國 立 交 通 大 學

電信工程學系

碩士論文

以韻律模型為基礎之中文韻律轉換研究

A Study on Model-based Prosody Conversion for Mandarin Chinese

研 究 生：宋柏毅

指導教授：陳信宏 博士

中華民國九十八年七月

以韻律模型為基礎之中文韻律轉換研究

A Study on Model-based Prosody Conversion for Mandarin Chinese

研 究 生：宋柏毅　　Student：Po-Yi Sung

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

國 立 交 通 大 學

電 信 工 程 學 系

碩 士 論 文

A Thesis

Submitted to Department 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 2009

Hsinchu, Taiwan, Republic of China

中華民國九十八年七月

以韻律模型為基礎之中文韻律轉換研究

研 究 生：宋柏毅　　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

本研究提出以韻律模型為基礎的中文韻律轉換方法，其系統架構可分為訓練以及轉換部份。在訓練部份，先以 A-PLM 演算法分別對來源以及目標語料標示韻律標記並建立韻律模型，接著建立彼此韻律標記上的轉換關係。本論文提出兩種轉換方法，在方法一中以線性轉換的方式預估目標韻律狀態，此方法不需特別用到平行語料；而在方法二中，以 MMSE(Minimum Mean Square Error)原則，建立來源與目標韻律標記的轉換關係，它需使用平行語料。在轉換部份，首先以 A-PLM 演算法標記欲轉換的語句，即可將得到的標記資訊透過轉換函式，預估目標語者的韻律標記；最後，藉由預估得到的目標語者標記資訊以及目標韻律模型還原音節基頻軌跡、音節長度以及音節能量位階，並利用目標語音原始之頻譜參數，以 STRAIGHT 合成器合成轉換之聲音。實驗結果證實，本論文所提出之方法在中央研究院 COSPRO 語料庫上轉換效果優於傳統轉換方法。以平行語料為基礎的方法中，方法二之轉換效果在不同轉換組別皆優於以高斯混合模型為基礎之轉換，而以非平行語料為基礎所推導的方法中，方法一則優於高斯正規化轉換。

A Study on Model-based Prosody Conversion for Mandarin Chinese

Student：Po-Yi Sung　　Advisor：Dr. Sin-Horng Chen

Department of Communication Engineering
National Chiao Tung University

Abstract

In this thesis, a novel model-based prosody conversion method for Mandarin speech is presented. In the training phase, the source and target speech datasets are first analyzed by the A-PLM method to label all utterances with prosody tags and to construct their own prosodic models; then, a mapping function is built to relate the prosodic phrase structure of the two speakers. Two schemes of building mapping function are proposed. Scheme 1 builds a linear mapping function to relate the source and target prosodic states. No parallel training datasets are needed. Scheme 2 builds a probabilistic mapping function to relate the source and target prosody tags. A set of parallel data is required to train the mapping function. In the conversion phase, the source utterance is first analyzed by the A-PLM method. The labeled prosody tags are then converted to the target prosody tags by the mapping function. The transformed syllable pitch contour, duration and energy level is lastly generated by the target prosodic model. Experimental results on the Sinica COSPRO corpus confirmed that the proposed method performed very well. The two proposed schemes outperformed the conventional methods of mean/variance transformation and GMM-based mapping conversion, respectively, for the cases without and with using parallel data.

致謝

首先，要感謝研究所這兩年來指導我，對我諄諄教誨的陳信宏老師以及王逸如老師。謝謝陳老師，因為您在研究上的指點以及提醒，使我的研究進度一直都很順利；感謝王老師，跟我們 meeting 時能不厭其煩的幫我們找錯誤，使我能用不同的角度分析思考事情。

接著要感謝的是一起為畢業打拼的痞子德，還好你每次在抽菸跟洗澡時都能想出不錯的研究方向，不然我們現在不知道會在幹麻；也感謝性獸，提供我許多的協助，使我在研究上能夠很順利的解決許多問題；謝謝楊智合、希群、巴金叔叔，在我們學弟對研究感到迷惘時給我們鼓勵；還有常常嚇我的輝哥，因為你的不斷提醒，讓我知道畢業沒那麼容易；博學多聞又很會打官腔的 Q 哥，總是能提供我許多不同的意見，雖然有時候很敷衍；常常帶我們去吃好料的普烏，讓我知道了很多好玩的景點；最佳新好男人小帥哥，你真的是好男人的代表，刻苦耐勞對女朋友又體貼，我真是該多跟你學習學習；常常陪我們熬夜的宥余，很少看到像你這麼認真的人了，有時候真搞不清楚到底你是碩二還是我是碩二；感謝帥哥承燁，在口試前還幫我合成聲音；常常陪我丟球的皓翔，記得要練強一點，不然都說我欺負你；天天都在跟妹聊天的撲馬，快點介紹給我認識吧！希臘人小卡，你程式那麼強，研究一定沒問題的；還有實驗室唯二的女生，jolin 跟雲舒，實驗室真的很久沒女生了，希望你們能成為典範，讓老師以後還會收女生。在這裡也要特別感謝一下的是胖胖，不厭其煩的幫我跑實驗數據，接下來就交給你啦，相信你一定可以做的比我更好的。

最後，要感謝的是我的爸媽還有我哥，謝謝你們對我的信任，以及在我失意時，不斷的鼓勵我，如果沒有這份信任以及背後的支持，我可能沒辦法這樣無後顧之憂的完成我的學業。

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究動機

人與人之間的溝通交流都是透過語言以及文字來進行，而在人與機器之間，卻還是必須要透過鍵盤以及滑鼠來進行操作。因此，若能夠使機器像人一般可以理解人類的語言，以及能夠發出像人類的聲音、說話的韻律，並以此與人溝通，則能夠使得操作機器時更人性化。語音合成系統即是為了能讓機器發出如同人類說話的聲音進而發展出的技術。雖然目前已經可以合成出品質頗佳的聲音，但是要表現出完全像是人類說話的韻律，以及抑揚頓挫等特性，卻仍處於積極研究的階段。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，使電腦能任意地轉換成不同語者說話韻律之特性，將能夠使合成的聲音更富有多樣性。

語言學家發現，語音的韻律結構是呈階層式的架構，同樣一段文字會隨著語者說話方式的不同，而有不同的斷句時機跟停頓時間長短等，尤其在唸愈長的語句中，語者之間說話方式的差異會愈明顯。因此本論文之研究動機即是以過去所提出的韻律模型，將來源語者(source speaker)與目標語者(target speaker)的韻律表現拆解成若干個影響因素，建立影響因素之間的對應關係，進而達到不同語者之間說話韻律以及特性之轉換。

1.2 文獻回顧

聲音轉換的技術最常用在語者聲音轉換上[1]，期望能將來源語者的聲音經由轉換後聽起來像是目標語者，為了達成此目的，過去的做法主要從頻譜轉換以及韻律轉換兩方面著手；除了上述的應用外，近年來，聲音轉換的技術也應用於以資料庫為基礎的文字轉語音系統(corpus-based text-to-speech system)之後端，以便於將合成的語音轉換成目標語者之語音[2,3]，藉由此項技術，當需要合成不同語者的聲音，則不用重新錄製大量新語者之語料，僅需要相對少量的訓練語料，用來建立轉換函式，即可合成新語者之聲音。此外，聲音轉換的

技術也應用在其它方面，例如將中性情緒的語音轉換成目標情緒語音[4, 5]，歌唱聲音的轉換[6]，以及以窄頻訊號預估寬頻訊號[7]。

聲音轉換技術主要可以分為頻譜轉換以及韻律轉換，圖 1-1 為聲音轉換之基本架構圖。如圖所示，主要分成訓練階段(training phase)與轉換階段(conversion phase)兩部份；在訓練階段，傳統上是先由來源語者與目標語者錄製平行語料，亦即來源與目標語者錄製相同的文本；接著，求取韻律以及頻譜有關的特徵向量。由於來源與目標語者所錄製的語句長度不同，必須透過演算法，如動態時軸校準(Dynamic Time Wrapping, DTW)來建立來源與目標語者特徵向量的對應關係，依據此對應關係訓練轉換函式(conversion function)以進行轉換。在轉換階段，則是先將輸入的來源語音信號抽取出特徵向量，並將此特徵向量經由轉換函式進行轉換得到估測之目標特徵向量，最後將估測之目標特徵向量藉由語音合成器合成聲音。

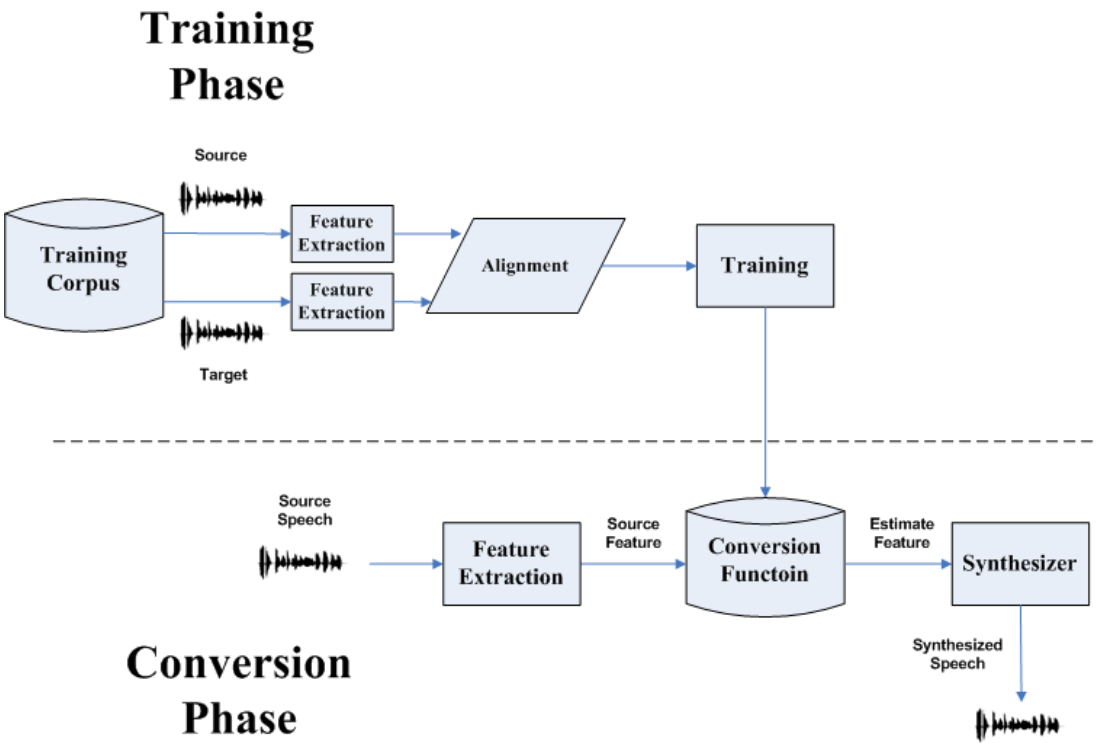

圖 1-1：聲音轉換基本架構圖

對聲音轉換而言，最重要的核心技術就是建立來源與目標語者之間的轉換函式，過去已有許多的方法被提出，主要的做法可以分為以下幾種：

- 以向量量化(Vector Quantization)為基礎之轉換

此方法最早由 M. Abe 等人[8]所提出，作法為先利用 DTW 演算法對來源與目標語者之訓練語料做校準，同時也分別對來源訓練語料與目標訓練語料做向量量化並建立碼本(codebook)，如此則可利用校準之結果找出來源碼字(codeword)與目標碼字之對應關係，並統計每個碼字對應所佔之權重；在轉換部分，先查詢來源碼本中哪個碼字與來源聲音之特徵參數最相近，即可利用查詢到之碼字以及其與目標碼字之對應權重，以線性組合建立轉換。此方法雖然簡單，但轉換之結果僅為碼字之線性組合，故為一不連續之轉換，聲音品質也不佳。

- 以高斯混合模型(Gaussian Mixture Model, GMM)為基礎之轉換

以 GMM 為基礎之轉換是由 Y. Stylianou 等人[1]所提出，之後有許多研究[9]都是以此方法為基礎對其做進一步的改進。其中，又以 A. Kain 等人[10]所提出的改進方式為目前聲音轉換最常被引用。其基本想法為利用高斯混合模型描繪來源與目標語者之特徵參數，並以此建立轉換函式。因為高斯混合模型為連續機率密度函數，故此方法解決了以向量量化為基礎之轉換不連續性問題。此一做法將於第三章做詳細之介紹。

- 以隱藏式馬可夫模型(Hidden Markov Model, HMM)為基礎之轉換

高斯混合模型雖然解決了轉換上不連續的問題，但是並沒有考慮到語音信號在時間上之相關性，因此 H. Duxans 等人[3]以及 C. H. Wu 等人[11]提出了以 HMM 為基礎之轉換，藉由聲音信號在 HMM 狀態上之變換，而以不同轉換函式進行轉換，以解決 GMM 對於時間獨立(time independent)之假設。

- 以分類迴歸樹(Classification and Regression Tree, CART)為基礎之轉換

除了最基本的語音特徵參數之外，語音學上的資訊(phonetic information)也為聲音的重要特性，例如聲母(initial)、韻母(final)以及聲調(tone)等。因此 H. Duxans[3]提出了以 CART 為基礎之轉換，在訓練轉換函式時除了基本的語音特徵參數外，進一步考慮到每個音框上的語音學資訊，並依設定的問題集以 CART 演算法分類，最後在

每一葉節點建立轉換函式，形成多轉換函式之轉換方法。

韻律轉換之目標是將來源語者之韻律參數，例如基頻值(pitch value)或基頻軌跡(pitch contour)、音長(duration)、停頓(pause)，以及能量位階(energy level)轉換成目標語者之韻律參數。在早期聲音轉換之研究上，主要以探討頻譜為主，對於韻律的轉換，則是以簡單的方法，如線性轉換來呈現之；而在探討韻律轉換之研究上，絕大多數仍是以基頻為主；其中，最普遍的是以高斯正規化的方式，找出線性對應的函式對來源語者之基頻值做轉換，此方法的優點在於簡單易於實現且不需要用到平行語料。近年來，開始有學者們嘗試用不同的方法於韻律轉換之研究，例如以高斯混合模型為基礎於基頻轉換[5]。

