Prosody Conversion from Neutral Speech to Emotional Speech,” IEEE Trans. Audio, Speech and Language Processing, Vol. 14, No.4, pp.1145-1154, July 2006.

[8] M. Abe, S. Nakamura, K. Shikano and H. Kuwabara, “Voice conversion through vector quantization,” in Proc. ICASSP, Istanbul, Turkey, Jun. 2000, pp.1847-1850.

[9] A. Kain and M. W. Macon, “Spectral Voice Conversion for Text-to-Speech Synthesis,” in Proc. Of ICASSP, vol.1, pp.285-288, Seattle, Washington, USA, May 1998.

[10] C. H. Wu, C. C. Hsia, T. H. Liu, and J. F. Wang, " Voice Conversion Using Duration-Embedded Bi-HMMs for Expressive Speech Synthesis," IEEE Trans. Audio, Speech, and Language Processing, Vol.14, No.4, pp.1109-1116, July, 2006.

[11] C. Y. Tseng, "Corpus Phonetic Investigations of Discourse Prosody and Higher Level Information," LANGUAGE AND LINGUISTICS, Institute of Linguistics, Vol.9, No.3, 2008.

[12] Chen-Yu Chiang, Sin-Horng Chen, Hsiu-Min and Yu, Yih-Ru Wang, "Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech," J. Acoust. Soc. Am., vol. 125, No. 2, pp. 1164-1183, Feb, 2009.

[13] Vassilios V. Digalakis and Leonardo G. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," IEEE Trans. Audio, Speech and Language Processing, Vol. 4, No.4, pp.294-300, July 1996.

[14] Bo-Shu Wu and Jen-Tzung Chien, "Robust Speech Recognition Using Discriminative Prior Statistics," Department of Computer Science and Information Engineering National Cheng Kung University, June 28, 2006.

[15] Yi-Chaio Wu and Yih-Ru Wang, "Speaker Recognition System for Intelligent Home Robot," Department of Communication Engineering, NCTU, August, 2011.

[16] C.-Y. Tseng, S.-H. Pin, Y.-L. Lee, H.-M. Wang, and Y.-C. Chen, "Fluent speech prosody: Framework and modeling," Speech Commun. special issue on quantitative prosody modeling for natural speech description and generation, 46, 284–309, 2005.

[17] S.-H. Chen and Y.-R. Wang, "Vector quantization of pitch information in Mandarin speech," IEEE Trans. Commun., vol. 38, no. 9, pp. 1317-1320, Sept. 1990.

[18] Chiao-Hua Hsieh, Sin-Horng Chen and Yih-Ru Wang, "A Modeling of Speaking Rate Influences on Mandarin Speech Prosody and its Application to TTS" Department of Communication Engineering, NCTU, July, 2012.