對於基頻轉換，過去的作法[12, 13]主要是以音框為單位，亦即對基頻值做轉換。此作法的缺點在於並未考量音框與音框之間時間上之關聯性，尤其影響以聲調語系為主的語言如中文。由於中文是有聲調的語言，而聲調帶有語意上之資訊。聲調主要特徵來自於音節中的基頻軌跡，因此基頻軌跡在轉換時若發生錯誤，則可能導致最後語意方面的誤解。故在一些文獻中，開始有以音段(segmental)或超音段(suprasegmental)為轉換單元[5, 14-17]，而不是單一基頻值。在[18]中，就先以time normalization以及moving average filter對音節基頻軌跡作量化，轉換方式則是基於音調碼本方法(tone codebook mapping method)做轉換。此外，[5]則是以pitch target model[19]對音節基頻軌跡作量化，再以GMM與CART為基礎的轉換方式對此特徵向量作轉換。本研究則是以正交化展開之三階係數描述音節的基頻軌跡，再對此係數作轉換。

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[20, 21]，因此近年來學者開始運用韻律以及語言學上的知識於基頻轉換。吳宗憲等人首先提出一簡化的階層式架構，將來源與目標語者的基頻軌跡由上層至下層拆解為句子(sentence)、詞(word)與次音節(sub-syllable)三個階層，對各階層間建立轉換函式，轉換時則以此不同階層的轉換函式進行轉換[14]。

1.3 研究方向

本論文以江振宇的韻律模型[22]為基礎，提出中文韻律轉換的方法。首先在訓練部份，利用A-PLM(Advanced unsupervised joint Prosody Labeling and Modeling)演算法分別對來源與目標語者之所有訓練語料做標記並建立其韻律模型，此韻律模型建立了語者的音節基頻軌跡、音長、能量位階之統計模型，而韻律標記資訊則描述了上層的韻律階層架構。

接著建立來源與目標語者在韻律標記上的對應函式；在轉換時，藉由韻律標記轉換函式以及目標語者的部份影響因素(Affecting Factor)，完成韻律轉換。最後，聲音的合成部份則是使用STRAIGHT(Speech Transformation and Representation using Adaptive Interpolation of weiGHTed Spectrum)[23]合成器進行合成；本研究主要是針對韻律轉換，因此在頻譜部份，則保有目標語音語句之頻譜參數進行合成。

1.4 章節概要說明

本論文共分為五章：

第一章　緒論：介紹本論文之研究動機與方向。

第二章　系統架構簡介：介紹本論文提出之轉換方法系統架構以及所採用之韻律模型。

第三章　以音節為基礎之韻律轉換：介紹兩種傳統的韻律轉換方法，以及本論文所提出兩種以韻律模型為基礎之轉換法。

第四章　實驗結果與分析：以客觀與主觀評估方式驗證轉換方法，並分析實驗結果。

第五章　結論與未來展望。

第二章 系統架構簡介

此章節首先介紹本論文所提出以韻律模型為基礎的轉換方法系統架構，接著介紹採用的韻律模型。

2.1 韻律轉換系統架構

圖 2-1 與圖 2-2 分別為本研究提出之韻律轉換系統架構圖之訓練階段(training phase)與轉換階段(conversion phase)。首先在訓練階段，分別對來源以及目標語料以音節為單位做切割，並藉由切割資訊抽取出韻律參數(prosodic features)，包括音節基頻軌跡、音節邊界的停頓時長(pause duration)、音節長度、能量位階以及音節邊界的 energy-dip level 等資訊；同時以文字處理器抽取出語言參數(linguistic features)，包括聲調、詞長、詞類之資訊。接著以 A-PLM 演算法，結合韻律參數以及語言參數分別訓練來源語者與目標語者各自之韻律模型，並標記韻律狀態(prosodic state)以及停頓標記(break type)。

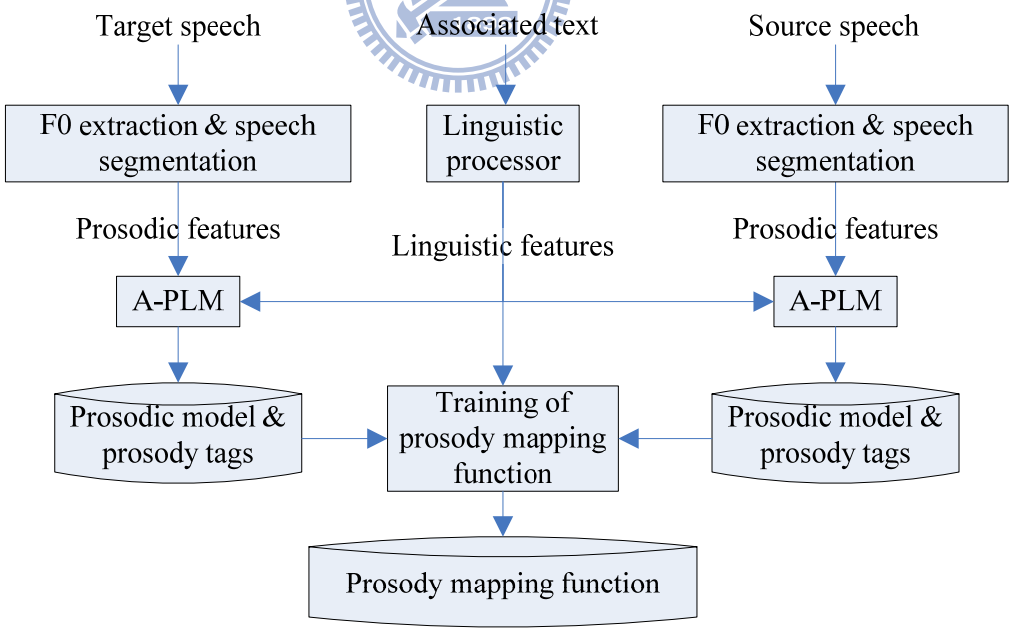

圖 2-1：訓練階段之系統架構圖

為了要使來源語者之說話韻律特性經由轉換之後能與目標語者之說話韻律特性相近，本論文採用的方法為統計來源與目標語者之間韻律狀態以及停頓標記的對應關係，以此建立韻

律標記之轉換函式(prosody mapping function)。在建立韻律轉換函式，本研究提出兩種方法，分別為轉換方法一與方法二。首先在方法一中，運用了高斯正規化的概念對來源與目標的韻律參數以及韻律狀態建立對應關係，因此，此方法並不需要特別用到平行語料，即可建立轉換函式；其次，在方法二則是以 MMSE 法則，建立來源以及目標韻律標記資訊之間的轉換關係，此該方法必需要使用平行語料。

在轉換階段，先將來源語音以音節為單位做切割，藉由切割位置抽取出韻律參數，同時將語料文字抽取出語言參數；接著利用來源語者之韻律模型對輸入語音標記其韻律狀態及停頓標記(source prosody tags)，之後即可將得到的標記資訊，透過轉換函式，得到預估的目標語者韻律標記(target prosody tags)；最後，藉由預估得到的目標語者標記資訊以及目標韻律模型還原音節基頻軌跡、音節長度以及音節能量位階，並利用目標語音原始之頻譜參數，以 STRAIGHT 合成器合成轉換之聲音。

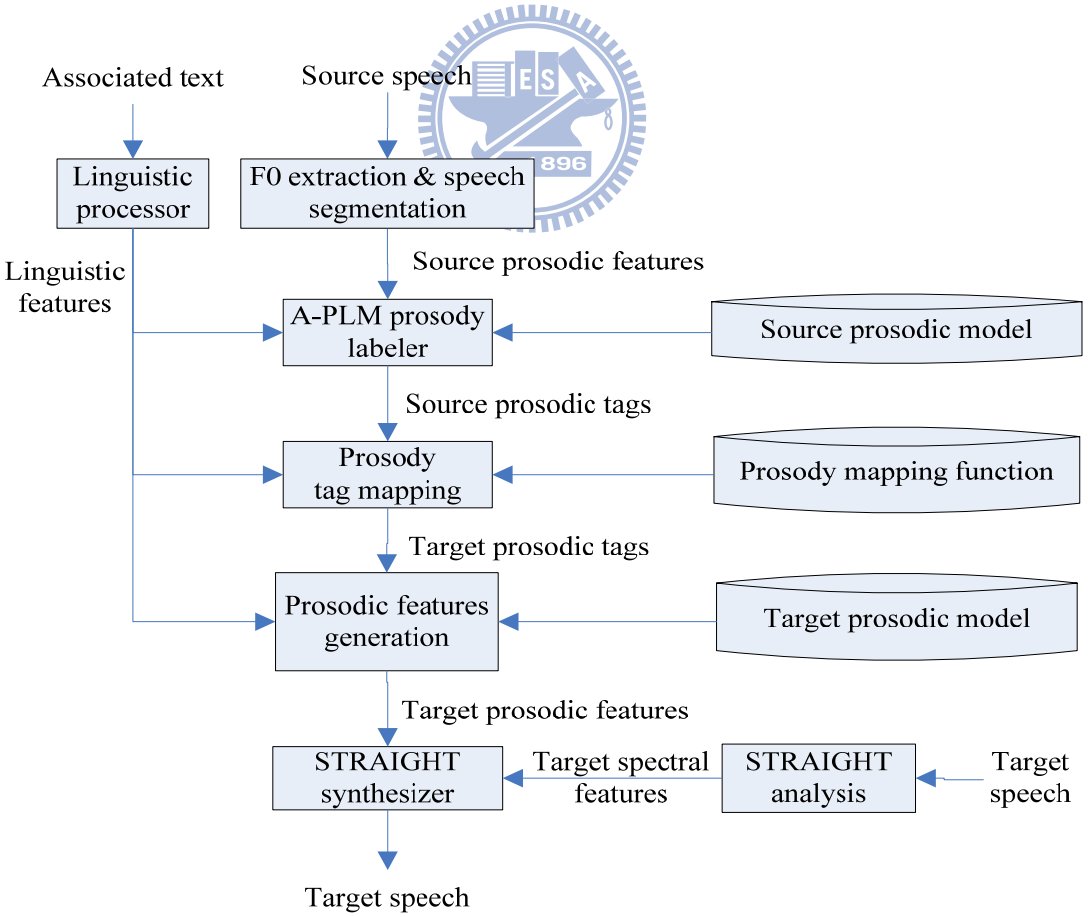

圖 2-2：轉換階段之系統架構圖

2.2 A-PLM 演算法

本論文所採用之 A-PLM 演算法可以針對一個未經人工事先標記好的語料庫，經由一連串參數最佳化的過程，同時做好韻律標記以及模型參數估測。圖 2-3 為在 A-PLM 演算法中所採用之中文韻律階層架構；此架構由四層所構成：音節(SYL)、韻律詞(PW)、韻律短語(PPh)、以及呼吸組/韻律句組(BG/PG)。

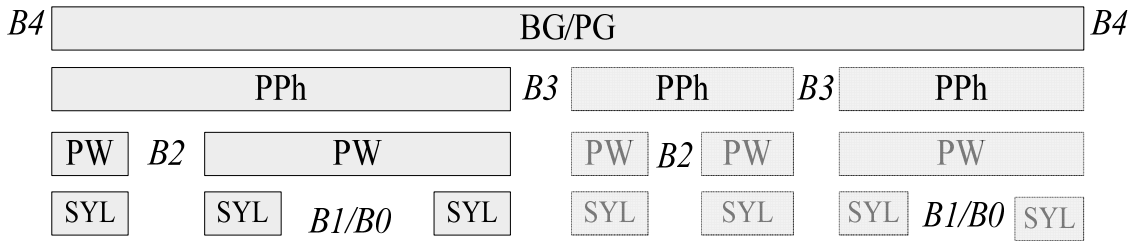

圖 2-3：A-PLM 演算法所採用之韻律階層架構

2.2.1 設計韻律模型

韻律標記問題可以視為，在給定語料庫之語音聲學參數集合 **A**，和相對應的語言參數集合 **L** 之下，要求取輸出的韻律標記集合 **T** 之最佳解，因此整個過程可以看成一個求取最佳參數解的過程，即

$$\mathbf{T}^{*}=\arg\max_{\mathbf{T}} P(\mathbf{T}|\mathbf{A},\mathbf{L})=\arg\max_{\mathbf{T}} P(\mathbf{T},\mathbf{A}|\mathbf{L}) \tag{2-1}$$

韻律標記集合包含了兩類很重要的漢語語音韻律資訊，第一類是階層韻律架構的音節邊界停頓標記(Break Type)，在本論文定義韻律邊界音節停頓標記集合 **B**={B0, B1, B2-1, B2-2, B2-3, B3, B4}，其中 B2-1、B2-2 及 B2-3 分別代表明顯音高重置(pitch reset)之韻律詞邊界、短停頓(short pause)之韻律詞邊界以及含有音節拉長效應(duration lengthening)之後的韻律詞邊界。另一類的韻律標記是音節的韻律狀態，在本方法中韻律狀態有 3 種，代表的意義分別是經過量化和正規化音節基頻韻律狀態 **p**、音長韻律狀態 **q** 和音節能量韻律狀態 **r**。正規化後的基頻會扣除掉音節層次對基頻的貢獻，即聲調和連音的影響因素會被扣掉，此時音節基頻的韻律狀態代表的是韻律詞、韻律短語、呼吸組/韻律句組對基頻的貢獻。至於音長或能

量強度則要分別扣除語句、聲調、基本音節類型或韻母類型的影響因素，使其分別表示最上面三層之韻律詞、韻律短語、呼吸組/韻律句組(PW, PPh, BG/PG)對音長和能量強度的貢獻。綜合以上，韻律標記集合**T**={**B**, **PS**}，其中**PS**={**p**, **q**, **r**}為韻律狀態標記集合。

聲學參數可分為兩類，其中一類的聲學參數和韻律狀態標記有很大的相關性，與音節邊界停頓標記的相關性很低或是獨立，屬於這類的聲學參數有音節基頻軌跡、音長和音節能量；另一類的聲學參數則用來說明音節邊界停頓標記，這類型的聲學參數和音節邊界停頓標記有很大的相關性，與韻律狀態標記的相關性很低或是獨立，屬於這類的聲學參數有音節邊界的停頓時長(pause duration)、音節邊界的 energy-dip level、正規化的能量差、正規化的基頻差(normalized pitch jump)以及正規化的音節長度拉長因子(normalized duration lengthening factor)等。根據上面的討論定義 **A** 包含音節基頻軌跡序列 **sp**、停頓時長序列 **pd**、energy-dip level 序列 **ed**、音節長度序列 **sd**、音節能量序列 **se**、正規化的音節內基頻差序列 **pj**，定義為：

$$pj_n = \left(\mathbf{sp}_{n+1}(1) - \boldsymbol{\beta}_{t_{n+1}}(1)\right) - \left(\mathbf{sp}_n(1) - \boldsymbol{\beta}_{t_n}(1)\right), \tag{2-2}$$

在此**x**(1)定義為向量 **x** 的第一維度，下標 n 表示為第 n 個音節，$\boldsymbol{\beta}_{t_n}$ 為聲調影響因素 t_n 的 affecting patterns(APs)，而正規化的音節長度拉長因子序列 **dl** 和 **df** 定義為：

$$dl_n = \left(sd_n - \gamma_{t_n} - \gamma_{s_n}\right) - \left(sd_{n\text{-}1} - \gamma_{t_{n-1}} - \gamma_{s_{n-1}}\right) \tag{2-3}$$

和

$$df_n = \left(sd_n - \gamma_{t_n} - \gamma_{s_n}\right) - \left(sd_{n+1} - \gamma_{t_{n+1}} - \gamma_{s_{n+1}}\right) \tag{2-4}$$

其中 γ_t 和 γ_s 分別表示聲調與基本音節類型影響因素在音長的 APs，因此聲學參數集合成為 **A**={**sp**,**sd**,**se**,**pd**,**ed**,**pj**,**dl**,**df**}。為了能夠更清楚的說明這些聲學參數，將 **A** 細分三個類別：音節韻律參數(Syllable Prosodic Feature) **X**={**sp**,**sd**,**se**}，音節內韻律參數(Inter-syllabic Prosodic Feature) **Y**={**pd**,**ed**}以及音節差韻律參數(Differential Prosodic Feature) **Z**={**pj**,**dl**,**df**}。

至於語言參數方面，首先用 **L** 來表示所有的語言參數集合。接下來將音節聲調、基本音節類型與韻母類型從 **L** 中獨立出來，這樣做的用意在於音節聲調、基本音節類型與韻母類型分別對音節基頻軌跡、音長與音節能量有顯著的影響。其次考慮到不同語句時，說話速度上的變動會造成音長的變化以及說話音量變動會造成能量的變化，再把兩個語句層次的正規化因子獨立出來。最後將上述這些從 **L** 中拿掉和獨立出來後剩餘的語言參數，定義為 reduced linguistic feature set **l**，為了能清楚的了解這些符號定義，將其列在表格 2-1。

表 2-1：韻律標記、韻律參數和語言參數的表示法

T : prosodic tag	**B** : break type	
	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	
	u: utterance sequence	

綜合上述之討論，可將 2-1 式改寫為

$$P(\mathbf{T},\mathbf{A}|\mathbf{L}) = P(\mathbf{A}|\mathbf{T},\mathbf{L})P(\mathbf{T}|\mathbf{L}) = P(\mathbf{X},\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{PS},\mathbf{L})P(\mathbf{B},\mathbf{PS}|\mathbf{L})$$
$$\approx P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})P(\mathbf{PS}|\mathbf{B})P(\mathbf{B}|\mathbf{L}) \quad (2\text{-}5)$$

其中 $P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})$ 稱為音節韻律模型(Syllable Prosodic Model)，$P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$ 稱為停頓聲學模型(Break-acoustic Model)，$P(\mathbf{PS}|\mathbf{B})$ 稱為韻律狀態模型(Prosodic State Model)，$P(\mathbf{B}|\mathbf{L})$ 稱為

break-syntax model。進一步將音節韻律模型 $P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})$ 分解成三個模型，分別模擬音節基頻軌跡序列 **sp**、音長序列 **sd** 和音節能量序列 **se**，並且假設 **sp**、**sd** 和 **se** 的變化在此只受到以下幾個影響因素控制：音節聲調 **t**、基本音節類型 **s**、韻母類型 **f**、語句 **u**、韻律狀態 $\mathbf{PS}=\{\mathbf{p},\mathbf{q},\mathbf{r}\}$ 和韻律邊界停頓 **B**，因此得到

$$\begin{aligned} p(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L}) &\approx p(\mathbf{sp}|\mathbf{B},\mathbf{p},\mathbf{t})\,p(\mathbf{sd}|\mathbf{q},\mathbf{t},\mathbf{s},\mathbf{u})\,p(\mathbf{se}|\mathbf{r},\mathbf{t},\mathbf{f},\mathbf{u}) \\ &\approx \prod_{n=1}^{N} p(\mathbf{sp}_n|B_{n-1}^{n},p_n,t_{n-1}^{n+1}) \prod_{n=1}^{N} p(sd_n|q_n,t_n,s_n,u_n) \prod_{n=1}^{N} p(se_n|r_n,t_n,f_n,u_n) \end{aligned} \tag{2-6}$$

先從 2-6 式的第一個模型看起，$\prod_{n=1}^{N} p(\mathbf{sp}_n|B_{n-1}^{n},p_n,t_{n-1}^{n+1})$ 是在模擬每個基頻軌跡時，假設所觀察到第 n 個音節之基頻軌跡 $\mathbf{sp}_n$ 會受到目前基頻韻律狀態 p_n、目前聲調 t_n 以及在給定韻律邊界停頓 B_{n-1} 和 B_n 時，前後各一個音節聲調 t_{n-1} 和 t_{n+1} 造成的連音影響，因此 $B_{n-1}^{n}=(B_{n-1},B_n)$，$t_{n-1}^{n+1}=(t_{n-1},t_n,t_{n+1})$。而 $\mathbf{sp}_n$ 則為第 n 個音節基頻軌跡，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，在此將 $\mathbf{sp}_n$ 寫成

$$\mathbf{sp}_n = \mathbf{sp}_n^{r} + \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\beta}^{f}_{B_{n-1},tp_{n-1}} + \boldsymbol{\beta}^{b}_{B_n,tp_n} + \boldsymbol{\mu} \quad \text{for } 1 \le n \le N \tag{2-7}$$

2-7 式的 $\boldsymbol{\beta}_x$ 表示音節基頻軌跡影響因素為 x 時的 AP，tp_n 是 tone pair $t_n^{n+1}=(t_n,t_{n+1})$，$\boldsymbol{\beta}^{f}_{B_{n-1},tp_{n-1}}$ 和 $\boldsymbol{\beta}^{b}_{B_n,tp_n}$ 分別是第 n-1 個和第 n+1 個音節所貢獻的前後音節影響效應的 APs，$\boldsymbol{\mu}$ 是 global mean 的 AP。每個語句的韻律邊界都有兩個特例，即為語句的開始與結束，分別以 B_b 和 B_e 表示之，因此 $\boldsymbol{\beta}^{f}_{B_b,t_1}=\boldsymbol{\beta}^{f}_{B_0,tp_0}$，$\boldsymbol{\beta}^{b}_{B_e,t_N}=\boldsymbol{\beta}^{b}_{B_N,tp_N}$ 為兩個特例的連音效應 APs，另外為了將韻律狀態的影響限制在目前音節的 log-F0 level，我們把 $\boldsymbol{\beta}_{p_n}$ 設定成在四維正交係數的第一維都是非零值。$\mathbf{sp}_n^{r}$ 是正規化後的 $\mathbf{sp}_n$，亦可稱為 $\mathbf{sp}_n$ 扣除 $\boldsymbol{\beta}_{t_n}$、$\boldsymbol{\beta}_{p_n}$、$\boldsymbol{\beta}^{f}_{B_{n-1},tp_{n-1}}$、$\boldsymbol{\beta}^{b}_{B_n,tp_n}$ 和 $\boldsymbol{\mu}$ 的殘餘值(residual)，圖 2-4 顯示出 $\mathbf{sp}_n$ 與這些影響因素之間的關係圖，藉由假設 $\mathbf{sp}_n^{r}$ 是一 zero-mean 的 normal distribution，即 $N(\mathbf{sp}_n^{r};\mathbf{0},\mathbf{R})$，則可以得到

$$P(\mathbf{sp}_n \mid p_n, B_{n-1}^{n}, t_{n-1}^{n+1}) = N(\mathbf{sp}_n; \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\beta}^{f}_{B_{n-1}, tp_{n-1}} + \boldsymbol{\beta}^{b}_{B_n, tp_n} + \boldsymbol{\mu}, \mathbf{R}) \quad \text{for } 1 \le n \le N \tag{2-8}$$

其中 **R** 定義為 $\mathbf{sp}_n^r$ 的共變數矩陣(covariance matrix)。

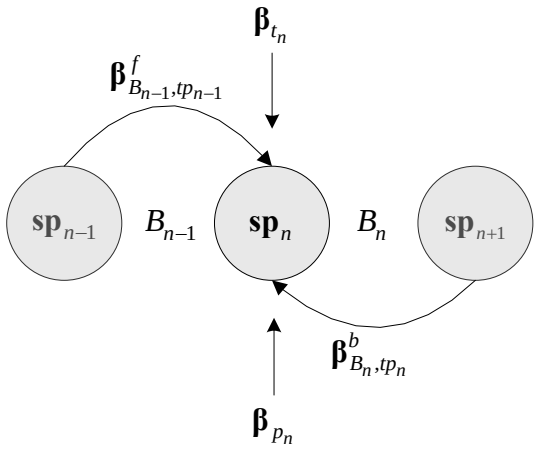

圖 2-4 觀察到的音節基頻軌跡與其影響因素的關係圖

2-6 式第二個模型：

$$P(sd_n \mid q_n, t_n, s_n, u_n) = N(sd_n; \gamma_{t_n} + \gamma_{q_n} + \gamma_{s_n} + \gamma_{u_n} + \mu_d, R_d) \tag{2-9}$$

模擬了音節長度 sd_n， μ_d 與 R_d 分別表示 global mean 與音長殘餘值的共變異數矩陣；而 2-6 式第三個模型

$$P(se_n \mid r_n, t_n, f_n, u_n) = N(se_n; \alpha_{t_n} + \alpha_{r_n} + \alpha_{f_n} + \alpha_{u_n} + \mu_e, R_e) \tag{2-10}$$

模擬了音節能量 se_n， μ_e 與 R_e 分別表示 global mean 與音節能量殘餘值的共變異數矩陣。

停頓聲學模型 $P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$ 進一步化簡如下：

$$P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L}) \approx P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{l}) \approx \prod_{n=1}^{N} P(pd_n, ed_n, pj_n, dl_n, df_n \mid B_n, \mathbf{l}_n) \tag{2-11}$$

其中 $P(pd_n, ed_n, pj_n, dl_n, df_n \mid B_n, \mathbf{l}_n)$ 是經由分類樹與決策樹(Classification and Regression Tree, CART)推導出來，其節點的分類標準是依據最大概似函數增益(Maximum Likelihood Gain)，

CART 演算法可以利用一個已經設計好的問題集，依據不同的韻律邊界停頓同時將所有音節的 pd_n、ed_n、pj_n、dl_n 和 df_n 做好分類。在此將 pd_n 以 gamma distribution 建構，而 ed_n、pj_n、dl_n 和 df_n 以 normal distribution 建構，因此 $p(pd_n,ed_n,pj_n,dl_n,df_n|B_n,\mathbf{l}_n)$ 會是一個 gamma distribution 和四個 normal distribution 的乘積。

韻律狀態模型可以進一步針對三種韻律狀態分解成三個子模型，表示為

$$P(\mathbf{PS}|\mathbf{B})\approx P(\mathbf{p}|\mathbf{B})P(\mathbf{q}|\mathbf{B})P(\mathbf{r}|\mathbf{B}) \tag{2-12}$$

而 $P(\mathbf{p}|\mathbf{B})$、$P(\mathbf{q}|\mathbf{B})$ 和 $P(\mathbf{r}|\mathbf{B})$ 可以用雙連文模型(Bigram Models)分別表示為

$$P(\mathbf{p}|\mathbf{B})\approx P(p_1)\left[\prod_{n=2}^{N}P(p_n|p_{n-1},B_{n-1})\right] \tag{2-13}$$

$$P(\mathbf{q}|\mathbf{B})\approx P(q_1)\left[\prod_{n=2}^{N}P(q_n|q_{n-1},B_{n-1})\right] \tag{2-14}$$

和

$$P(\mathbf{r}|\mathbf{B})\approx P(r_1)\left[\prod_{n=2}^{N}P(r_n|r_{n-1},B_{n-1})\right] \tag{2-15}$$

至於 break-syntax 模型 $P(\mathbf{B}|\mathbf{l})$，若能將每個音節邊界分開模擬，還可以再化簡為

$$P(\mathbf{B}|\mathbf{l})=\prod_{n=1}^{N-1}P(B_n|\mathbf{l}_n) \tag{2-16}$$

其中 $P(B_n|\mathbf{l}_n)$ 可以經由 CART 演算法得到。

2.2.2 A-PLM 法標記及訓練韻律模型

A-PLM 法在同時估測 8 個韻律模型的參數及對所有語句做韻律標記的過程中，是根據 ML 法則做一連串的最佳化程序直到收斂為止，整個演算過程分為兩部份：初始化和疊代。初始化過程會對所有語句做初始的韻律標記，以及對 2.2.1 節所討論的 8 個子模型做初始的韻律參數估測；而在疊代的過程中會先對所有語句定義一概似函數(Likelihood Function)

$$Q=\left(\prod_{n=1}^{N}P(\mathbf{sp}_n\mid p_n,B_{n\text{-}1}^{n},t_{n-1}^{n+1})\,p(sd_n\mid q_n,t_n,s_n,u_n)\,p(se_n\mid r_n,t_n,f_n,u_n)\right)$$
$$\left(P(p_1)P(q_1)P(r_1)\prod_{n=2}^{N}P(p_n\mid p_{n-1},B_{n-1})P(q_n\mid q_{n-1},B_{n-1})P(r_n\mid r_{n-1},B_{n-1})\right) \tag{2-17}$$
$$\left(\prod_{n=1}^{N-1}\left(p(pd_n,ed_n,pj_n,dl_n,df_n\mid B_n,\mathbf{l}_n)P(B_n\mid\mathbf{l}_n)\right)\right)$$

接著利用一個多重步驟的疊代程序，反覆更新所有的韻律標記和 8 個韻律子模型的參數，細節可參考[22]。

第三章 以音節為基礎之韻律轉換

本論文將會對音節之基頻軌跡、長度以及能量作轉換，故在本章中，首先將會介紹對音節基頻軌跡量化之演算法，接著詳細介紹兩種以音節為基礎的韻律轉換方法，並以此作為與本論文提出方法的比較基準，最後提出以 A-PLM 產生之韻律模型為基礎的兩種韻律轉換方法。

3.1 基頻軌跡量化

為了呈現以音節為基本單元之基頻轉換，也就是對音節的基頻軌跡做轉換，在本論文以正交化展開之三階係數[24]，來描述音節之基頻軌跡變化曲線。之所以選用正交化係數作為韻律轉換的參數主要是因為以下兩點原因：首先，在過去的研究中[21, 24]，正交化係數已經成功的用來描述音節的基頻軌跡；此外，在本研究，實驗也證實了正交化係數的確適合做為音節基頻軌跡的轉換。正交化係數展開的數學式如下：

$$a_j = \frac{1}{N+1}\sum_{i=0}^{N} f(i)\cdot \Phi_j(\frac{i}{N}) \tag{3-1}$$

其中，a_j 代表第 j 階的四維正交化參數；$f(i)$，$0 \le i \le N$，表示以音框為單位之原始基頻軌跡，$N+1$ 為音節基頻軌跡的長度；$\Phi_j(\frac{i}{N})$，$0 \le j \le 3$，為四個勒讓德多項式(Legendre polynomial)的基底，定義如下：

$$\begin{aligned}
&\Phi_0(\frac{i}{N}) = 1,\\
&\Phi_1(\frac{i}{N}) = \left[\frac{12\cdot N}{N+2}\right]^{\frac{1}{2}}\cdot\left[\left(\frac{i}{N}\right)-\frac{1}{2}\right],\\
&\Phi_2(\frac{i}{N}) = \left[\frac{180\cdot N^3}{(N-1)(N+2)(N+3)}\right]^{\frac{1}{2}}\cdot\left[\left(\frac{i}{N}\right)^2-\left(\frac{i}{N}\right)+\frac{N-1}{6\cdot N}\right],\\
&\Phi_3(\frac{i}{N}) = \left[\frac{2800\cdot N^5}{(N-1)(N-2)(N+2)(N+3)(N+4)}\right]^{\frac{1}{2}}\times\\
&\qquad\left[\left(\frac{i}{N}\right)^3-\frac{3}{2}\left(\frac{i}{N}\right)^2+\frac{6\cdot N^2-3\cdot N+2}{10\cdot N^2}\left(\frac{i}{N}\right)-\frac{(N-1)(N-2)}{20\cdot N^2}\right].
\end{aligned} \tag{3-2}$$

藉由正交化展開，可利用四維參數表示一個音節基頻軌跡；在基頻轉換所使用的向量，即為此四維參數。轉換後的四維參數可使用下列數學式還原音節基頻軌跡：

$$f'(i) = \sum_{j=0}^{3} a_j \cdot \Phi_j(\frac{i}{N}) \text{ 。} \quad (3\text{-}3)$$

3.2 傳統韻律轉換方法簡介

在本節將會介紹兩種傳統的韻律轉換方法，並以此作為與本論文提出方法的比較基準。

3.2.1 高斯正規化轉換

韻律轉換的方法中，最常使用的轉換方法為高斯正規化(Gaussian Normalization)的方式，也就是對平均值與變異數做一線性轉換，此方法亦稱為(Mean/Variance Transformation)。此方法的優點在於簡單易於實作，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 $\mathbf{y}_n$ 分別表示來源語者與目標語者在第 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：

$$P(\mathbf{x}_n) = N(\mathbf{x}_n; \boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma}_{\mathbf{xx}}) \text{ and } P(\mathbf{y}_n) = N(\mathbf{y}_n; \boldsymbol{\mu}_\mathbf{y}, \boldsymbol{\Sigma}_{\mathbf{yy}}) \quad (3\text{-}4)$$

其中，$\boldsymbol{\mu}_\mathbf{x}$ 與 $\boldsymbol{\mu}_\mathbf{y}$ 分別為來源和目標語者的期望值向量；$\boldsymbol{\Sigma}_{\mathbf{xx}}$ 與 $\boldsymbol{\Sigma}_{\mathbf{yy}}$ 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：

$$\hat{\mathbf{y}}_n = (\boldsymbol{\Sigma}_{\mathbf{yy}})^{\frac{1}{2}} (\boldsymbol{\Sigma}_{\mathbf{xx}})^{\frac{-1}{2}} (\mathbf{x}_n - \boldsymbol{\mu}_\mathbf{x}) + \boldsymbol{\mu}_\mathbf{y} \quad (3\text{-}5)$$

其中，$\hat{\mathbf{y}}_n$ 為轉換後第 n 個音節的韻律參數。然而此方法的缺點在於，當來源語料與目標語料為平行語料時，並無法有效的利用平行語料之間的相關性來做轉換，使得轉換後的效果無法進一步提升。

3.2.2 聯合高斯混合模型轉換

為了能建立來源語者與目標語者之間的關聯性，提升轉換的效果，以高斯混合模型為基礎的轉換方法[10]，可以有效的利用平行語料，建立來源語者與目標語者的相關特性。此方法在頻譜轉換的研究上，已被驗證有相當不錯的效果；過去也有學者將此方法應用於基頻轉換[5, 14]。

令 $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n\}$ 與 $\mathbf{Y}=\{\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_n\}$ 分別為來源以及目標語者的韻律參數序列，而 $\mathbf{Z}=[\mathbf{X}^{\mathrm{T}},\mathbf{Y}^{\mathrm{T}}]^{\mathrm{T}}$ 為一組韻律參數向量對，“T”為矩陣轉置符號，$\mathbf{x}$ 與 $\mathbf{y}$ 之向量維度皆為 d。值得注意的是，此方法所使用的語料為平行語料，因此來源與目標語者的音節數必定相同，故不需要做額外的校準(Align)步驟。以下先以單一高斯分佈來說明與推導此方法。首先，此方法假設 $[\mathbf{X}^{\mathrm{T}},\mathbf{Y}^{\mathrm{T}}]^{\mathrm{T}}$ 的聯合機率分布符合高斯分佈，其機率密度函數為：

$$P(\mathbf{z})=P(\mathbf{x},\mathbf{y})=\frac{1}{(2\pi)^d\sqrt{|\Sigma^{\mathbf{zz}}|}}\exp\left(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}^{\mathbf{z}})^T(\Sigma^{\mathbf{zz}})^{-1}(\mathbf{z}-\boldsymbol{\mu}^{\mathbf{z}})\right) \tag{3-6}$$

其中，$\boldsymbol{\mu}^{\mathbf{z}}=[(\boldsymbol{\mu}^{\mathbf{x}})^{\mathbf{T}},(\boldsymbol{\mu}^{\mathbf{y}})^{\mathbf{T}}]^{\mathbf{T}}$，$\boldsymbol{\Sigma}^{\mathbf{zz}}=\begin{bmatrix}\boldsymbol{\Sigma}^{\mathbf{xx}} & \boldsymbol{\Sigma}^{\mathbf{xy}}\\ \boldsymbol{\Sigma}^{\mathbf{yx}} & \boldsymbol{\Sigma}^{\mathbf{yy}}\end{bmatrix}$，在此希望找到轉換函式 $F(\mathbf{x})$，使得目標韻律向量序列與轉換後之值，能夠有最小的均方差(mean square error)，也就是使 $\varepsilon_{mse}=E\left[\left\|\mathbf{y}-F(\mathbf{x})\right\|^2\right]$ 為最小。根據最小均方差(Minimum Mean Square Error, MMSE)之法則，當 ε_{mse} 有最小值時，轉換函式 $F(\mathbf{x})=E[\mathbf{y}\,|\,\mathbf{x}]$，而其條件機率密度函式為：

$$P(\mathbf{y}\,|\,\mathbf{x})=\frac{P(\mathbf{x},\mathbf{y})}{P(\mathbf{x})}=\frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma^{\mathbf{YY}}-\Sigma^{\mathbf{YX}}(\Sigma^{\mathbf{XX}})^{-1}\Sigma^{\mathbf{XY}}|}}\exp\left(-\frac{1}{2}\mathbf{U}\right) \tag{3-7}$$

其中，

$\mathbf{U}=(\mathbf{y}-(\boldsymbol{\mu}^{\mathbf{y}}+\Sigma^{\mathbf{YX}}(\Sigma^{\mathbf{XX}})^{-1}(\mathbf{x}-\boldsymbol{\mu}^{\mathbf{x}})))^T(\Sigma^{\mathbf{YY}}-\Sigma^{\mathbf{YX}}(\Sigma^{\mathbf{XX}})^{-1}\Sigma^{\mathbf{XY}})^{-1}(\mathbf{y}-(\boldsymbol{\mu}^{\mathbf{y}}+\Sigma^{\mathbf{YX}}(\Sigma^{\mathbf{XX}})^{-1}(\mathbf{x}-\boldsymbol{\mu}^{\mathbf{x}})))$，因此可以得到轉換函式：

$$F(\mathbf{x}) = E[\mathbf{y} \mid \mathbf{x}] = (\boldsymbol{\mu}^{\mathbf{y}} + \Sigma^{\mathbf{yx}}(\Sigma^{\mathbf{xx}})^{-1}(\mathbf{x} - \boldsymbol{\mu}^{\mathbf{x}}))\ 。 \tag{3-8}$$

上式是以單一 mixture 描述 $\mathbf{x}$ 與 $\mathbf{y}$ 之聯合機率分佈，以此為基礎，進一步以多個 mixture 之高斯混合模型建構 $\mathbf{Z}$ 的機率分佈：

$$P(\mathbf{z}_n) = P(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{M} \alpha_i N(\mathbf{z}_n; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{3-9}$$

其中，$N(\mathbf{z}_n; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ 為第 i 個 mixture 的高斯機率分佈；$\boldsymbol{\mu}_i = [(\boldsymbol{\mu}_i^{\mathbf{x}})^{\mathbf{T}}, (\boldsymbol{\mu}_i^{\mathbf{y}})^{\mathbf{T}}]^{\mathbf{T}}$ 以及 $\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{\mathbf{xx}} & \boldsymbol{\Sigma}_i^{\mathbf{xy}} \\ \boldsymbol{\Sigma}_i^{\mathbf{yx}} & \boldsymbol{\Sigma}_i^{\mathbf{yy}} \end{bmatrix}$ 則分別為第 i 個 mixture 之期望值向量與共變異數矩陣；α_i 為每個 mixture 的權重，且 $\sum_{i=1}^{M} \alpha_i = 1$；$M$ 為總共的 mixture 數；利用 EM(Expectation-Maximization)演算法，可以估計出高斯混合模型的參數。

而基於高斯混合模型的轉換函式，可以推導如下：

$$\begin{aligned} F(\mathbf{x}) = E[\mathbf{y} \mid \mathbf{x}] &= \int_{\mathbf{y}} \mathbf{y} P(\mathbf{y} \mid \mathbf{x}) d\mathbf{y} = \int_{\mathbf{y}} \mathbf{y} \frac{P(\mathbf{x}, \mathbf{y})}{P(\mathbf{x})} d\mathbf{y} = \int_{\mathbf{y}} \mathbf{y} \frac{\sum_{i}^{M} \alpha_i N_i(\mathbf{x}, \mathbf{y})}{\sum_{i}^{M} \alpha_i N_i(\mathbf{x})} d\mathbf{y} \\ &= \int_{\mathbf{y}} \mathbf{y} \sum_{i}^{M} \left[\frac{\alpha_i N_i(\mathbf{x})}{\sum_{i} \alpha_i N_i(\mathbf{x})} N_i(\mathbf{y} \mid \mathbf{x}) \right] d\mathbf{y} = \int_{\mathbf{y}} \mathbf{y} \sum_{i}^{M} P(i \mid \mathbf{x}) N_i(\mathbf{y} \mid \mathbf{x}) d\mathbf{y} \\ &= \sum_{i}^{M} \left[P(i \mid \mathbf{x}) \int_{\mathbf{y}} \mathbf{y} N_i(\mathbf{y} \mid \mathbf{x}) d\mathbf{y} \right] = \sum_{i}^{i} P(i \mid \mathbf{x}) E[\mathbf{y} \mid \mathbf{x}, i] \\ &= \sum_{i=1}^{M} P(i \mid \mathbf{x}_n) \left[\boldsymbol{\mu}_i^{\mathbf{y}} + \boldsymbol{\Sigma}_i^{\mathbf{xy}} (\boldsymbol{\Sigma}_i^{\mathbf{xx}})^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_i^{\mathbf{x}}) \right] \end{aligned} \tag{3-10}$$

其中，$P(i \mid \mathbf{x}_n) = \dfrac{\alpha_i N(\mathbf{x}_n; \boldsymbol{\mu}_i^{\mathbf{x}}, \boldsymbol{\Sigma}_i^{\mathbf{xx}})}{\sum_{i=1}^{M} \alpha_i N(\mathbf{x}_n; \boldsymbol{\mu}_i^{\mathbf{x}}, \boldsymbol{\Sigma}_i^{\mathbf{xx}})}$ 表示 $\mathbf{x}_n$ 屬於第 i 個 mixture 之機率。

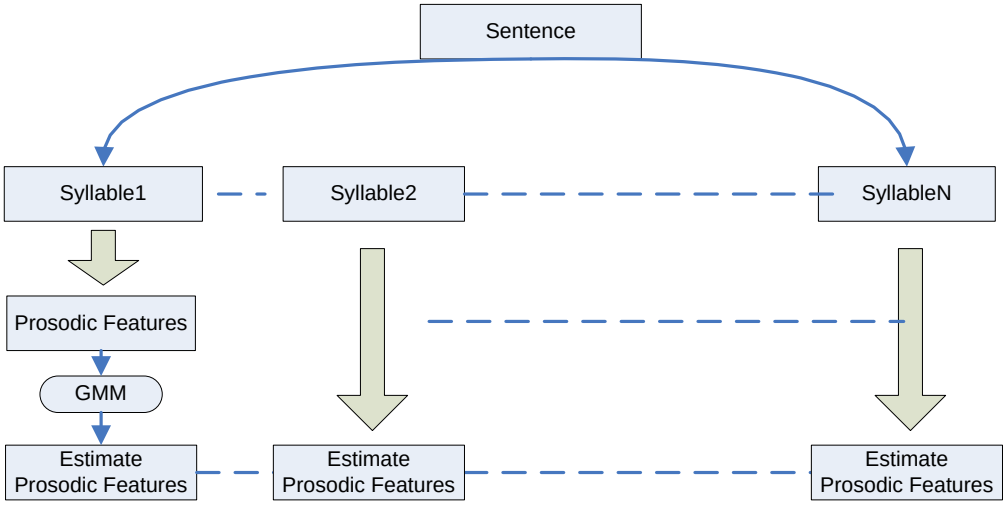

圖 3-1：GMM 轉換之概念圖

此轉換方法之概念如圖 3-1 所示。在此轉換函式中，每個來源韻律參數都可獨自代入轉換，得到預估之目標韻律參數，這也意謂著每個參數之間是互為獨立的；然而，此方法缺點在於沒有考慮音節與音節之間的關聯性，且無法有效利用語言參數於韻律轉換中。

3.3 以韻律模型為基礎之基頻轉換

音節基頻軌跡反應語者說話音調的高低起伏變化，尤其在中文裡，音節之聲調表現在基頻軌跡中，因此基頻軌跡在所有韻律參數中扮演著重要的角色。本節將介紹提出的兩種基頻轉換方法。

3.3.1 基頻轉換方法一

在第二章介紹了音節基頻軌跡模型(如 2-8 式)，可將來源與目標語者的音節基頻軌跡特徵向量分別表示成一個高斯分佈，如下數學式：

$$P(\mathbf{x}_n \mid p_n^{\mathbf{x}}, B_{n-1}^{\mathbf{x},n}, t_{n-1}^{n+1}) = N(\mathbf{x}_n; \boldsymbol{\beta}_{t_n}^{\mathbf{x}} + \boldsymbol{\beta}_{p_n^{\mathbf{x}}}^{\mathbf{x}} + \boldsymbol{\beta}_{B_{n-1}^{\mathbf{x}}, tp_{n-1}}^{\mathbf{x},f} + \boldsymbol{\beta}_{B_n^{\mathbf{x}}, tp_n}^{\mathbf{x},b} + \boldsymbol{\mu}_{\mathbf{x}}, \mathbf{R}_{\mathbf{x}}) \tag{3-11}$$

及

$$P(\mathbf{y}_n \mid p_n^{\mathbf{y}}, B_{n-1}^{\mathbf{y},n}, t_{n-1}^{n+1}) = N(\mathbf{y}_n; \boldsymbol{\beta}_{t_n}^{\mathbf{y}} + \boldsymbol{\beta}_{p_n^{\mathbf{y}}}^{\mathbf{y}} + \boldsymbol{\beta}_{B_{n-1}^{\mathbf{y}}, tp_{n-1}}^{\mathbf{y},f} + \boldsymbol{\beta}_{B_n^{\mathbf{y}}, tp_n}^{\mathbf{y},b} + \boldsymbol{\mu}_{\mathbf{y}}, \mathbf{R}_{\mathbf{y}}) \tag{3-12}$$

其中，$\mathbf{x}_n$ 及 $\mathbf{y}_n$ 分別表示來源語者及目標語者第 n 個音節基頻軌跡正交化參數；上標與下標符號“$\mathbf{x}$”與“$\mathbf{y}$”分別表示來源語者與目標語者。藉由以上兩式，並運用高斯正規化轉換的概念，可以得到轉換函式：

$$\begin{aligned}\hat{\mathbf{y}}_n = (\mathbf{R}_{\mathbf{y}})^{\frac{1}{2}} (\mathbf{R}_{\mathbf{x}})^{\frac{-1}{2}} \left\{ \mathbf{x}_n - (\boldsymbol{\beta}_{t_n}^{\mathbf{x}} + \boldsymbol{\beta}_{p_n^{\mathbf{x}}}^{\mathbf{x}} + \boldsymbol{\beta}_{B_{n-1}^{\mathbf{x}}, tp_{n-1}}^{\mathbf{x},f} + \boldsymbol{\beta}_{B_n^{\mathbf{x}}, tp_n}^{\mathbf{x},b} + \boldsymbol{\mu}_{\mathbf{x}}) \right\} \\ + (\boldsymbol{\beta}_{t_n}^{\mathbf{y}} + \boldsymbol{\beta}_{p_n^{\mathbf{y}}}^{\mathbf{y}} + \boldsymbol{\beta}_{B_{n-1}^{\mathbf{y}}, tp_{n-1}}^{\mathbf{y},f} + \boldsymbol{\beta}_{B_n^{\mathbf{y}}, tp_n}^{\mathbf{y},b} + \boldsymbol{\mu}_{\mathbf{y}})\end{aligned} \tag{3-13}$$

與高斯正規化轉換方式相比較，此式子若僅考量語者說話的影響因素，即 $\boldsymbol{\mu}_{\mathbf{x}}$ 與 $\boldsymbol{\mu}_{\mathbf{y}}$，則該式可化簡成 3-5 式。由此可知，方法一可以視為高斯正規化轉換方式的廣義表示方式(General Form)。

接著進一步化簡上式；首先，在此轉換函式中，值得注意的是，$\mathbf{x}_n - (\boldsymbol{\beta}_{t_n}^{\mathbf{x}} + \boldsymbol{\beta}_{p_n^{\mathbf{x}}}^{\mathbf{x}} + \boldsymbol{\beta}_{B_{n-1}^{\mathbf{x}}, tp_{n-1}}^{\mathbf{x},f} + \boldsymbol{\beta}_{B_n^{\mathbf{x}}, tp_n}^{\mathbf{x},b} + \boldsymbol{\mu}_{\mathbf{x}})$ 這一項已變成為基頻的殘存值；因為在基頻上有意義的影響因素都已被扣除，所以殘存值的機率分布傾向於一個變異數極小的白雜訊；除此之外，來源基頻與目標基頻的殘存值相關性極小，因此不適合將來源基頻殘存值用來估計目標基頻殘存值。基於以上兩個原因，在此將 $\mathbf{x}_n - (\boldsymbol{\beta}_{t_n}^{\mathbf{x}} + \boldsymbol{\beta}_{p_n^{\mathbf{x}}}^{\mathbf{x}} + \boldsymbol{\beta}_{B_{n-1}^{\mathbf{x}}, tp_{n-1}}^{\mathbf{x},f} + \boldsymbol{\beta}_{B_n^{\mathbf{x}}, tp_n}^{\mathbf{x},b} + \boldsymbol{\mu}_{\mathbf{x}})$ 項從轉換函式中移除，並將 3-13 式簡化成如下：

$$\hat{\mathbf{y}}_n = \boldsymbol{\beta}_{t_n}^{\mathbf{y}} + \boldsymbol{\beta}_{p_n^{\mathbf{y}}}^{\mathbf{y}} + \boldsymbol{\beta}_{B_{n-1}^{\mathbf{y}}, tp_{n-1}}^{\mathbf{y},f} + \boldsymbol{\beta}_{B_n^{\mathbf{y}}, tp_n}^{\mathbf{y},b} + \boldsymbol{\mu}_{\mathbf{y}} \text{ 。} \tag{3-14}$$

基本上，在 3-14 式所表示的就是目標基頻軌跡模型的期望值，但在轉換時，為了要估計 $\hat{\mathbf{y}}_n$，仍然需要知道語言參數 t_{n-1}^{n+1}，以及韻律標記資訊 $p_n^{\mathbf{y}}$, $B_{n-1}^{\mathbf{y}}$, $B_n^{\mathbf{y}}$，而這些資訊需要經由來源語者的韻律標記來預估；其中，轉換聲音之聲調(tone)以及聲調組合(tone pair)，必定與來源語者一致，因此可以直接以來源語者之 t_{n-1}^{n+1} 取代；此方法也假設目標語者與來源語者有相同的說

話風格(speaking style)，也就是假設：

$$B_{n-1}^{\mathbf{y}} = B_{n-1}^{\mathbf{x}} \quad \text{以及} \quad B_n^{\mathbf{y}} = B_n^{\mathbf{x}} \tag{3-15}$$

對於韻律狀態(Prosodic State)的轉換，目的是建立來源語者的韻律狀態 $p_n^{\mathbf{x}}$ 與目標語者的韻律狀態 $p_n^{\mathbf{y}}$ 之間的關係。同樣的因為假設目標與來源語者有相同的說話風格，因此假設：

$$p_n^{\mathbf{y}} = p_n^{\mathbf{x}} \tag{3-16}$$

但是同時考慮到來源與目標語者之韻律狀態並非固定為一對一之對應關係，因此更進一步以高斯正規化的方式，對來源語者的韻律狀態做線性轉換，再去尋找轉換後的狀態值，最接近哪一個目標韻律狀態碼字，如下數學式所示：

$$\hat{p}_n^{\mathbf{y}} = \arg\min_i \left(\frac{\boldsymbol{\beta}_{p_n^{\mathbf{x}}}^{\mathbf{x}}(1)}{\sigma_{\mathbf{x}}} \sigma_{\mathbf{y}} - \boldsymbol{\beta}_{p_n=i}^{\mathbf{y}}(1) \right)^2 \tag{3-17}$$

其中，$\boldsymbol{\beta}_{p_n^{\mathbf{x}}}^{\mathbf{x}}(1)$ 代表 $\boldsymbol{\beta}_{p_n^{\mathbf{x}}}^{\mathbf{x}}$ 的第一維數值；$\sigma_{\mathbf{x}}$ 與 $\sigma_{\mathbf{y}}$ 分別表示訓練語料中音節基頻軌跡參數扣除了韻律狀態以外的影響因素(Affecting Factor)的標準差，如下數學式所述：

$$\sigma_{\mathbf{x}} = Std(\mathbf{x}_n(1) - \boldsymbol{\beta}_{t_n}^{\mathbf{x}}(1) - \boldsymbol{\beta}_{B_{n-1}^{\mathbf{x}}, tp_{n-1}}^{\mathbf{x},f}(1) - \boldsymbol{\beta}_{B_n^{\mathbf{x}}, tp_n}^{\mathbf{x},b}(1) - \boldsymbol{\mu}_{\mathbf{x}}(1)) \tag{3-18}$$

以及

$$\sigma_{\mathbf{y}} = Std(\mathbf{y}_n(1) - \boldsymbol{\beta}_{t_n}^{\mathbf{y}}(1) - \boldsymbol{\beta}_{B_{n-1}^{\mathbf{y}}, tp_{n-1}}^{\mathbf{y},f}(1) - \boldsymbol{\beta}_{B_n^{\mathbf{y}}, tp_n}^{\mathbf{y},b}(1) - \boldsymbol{\mu}_{\mathbf{y}}(1)) \tag{3-19}$$

值得注意的是，在此假設韻律狀態的影響因素向量只有在第一維有值，這是因為韻律狀態所代表的是韻律上層的資訊，因此它只會對音節基頻軌跡的層級(level)有影響。最後，可以將 3-14 式改寫成

$$\hat{\mathbf{y}}_n = \boldsymbol{\beta}_{t_n}^{\mathbf{y}} + \boldsymbol{\beta}_{\hat{p}_n^{\mathbf{y}}}^{\mathbf{y}} + \boldsymbol{\beta}_{B_{n-1}^{\mathbf{x}}, tp_{n-1}}^{\mathbf{y},f} + \boldsymbol{\beta}_{B_n^{\mathbf{x}}, tp_n}^{\mathbf{y},b} + \boldsymbol{\mu}_{\mathbf{y}} \tag{3-20}$$

此方法繼承了高斯正規化轉換方式的優點，即不需要特別準備平行的訓練語料;此外，在更精細的考量各影響因素後，可以預期轉換後的效果能有所提升。

3.3.2 基頻轉換方法二

在上小節提出的方法一中，有兩個缺點。首先，方法一強烈的假設了來源與目標語者有相同的韻律片語結構，即假設彼此的說話方式相似，因此方法一令轉換聲音與來源語料有相同的停頓標記(如 3-15 式)；第二，在轉換語音時，給定韻律狀態的方式是以簡單的線性轉換，即 hard decision 的方式決定(如 3-17 式)。當轉換的句子長度較短時，這些假設是合理的，因為較短的句子相對而言有較簡單的韻律片語結構，亦即唸短句時，來源與目標語者的說話方式應該相似；相反的，當轉換的句子長度較長時，來源語者與目標語者的說話特性則會呈現出來，例如相同的語句，但兩者在不同的位置發生停頓，使得來源與目標語者該句話的韻律片語結構差異變大，進而造成轉換上的效果不好。此影響尤其對來源與目標語者有不同說話方式。

為了克服這兩項缺點，在此提出了方法二的轉換。首先，採用 MMSE 為準則以 $\mathbf{x}_n$ 預估 $\mathbf{y}_n$：

$$\hat{\mathbf{y}}_n = E[\mathbf{y}_n \mid \mathbf{x}_n] = \int \mathbf{y}_n P(\mathbf{y}_n \mid \mathbf{x}_n) d\mathbf{y}_n \tag{3-21}$$

其中，

$$P(\mathbf{y}_n \mid \mathbf{x}_n) = \sum_{\mathbf{B}^{\mathbf{y}}} \sum_{\mathbf{p}^{\mathbf{y}}} P(\mathbf{y}_n, \mathbf{p}^{\mathbf{y}}, \mathbf{B}^{\mathbf{y}} \mid \mathbf{x}_n, \mathbf{p}^{\mathbf{x}^*}, \mathbf{B}^{\mathbf{x}^*}, \mathbf{L}) \tag{3-22}$$

$\mathbf{p}^{\mathbf{y}}$ 以及 $\mathbf{B}^{\mathbf{y}}$ 分別為目標語句之韻律狀態與停頓標記序列；$\mathbf{p}^{\mathbf{x}^*}$ 以及 $\mathbf{B}^{\mathbf{x}^*}$ 分別為來源語句之韻律狀態與停頓標記序列；$\mathbf{L}=\{ t_{n-1}^{n+1},\ \mathbf{l}_n \}$ 為與來源語句相關的語言參數集合。接著將 $P(\mathbf{y}_n, \mathbf{p}^{\mathbf{y}}, \mathbf{B}^{\mathbf{y}} \mid \mathbf{x}_n, \mathbf{p}^{\mathbf{x}^*}, \mathbf{B}^{\mathbf{x}^*}, \mathbf{L})$ 簡化如下：

$$
\begin{aligned}
&P(\mathbf{y}_n, \mathbf{p}^{\mathbf{y}}, \mathbf{B}^{\mathbf{y}} \mid \mathbf{x}_n, \mathbf{p}^{\mathbf{x}^*}, \mathbf{B}^{\mathbf{x}^*}, \mathbf{L}) \\
&= P(\mathbf{y}_n \mid \mathbf{p}^{\mathbf{y}}, \mathbf{B}^{\mathbf{y}}, \mathbf{x}_n, \mathbf{p}^{\mathbf{x}^*}, \mathbf{B}^{\mathbf{x}^*}, \mathbf{L}) P(\mathbf{p}^{\mathbf{y}}, \mathbf{B}^{\mathbf{y}} \mid \mathbf{x}_n, \mathbf{p}^{\mathbf{x}^*}, \mathbf{B}^{\mathbf{x}^*}, \mathbf{L}) \\
&\approx P(\mathbf{y}_n \mid p_n^{\mathbf{y}}, B_{n-1}^{\mathbf{y},n}, t_{n-1}^{n+1}) P(\mathbf{p}^{\mathbf{y}}, \mathbf{B}^{\mathbf{y}} \mid \mathbf{p}^{\mathbf{x}^*}, \mathbf{B}^{\mathbf{x}^*}, \mathbf{L})
\end{aligned}
\tag{3-23}
$$

其中，$P(\mathbf{y}_n \mid p_n^{\mathbf{y}}, B_{n-1}^{\mathbf{y},n}, t_{n-1}^{n+1})$ 為目標音節基頻軌跡模型；$P(\mathbf{p}^{\mathbf{y}}, \mathbf{B}^{\mathbf{y}} \mid \mathbf{p}^{\mathbf{x}^*}, \mathbf{B}^{\mathbf{x}^*}, \mathbf{L})$ 為韻律標記對應模型(prosody-tag mapping model)。前者之簡化是為了只考慮對於 $\mathbf{y}_n$ 有最重要的影響因素，而後者是為了只對韻律標記做轉換。

將 3-22 及 3-23 式代入 3-21 式中，可進一步推導為：

$$
\begin{aligned}
\hat{\mathbf{y}}_n &= \sum_{\mathbf{p}^{\mathbf{y}}} \sum_{\mathbf{B}^{\mathbf{y}}} P(\mathbf{p}^{\mathbf{y}}, \mathbf{B}^{\mathbf{y}} \mid \mathbf{p}^{\mathbf{x}^*}, \mathbf{B}^{\mathbf{x}^*}, \mathbf{L}) \int \mathbf{y}_n P(\mathbf{y}_n \mid p_n^{\mathbf{y}}, B_{n-1}^{\mathbf{y},n}, t_{n-1}^{n+1}) d\mathbf{y}_n \\
&\approx \sum_{p_n^{\mathbf{y}}} \sum_{B_{n-1}^{\mathbf{y},n}} P(p_n^{\mathbf{y}}, B_{n-1}^{\mathbf{y},n} \mid p_1^{\mathbf{y},n-1}, B_1^{\mathbf{y},n-2}, \mathbf{p}^{\mathbf{x}^*}, \mathbf{B}^{\mathbf{x}^*}, \mathbf{L}) E[\mathbf{y}_n \mid p_n^{\mathbf{y}}, B_{n-1}^{\mathbf{y},n}, t_{n-1}^{n+1}]
\end{aligned}
\tag{3-24}
$$

3-24 式的化簡，是為了以從語句開始到時間點 n 為止的資訊，預估出 $\mathbf{y}_n$。其物理意義可以解釋為：總共有 $\mathbf{p}^{\mathbf{y}} \times \mathbf{B}^{\mathbf{y}}$ 個轉換函式 $E[\mathbf{y}_n \mid p_n^{\mathbf{y}}, B_{n-1}^{\mathbf{y},n}, t_{n-1}^{n+1}]$，每個轉換函式分別給予不同的權重 $P(\mathbf{p}^{\mathbf{y}}, \mathbf{B}^{\mathbf{y}} \mid \mathbf{p}^{\mathbf{x}^*}, \mathbf{B}^{\mathbf{x}^*}, \mathbf{L})$ 後作相加(weighted summation)，進而得到預估的基頻軌跡；此權重則表示利用來源語句的韻律標記資訊($\mathbf{p}^{\mathbf{x}^*}, \mathbf{B}^{\mathbf{x}^*}$)以及語言參數 $\mathbf{L}$，預估目標語者的韻律標記資訊($\mathbf{p}^{\mathbf{y}}, \mathbf{B}^{\mathbf{y}}$)所得到的機率值即為該權重。方法二與方法一主要的差異在於，前者是用機率統計式的方式，亦即以軟式決策(soft decision)的方式去預估目標韻律標記，不同於後者是採用硬式決策(hard decision)的概念。

由於在 2.2 節中假設 $P(\mathbf{y}_n \mid p_n^{\mathbf{y}}, B_{n-1}^{\mathbf{y},n}, t_{n-1}^{n+1})$ 為一個高斯分佈，因此 3-24 式可改寫成

$$
\begin{aligned}
\hat{\mathbf{y}}_n = \sum_{p_n^{\mathbf{y}}} \sum_{B_{n-1}^{\mathbf{y},n}} & P(p_n^{\mathbf{y}}, B_{n-1}^{\mathbf{y},n} \mid p_1^{\mathbf{y},n-1}, B_1^{\mathbf{y},n-2}, \mathbf{p}^{\mathbf{x}^*}, \mathbf{B}^{\mathbf{x}^*}, \mathbf{L}) \\
& \cdot (\boldsymbol{\beta}_{t_n}^{\mathbf{y}} + \boldsymbol{\beta}_{p_n^{\mathbf{y}}}^{\mathbf{y}} + \boldsymbol{\beta}_{B_{n-1}^{\mathbf{y}}, tp_{n-1}}^{\mathbf{y},f} + \boldsymbol{\beta}_{B_n^{\mathbf{y}}, tp_n}^{\mathbf{y},b} + \boldsymbol{\mu}_{\mathbf{y}})
\end{aligned}
\tag{3-25}
$$

韻律標記對應函式可以用遞迴的方式得到：

$$P(p_n^{\mathbf{y}}, B_{n-1}^{\mathbf{y},n} | p_1^{\mathbf{y},n-1}, B_1^{\mathbf{y},n-2}, \mathbf{p}^{\mathbf{x}^*}, \mathbf{B}^{\mathbf{x}^*}, \mathbf{L})$$

$$= \begin{cases} P(p_1^{\mathbf{y}} | B_1^{\mathbf{y}}, p_1^{\mathbf{x}^*,2}, B_1^{\mathbf{x}^*}) P(B_1^{\mathbf{y}} | B_1^{\mathbf{x}^*}, \mathbf{l}_1) & n=1 \\ \sum_{p_{n-1}^{\mathbf{y}}} \left\{ \begin{matrix} P(p_1^{\mathbf{y}} | B_1^{\mathbf{y}}, p_1^{\mathbf{x}^*,2}, B_1^{\mathbf{x}^*}) P(B_1^{\mathbf{y}} | B_1^{\mathbf{x}^*}, \mathbf{l}_1) \times \\ P(p_2^{\mathbf{y}} | p_1^{\mathbf{y}}, B_1^{\mathbf{y},2}, p_1^{\mathbf{x}^*,3}, B_2^{\mathbf{x}^*}) P(B_2^{\mathbf{y}} | B_2^{\mathbf{x}^*}, \mathbf{l}_2) \end{matrix} \right\} & n=2 \\ \sum_{p_{n-1}^{\mathbf{y}}} \sum_{B_{n-2}^{\mathbf{y}}} \left\{ \begin{matrix} P(p_{n-1}^{\mathbf{y}}, B_{n-2}^{\mathbf{y},n-1} | p_1^{\mathbf{y},n-2}, B_1^{\mathbf{y},n-3}, \mathbf{p}^{\mathbf{x}^*}, \mathbf{B}^{\mathbf{x}^*}, \mathbf{L}) \times \\ P(p_n^{\mathbf{y}} | p_{n-1}^{\mathbf{y}}, B_{n-1}^{\mathbf{y},n}, p_{n-1}^{\mathbf{x}^*,n+1}, B_n^{\mathbf{x}^*}) P(B_n^{\mathbf{y}} | B_n^{\mathbf{x}^*}, \mathbf{l}_n) \end{matrix} \right\} & 3 \le n \le N \end{cases} \quad (3\text{-}26)$$

其中，$P(p_n^{\mathbf{y}} | p_{n-1}^{\mathbf{y}}, B_{n-1}^{\mathbf{y},n}, p_{n-1}^{\mathbf{x}^*,n+1}, B_n^{\mathbf{x}^*})$ 為韻律狀態對應函式(prosodic state mapping function)，藉由前一個轉換的韻律狀態 $p_{n-1}^{\mathbf{y}}$、對應音節其相鄰的停頓標記 $B_{n-1}^{\mathbf{y},n}$，以及來源音節相鄰的韻律標記 $p_{n-1}^{\mathbf{x}^*,n+1}$ 與 $B_n^{\mathbf{x}^*}$，預估現在轉換音節的韻律狀態 $p_n^{\mathbf{y}}$；而 $P(B_n^{\mathbf{y}} | B_n^{\mathbf{x}^*}, \mathbf{l}_n)$ 為停頓標記對應函式(break mapping function)，藉由來源音節的停頓標記 $B_n^{\mathbf{x}^*}$，以及前後文的語言參數 $\mathbf{l}_n$，預估現在轉換音節的停頓標記 $B_n^{\mathbf{y}}$。

在實作上，因為 $\{ p_{n-1}^{\mathbf{y}}, B_{n-1}^{\mathbf{y},n}, p_{n-1}^{\mathbf{x}^*,n+1}, B_n^{\mathbf{x}^*} \}$ 以及 $\{ B_n^{\mathbf{x}^*}, \mathbf{l}_n \}$ 的空間集合太大，可能會造成某些空間組合的資料量過少，因此在本研究中採用 CART 演算法，分別藉由問題集對兩個空間，依據最大概似函數增益的判定原則分裂節點；最後每個葉節點分類成 $C(p_{n-1}^{\mathbf{y}}, B_{n-1}^{\mathbf{y},n}, p_{n-1}^{\mathbf{x}^*,n+1}, B_n^{\mathbf{x}^*})$ 以及 $C(B_n^{\mathbf{x}^*}, \mathbf{l}_n)$。因此 3-26 式中的兩個對應函式可以改寫成：

$$P(p_n^{\mathbf{y}} | p_{n-1}^{\mathbf{y}}, B_{n-1}^{\mathbf{y},n}, p_{n-1}^{\mathbf{x}^*,n+1}, B_n^{\mathbf{x}^*}) \approx P\left(p_n^{\mathbf{y}} | C(p_{n-1}^{\mathbf{y}}, B_{n-1}^{\mathbf{y},n}, p_{n-1}^{\mathbf{x}^*,n+1}, B_n^{\mathbf{x}^*})\right) \quad (3\text{-}27)$$

以及

$$P(B_n^{\mathbf{y}} | B_n^{\mathbf{x}^*}, \mathbf{l}_n) \approx P\left(B_n^{\mathbf{y}} | C(B_n^{\mathbf{x}^*}, \mathbf{l}_n)\right)。 \quad (3\text{-}28)$$

3.4 以韻律模型為基礎之音長與能量轉換

前一小節探討了音節基頻軌跡之轉換，我們也可以將前文所提及基頻軌跡轉換方法的想法套用於音節長度與能量之轉換。音節的長度反應出說話速度的快慢，而音節之能量則代表聲音的大小聲，二者都影響聽者的感覺，因此若能將此二種韻律參數藉由韻律模型做轉換，則轉換後之聲音更能突顯目標語者的說話特性。本節介紹以韻律模型為基礎之音節長度及能量轉換方法。

3.4.1 音長與能量轉換方法一

如同 2-8 式之音節基頻軌跡模型，2-9 與 2-10 式描繪了音節長度與能量，能將來源語者與目標語者之音長以及能量分別以高斯分佈表示：

$$P(sd_n^{\mathbf{x}} \mid q_n^{\mathbf{x}}, t_n, s_n, u_n^{\mathbf{x}}) = N(sd_n^{\mathbf{x}}; \gamma_{t_n}^{\mathbf{x}} + \gamma_{q_n^{\mathbf{x}}}^{\mathbf{x}} + \gamma_{s_n}^{\mathbf{x}} + \gamma_{u_n^{\mathbf{x}}}^{\mathbf{x}} + \mu_d^{\mathbf{x}}, R_d^{\mathbf{x}}) \tag{3-29}$$

$$P(se_n^{\mathbf{x}} \mid r_n^{\mathbf{x}}, t_n, f_n, u_n^{\mathbf{x}}) = N(se_n^{\mathbf{x}}; \alpha_{t_n}^{\mathbf{x}} + \alpha_{r_n^{\mathbf{x}}}^{\mathbf{x}} + \alpha_{f_n}^{\mathbf{x}} + \alpha_{u_n^{\mathbf{x}}}^{\mathbf{x}} + \mu_e^{\mathbf{x}}, R_e^{\mathbf{x}}) \tag{3-30}$$

及

$$P(sd_n^{\mathbf{y}} \mid q_n^{\mathbf{y}}, t_n, s_n, u_n^{\mathbf{y}}) = N(sd_n^{\mathbf{y}}; \gamma_{t_n}^{\mathbf{y}} + \gamma_{q_n^{\mathbf{y}}}^{\mathbf{y}} + \gamma_{s_n}^{\mathbf{y}} + \gamma_{u_n^{\mathbf{y}}}^{\mathbf{y}} + \mu_d^{\mathbf{y}}, R_d^{\mathbf{y}}) \tag{3-31}$$

$$P(se_n^{\mathbf{y}} \mid r_n^{\mathbf{y}}, t_n, f_n, u_n^{\mathbf{y}}) = N(se_n^{\mathbf{y}}; \alpha_{t_n}^{\mathbf{y}} + \alpha_{r_n^{\mathbf{y}}}^{\mathbf{y}} + \alpha_{f_n}^{\mathbf{y}} + \alpha_{u_n^{\mathbf{y}}}^{\mathbf{y}} + \mu_e^{\mathbf{y}}, R_e^{\mathbf{y}}) \tag{3-32}$$

其中，sd_n 及 se_n 分別表示語者第 n 個音節之音長與能量；上標與下標符號“**x**”與“**y**”分別表示來源語者與目標語者，其餘之符號表示可參照表 2-1。藉由 3-29 與 3-31 式，以高斯正規化轉換方式，可以得到音長轉換函式：

$$\hat{sd}_n^{\mathbf{y}} = (R_d^{\mathbf{y}})^{\frac{1}{2}}(R_d^{\mathbf{x}})^{\frac{-1}{2}}\left\{sd_n^{\mathbf{x}} - (\gamma_{t_n}^{\mathbf{x}} + \gamma_{q_n^{\mathbf{x}}}^{\mathbf{x}} + \gamma_{s_n}^{\mathbf{x}} + \gamma_{u_n^{\mathbf{x}}}^{\mathbf{x}} + \mu_d^{\mathbf{x}})\right\} + (\gamma_{t_n}^{\mathbf{y}} + \gamma_{q_n^{\mathbf{y}}}^{\mathbf{y}} + \gamma_{s_n}^{\mathbf{y}} + \gamma_{u_n^{\mathbf{y}}}^{\mathbf{y}} + \mu_d^{\mathbf{y}}) \tag{3-33}$$

利用 3-30 與 3-32 式，可得到能量轉換函式：

$$\hat{se}_n^{\mathbf{y}} = (R_e^{\mathbf{y}})^{\frac{1}{2}}(R_e^{\mathbf{x}})^{\frac{-1}{2}}\left\{se_n^{\mathbf{x}} - (\alpha_{t_n}^{\mathbf{x}} + \alpha_{r_n^{\mathbf{x}}}^{\mathbf{x}} + \alpha_{f_n}^{\mathbf{x}} + \alpha_{u_n^{\mathbf{x}}}^{\mathbf{x}} + \mu_e^{\mathbf{x}})\right\} + (\alpha_{t_n}^{\mathbf{y}} + \alpha_{r_n^{\mathbf{y}}}^{\mathbf{y}} + \alpha_{f_n}^{\mathbf{y}} + \alpha_{u_n^{\mathbf{y}}}^{\mathbf{y}} + \mu_e^{\mathbf{y}}) \tag{3-34}$$

基於與 3.3.1 節相同之想法，分別從 3-33 以及 3-34 式中將 $sd_n^{\mathbf{x}} - (\gamma_{t_n}^{\mathbf{x}} + \gamma_{q_n^{\mathbf{x}}}^{\mathbf{x}} + \gamma_{s_n}^{\mathbf{x}} + \gamma_{u_n^{\mathbf{x}}}^{\mathbf{x}} + \mu_d^{\mathbf{x}})$ 與 $se_n^{\mathbf{x}} - (\alpha_{t_n}^{\mathbf{x}} + \alpha_{r_n^{\mathbf{x}}}^{\mathbf{x}} + \alpha_{f_n}^{\mathbf{x}} + \alpha_{u_n^{\mathbf{x}}}^{\mathbf{x}} + \mu_e^{\mathbf{x}})$ 項移除，並將轉換函式簡化如下：

$$\hat{sd}_n^{\mathbf{y}} = \gamma_{t_n}^{\mathbf{y}} + \gamma_{q_n^{\mathbf{y}}}^{\mathbf{y}} + \gamma_{s_n}^{\mathbf{y}} + \gamma_{u_n^{\mathbf{y}}}^{\mathbf{y}} + \mu_d^{\mathbf{y}} \tag{3-35}$$

$$\hat{se}_n^{\mathbf{y}} = \alpha_{t_n}^{\mathbf{y}} + \alpha_{r_n^{\mathbf{y}}}^{\mathbf{y}} + \alpha_{f_n}^{\mathbf{y}} + \alpha_{u_n^{\mathbf{y}}}^{\mathbf{y}} + \mu_e^{\mathbf{y}} \tag{3-36}$$

在轉換時，為了要估計 $\hat{sd}_n^{\mathbf{y}}$ 與 $\hat{se}_n^{\mathbf{y}}$，仍需要知道語言參數 t_n、s_n、f_n，以及韻律狀態標記 $q_n^{\mathbf{y}}$ 與 $r_n^{\mathbf{y}}$，而這些資訊需要經由來源語者預估；其中，所有轉換聲音之語言參數，必定與來源語者一致，因此直接以來源語者之 t_n、s_n、f_n 取代；對於韻律狀態的轉換，在此直接假設：

$$q_n^{\mathbf{y}} = q_n^{\mathbf{x}}, \quad r_n^{\mathbf{y}} = r_n^{\mathbf{x}} \tag{3-37}$$

同樣考慮到來源及目標語者之音長與能量韻律狀態並非固定為一對一之對應關係，因此以高斯正規化的方式，對來源韻律狀態做線性轉換，並尋找轉換後的狀態值，最接近哪一個目標韻律狀態碼字，如下數學式所示：

$$\hat{q}_n^{\mathbf{y}} = \arg\min_i \left(\frac{\gamma_{q_n^{\mathbf{x}}}^{\mathbf{x}}}{\sigma_q^{\mathbf{x}}} \sigma_q^{\mathbf{y}} - \gamma_{q_n=i}^{\mathbf{y}} \right)^2 \tag{3-38}$$

$$\hat{r}_n^{\mathbf{y}} = \arg\min_i \left(\frac{\alpha_{r_n^{\mathbf{x}}}^{\mathbf{x}}}{\sigma_r^{\mathbf{x}}} \sigma_r^{\mathbf{y}} - \alpha_{r_n=i}^{\mathbf{y}} \right)^2 \tag{3-39}$$

其中 σ_q 與 σ_r 分別表示訓練語料中音節長度與音節能量扣除了韻律狀態以外之影響因素(Affecting Factor)的標準差，如下數學式所述：

$$\sigma_q^{\mathbf{x}} = Std(sd_n^{\mathbf{x}} - \gamma_{t_n}^{\mathbf{x}} - \gamma_{s_n}^{\mathbf{x}} - \gamma_{u_n^{\mathbf{x}}}^{\mathbf{x}} - \mu_d^{\mathbf{x}}) \tag{3-40}$$

$$\sigma_q^{\mathbf{y}} = Std(sd_n^{\mathbf{y}} - \gamma_{t_n}^{\mathbf{y}} - \gamma_{s_n}^{\mathbf{y}} - \gamma_{u_n^{\mathbf{y}}}^{\mathbf{y}} - \mu_d^{\mathbf{y}}) \tag{3-41}$$

$$\sigma_r^{\mathbf{x}} = Std(se_n^{\mathbf{x}} - \alpha_{t_n}^{\mathbf{x}} - \alpha_{f_n}^{\mathbf{x}} - \alpha_{u_n^{\mathbf{x}}}^{\mathbf{x}} - \mu_e^{\mathbf{x}}) \tag{3-42}$$

$$\sigma_r^{\mathbf{y}} = Std(se_n^{\mathbf{y}} - \alpha_{t_n}^{\mathbf{y}} - \alpha_{f_n}^{\mathbf{y}} - \alpha_{u_n^{\mathbf{y}}}^{\mathbf{y}} - \mu_e^{\mathbf{y}}) \tag{3-43}$$

值得一提的是，對於語句層次之影響因素 $\gamma_{u_n^{\mathbf{y}}}^{\mathbf{y}}$ 與 $\alpha_{u_n^{\mathbf{y}}}^{\mathbf{y}}$，因每段語句錄音之說話速度與能量較沒有一致性，故很難在來源與目標語者之間找出彼此相關性；因此，我們假設預測語句層次影響因素 $\hat{\gamma}_{u_n^{\mathbf{y}}}^{\mathbf{y}}$ 與 $\hat{\alpha}_{u_n^{\mathbf{y}}}^{\mathbf{y}}$ 以訓練語料所估計得到每一句 $\gamma_{u_n^{\mathbf{y}}}^{\mathbf{y}}$ 與 $\alpha_{u_n^{\mathbf{y}}}^{\mathbf{y}}$ 的總平均值來表示之，即：

$$\hat{\gamma}_{u_n^{\mathbf{y}}}^{\mathbf{y}} = \frac{1}{N} \sum_{n=1}^{N} \gamma_{u_n^{\mathbf{y}}}^{\mathbf{y}} \tag{3-44}$$

$$\hat{\alpha}_{u_n^{\mathbf{y}}}^{\mathbf{y}} = \frac{1}{N} \sum_{n=1}^{N} \alpha_{u_n^{\mathbf{y}}}^{\mathbf{y}} \tag{3-45}$$

最後，綜合 3-35 至 3-45 式，可將轉換函式改寫為：

$$\hat{sd}_n^{\mathbf{y}} = \gamma_{t_n}^{\mathbf{y}} + \gamma_{\hat{q}_n^{\mathbf{y}}}^{\mathbf{y}} + \gamma_{s_n}^{\mathbf{y}} + \hat{\gamma}_{u_n^{\mathbf{y}}}^{\mathbf{y}} + \mu_d^{\mathbf{y}} \tag{3-46}$$

$$\hat{se}_n^{\mathbf{y}} = \alpha_{t_n}^{\mathbf{y}} + \alpha_{\hat{r}_n^{\mathbf{y}}}^{\mathbf{y}} + \alpha_{f_n}^{\mathbf{y}} + \hat{\alpha}_{u_n^{\mathbf{y}}}^{\mathbf{y}} + \mu_e^{\mathbf{y}} \tag{3-47}$$

3.4.2 音長與能量轉換方法二

運用 3.3.2 節之概念，同樣也可將音長以及能量模型以 soft decision 之方式轉換。為了方便起見，我們僅推導音長的轉換函式，能量的轉換函式推導與音長轉換函式推導相同。以 MMSE 為準則推導如下：

$$\hat{sd}_n^{\mathrm{y}} = E[sd_n^{\mathrm{y}} \mid sd_n^{\mathrm{x}}] = \int sd_n^{\mathrm{y}} P(sd_n^{\mathrm{y}} \mid sd_n^{\mathrm{x}}) d(sd_n^{\mathrm{y}}) \tag{3-48}$$

其中，

$$P(sd_n^{\mathrm{y}} \mid sd_n^{\mathrm{x}}) = \sum_{\mathbf{q}^{\mathrm{y}}} P(sd_n^{\mathrm{y}}, \mathbf{q}^{\mathrm{y}} \mid sd_n^{\mathrm{x}}, u_n^{\mathrm{y}}, \mathbf{q}^{\mathrm{x}^*}, \mathbf{B}^{\mathrm{x}^*}, \mathbf{L}) \text{ ，} \tag{3-49}$$

$\mathbf{L}=\{t_n,\ s_n\}$，進一步將 3-49 式簡化如下：

$$\begin{aligned} & P(sd_n^{\mathrm{y}}, \mathbf{q}^{\mathrm{y}} \mid sd_n^{\mathrm{x}}, u_n^{\mathrm{y}}, \mathbf{q}^{\mathrm{x}^*}, \mathbf{B}^{\mathrm{x}^*}, \mathbf{L}) \\ &= P(sd_n^{\mathrm{y}} \mid \mathbf{q}^{\mathrm{y}}, sd_n^{\mathrm{x}}, u_n^{\mathrm{y}}, \mathbf{q}^{\mathrm{x}^*}, \mathbf{B}^{\mathrm{x}^*}, \mathbf{L}) P(\mathbf{q}^{\mathrm{y}} \mid sd_n^{\mathrm{x}}, u_n^{\mathrm{y}}, \mathbf{q}^{\mathrm{x}^*}, \mathbf{B}^{\mathrm{x}^*}, \mathbf{L}) \\ &\approx P(sd_n^{\mathrm{y}} \mid q_n^{\mathrm{y}}, t_n, s_n, u_n^{\mathrm{y}}) P(\mathbf{q}^{\mathrm{y}} \mid \mathbf{q}^{\mathrm{x}^*}, \mathbf{B}^{\mathrm{x}^*}) \end{aligned} \tag{3-50}$$

其中，$P(sd_n^{\mathrm{y}} \mid \gamma_n^{\mathrm{y}}, t_n, s_n, u_n^{\mathrm{y}})$ 為目標音節長度模型。將 3-49 及 3-50 式代入 3-48 式中，可進一步推導為：

$$\begin{aligned} \hat{sd}_n^{\mathrm{y}} &= \sum_{\mathbf{q}^{\mathrm{y}}} P(\mathbf{q}^{\mathrm{y}} \mid \mathbf{q}^{\mathrm{x}^*}, \mathbf{B}^{\mathrm{x}^*}) \int sd_n^{\mathrm{y}} P(sd_n^{\mathrm{y}} \mid q_n^{\mathrm{y}}, t_n, s_n, u_n^{\mathrm{y}}) d(sd_n^{\mathrm{y}}) \\ &\approx \sum_{\mathbf{q}^{\mathrm{y}}} P(q_n^{\mathrm{y}} \mid q_1^{\mathrm{y},n-1}, \mathbf{q}^{\mathrm{x}^*}, \mathbf{B}^{\mathrm{x}^*}) E[sd_n^{\mathrm{y}} \mid q_n^{\mathrm{y}}, t_n, s_n, u_n^{\mathrm{y}}] \end{aligned} \tag{3-51}$$

在 3-51 式之 $P(q_n^{\mathrm{y}} \mid q_1^{\mathrm{y},n-1}, \mathbf{q}^{\mathrm{x}^*}, \mathbf{B}^{\mathrm{x}^*})$ 可以用遞迴的方式得到：

$$\begin{aligned} & P(q_n^{\mathrm{y}} \mid q_1^{\mathrm{y},n-1}, \mathbf{q}^{\mathrm{x}^*}, \mathbf{B}^{\mathrm{x}^*}) \\ &= \begin{cases} P(q_1^{\mathrm{y}} \mid q_1^{\mathrm{x}^*,2}, B_1^{\mathrm{x}^*}) & n=1 \\ \sum_{q_{n-1}^{\mathrm{y}}} \left\{ \begin{matrix} P(q_{n-1}^{\mathrm{y}} \mid q_1^{\mathrm{y},n-1}, \mathbf{q}^{\mathrm{x}^*}, \mathbf{B}^{\mathrm{x}^*}) \times \\ P(q_n^{\mathrm{y}} \mid q_{n-1}^{\mathrm{y}}, q_{n-1}^{\mathrm{x}^*,n+1}, B_n^{\mathrm{x}^*}) \end{matrix} \right\} & 2 \le n \le N \end{cases} \end{aligned} \tag{3-52}$$

其中，$P(q_n^{\mathbf{y}}|q_{n-1}^{\mathbf{y}},q_{n-1}^{\mathbf{x}^*,n+1},B_n^{\mathbf{x}^*})$ 藉由前一個轉換的韻律狀態 $q_{n-1}^{\mathbf{y}}$，以及來源音節相鄰的韻律標記 $q_{n-1}^{\mathbf{x}^*,n+1}$ 與 $B_n^{\mathbf{x}^*}$，預估現在轉換音節的韻律狀態 $q_n^{\mathbf{y}}$。由於在 2.2.1 節中假設音長模型為一個高斯分佈，因此可將 3-51 式之轉換函式寫為：

$$\hat{sd}_n = \sum_{q^{\mathbf{y}}} P(q_n^{\mathbf{y}} \mid q_1^{\mathbf{y},n-1},\mathbf{q}^{\mathbf{x}^*},\mathbf{B}^{\mathbf{x}^*})(\gamma_{t_n}^{\mathbf{y}} + \gamma_{q_n^{\mathbf{y}}}^{\mathbf{y}} + \gamma_{s_n}^{\mathbf{y}} + \hat{\gamma}_{u_n^{\mathbf{y}}}^{\mathbf{y}} + \mu_d^{\mathbf{y}}) \tag{3-53}$$

同理，經由上述相同的推導過程，音節能量的轉換函式可寫為：

$$\hat{se}_n = \sum_{r^{\mathbf{y}}} P(r_n^{\mathbf{y}} \mid r_1^{\mathbf{y},n-1},\mathbf{r}^{\mathbf{x}^*},\mathbf{B}^{\mathbf{x}^*})(\alpha_{t_n}^{\mathbf{y}} + \alpha_{r_n^{\mathbf{y}}}^{\mathbf{y}} + \alpha_{f_n}^{\mathbf{y}} + \hat{\alpha}_{u_n^{\mathbf{y}}}^{\mathbf{y}} + \mu_e^{\mathbf{y}}) \tag{3-54}$$

其中，$\hat{\gamma}_{u_n^{\mathbf{y}}}^{\mathbf{y}}$ 與 $\hat{\alpha}_{u_n^{\mathbf{y}}}^{\mathbf{y}}$ 的估計方式同 3-44 與 3-45 式。

第四章 實驗結果與分析

在本章中，我們以客觀與主觀的評估方式對傳統方法與所提出的韻律轉換方法做比較，實驗的轉換組別分別為：M1→M2、F1→M1、M2→F2 以及 F2→F1。首先，於客觀評量，我們以 NMSE(Normalized Mean Square Error)評估轉換後的結果，並且進一步分析各方法對不同語者說話特性(speaking style)所呈現之效能。最後，以主觀評估方式評估基頻轉換後的聲音。

4.1 實驗環境設定

實驗所使用的語料庫為中央研究院之 COSPRO-03(Mandarin Continuous Speech Prosody Corpora)語料庫[25]，包含 2 男 3 女，共五位語者所錄製的韻律平衡平行語料。此語料庫又分成直述句、感嘆句以及疑問句三大類別，本研究則以直述句做為研究的語料。五位語者中選擇兩位男性語者(語料庫中編號為 M002、M003，本論文分別以 M1 以及 M2 表示)，以及兩位女性語者 F002、F004(本論文分別以 F1 以及 F2 表示)所錄製的音檔；經由前處理(移除文句中多字或少字的語句)，每位語者分別使用 757 個音檔，總共 24369 個音節，音檔皆為 16kHz 之取樣率及 16-bit 之 wav 格式，作為訓練與測試語料；此語料庫本身即附有音素、音節的切割資訊，切割資訊是由 Hidden Markov Model Tool Kit(HTK)[26]切割並經由人工手動校正；音節基頻軌跡則是由 Wavesurfer 軟體所提供的 ESPS[27](Entropic Signal Processing System)演算法，對每個音框求取基頻數值，接著再利用音節的切割資訊對每個音節的基頻軌跡取對數之後，做正交化展開抽取音節基頻軌跡參數向量，並同時計算每個音節之音節長度、能量位階，以及音節停頓長度。

最後，為了加入語言參數，在此將語料庫中所有的文字語料藉由中央研究院斷詞系統(CKIP)[28]，得到音節聲調、詞性及詞長的資訊。

4.2 基頻轉換之客觀性評估

以 NMSE 的評估標準如下：

$$\varepsilon_{\text{norm mse}} = \frac{\frac{1}{N}\sum_{n=1}^{N}\left\|\mathbf{y}_n - \hat{\mathbf{y}}_n\right\|^2}{\frac{1}{N}\sum_{n=1}^{N}\left\|\mathbf{y}_n - \mathbf{x}_n\right\|^2} \tag{4-1}$$

其中，N為測試語料中所有的音節數，$\mathbf{x}_n$、$\mathbf{y}_n$與$\hat{\mathbf{y}}_n$分別為來源、目標以及轉換後基頻軌跡參數。本實驗以5褶的交疊確核(5-fold cross-validation)對傳統方法，亦即高斯正規化轉換(*MV*)、聯合高斯混合模型轉換(*GMM*)，以及所提出的基頻轉換方法一(*M_1*)與方法二(*M_2*)作比較；值得注意的是，方法一的韻律狀態預估有兩種方式，一種是以 3-16 式預估韻律狀態(*M_1*)，另一種是以 3-17 式預估(*M_1_adv*)。對於*GMM*轉換的mixture數設定，實驗得知當mixture數為 16 時可得到最好的轉換效果。

表 4-1 展示了五種轉換方法對不同語者轉換組別的實驗結果。從表中發現，以平行語料所推導得到的兩種轉換方法，*M_2* 以及 *GMM*，轉換的效果皆比不需平行語料所推導得到的轉換方法，*MV*、*M_1* 以及 *M_1_adv* 來得好。此結果顯示，如果能有效的利用平行語料之間的相關性，將可以大幅的改善轉換的效能。此外，以平行語料為基礎的方法中，*M_2* 之轉換效果在不同轉換組別皆優於 *GMM*，而以非平行語料為基礎所推導的方法中，*M_1* 以及 *M_1_adv* 則均優於 *MV*，這也顯示了本論文提出的以韻律模型為基礎的音節基頻軌跡轉換方法，在 NMSE 的評估下確實可得到較佳之效能。比較 *M_1_adv* 與 *M_1*，結果也說明來源語者與目標語者間的韻律狀態，以 3-17 式方式去作韻律狀態間的映射，會比 3-16 式一對一映射方式，進一步改進轉換的效果。若是觀察同一性別轉換組別(F2→F1 與 M1→M2)，我們觀察到 F2 轉 F1 的 NMSE 比 M1 轉 M2 來的小很多；不同性別轉換組別(F1→M1 與 M2→F2)，F1 轉 M1 的 NMSE 比 M2 轉 F2 來的小，可能的原因在於 M1→M2 與 M2→F2 兩個轉換組別中，來源語者與目標語者的說話特性差異較大，間接影響轉換方法的效能。為了驗證此，我們在 4.3 小節將進一步對此現象做分析。

表 4-1：五種轉換方法對四組轉換組別的客觀評估(NMSE)結果

	F1→M1	**M1→M2**	**M2→F2**	**F2→F1**
MV	0.0257	1.0561	0.0647	0.1498
M_1	0.0246	0.9245	0.0584	0.1419
M_1_adv	0.0230	0.8818	0.0554	0.1206
GMM	0.0204	0.7586	0.0443	0.1186
M_2	0.0198	0.7489	0.0355	0.1049

圖 4-1 為 F2 轉 F1 的基頻軌跡轉換範例，每一列對應一種轉換方式，由上至下分別為 *M_2*、*M_1_adv*、*GMM* 以及 *MV*，而每列之粗虛線、細虛線與實線分別代表轉換、來源以及目標基頻軌跡，垂直線代表音節邊界。從最底下的轉換圖中可以發現，高斯正規化的轉換方式，只是在對來源語者的基頻軌跡做上下平移的動作，即改變其音高(pitch mean)，而其基頻軌跡形狀則與來源基頻軌跡相似；當來源與目標基頻軌跡之間的形狀差異很大時，這將會造成較大的轉換誤差；此情況在第三個轉換圖，也就是 *GMM* 的方法中則有進一步之改善。而在第一與第二個轉換圖中，發現 *M_2* 與 *M_1_adv* 方法，其轉換後的基頻軌跡形狀與目標軌跡之形狀極為相近。

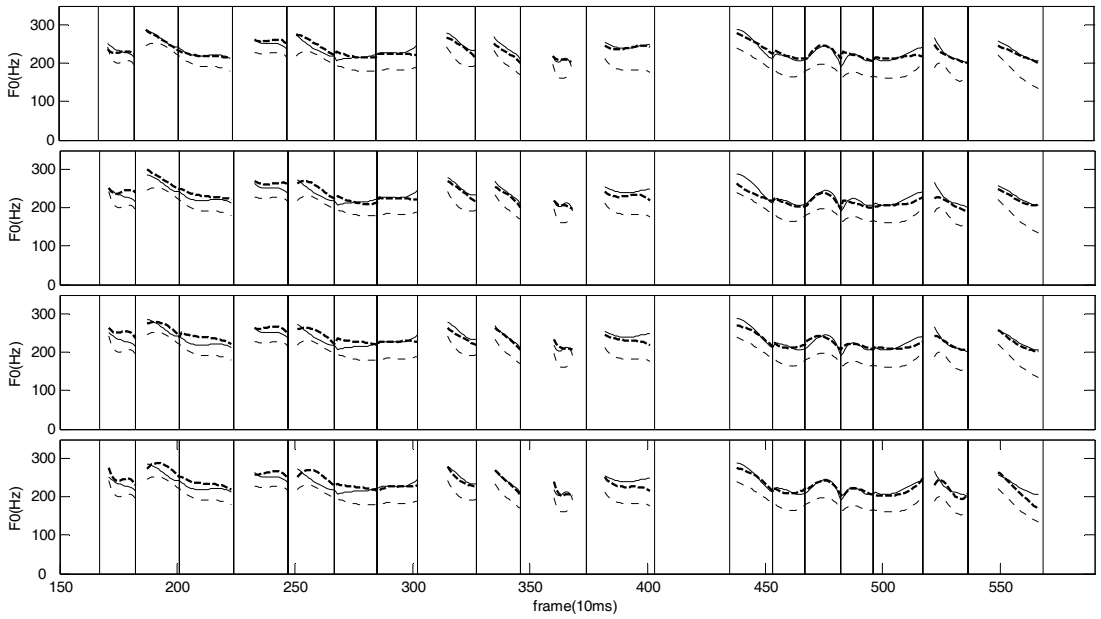

圖 4-1：F2→F1 之基頻軌跡轉換圖，內容為「著重於兼顧人文社會學科，各領域的完整性」

為了進一步證明上述的論點，表 4-2 計算了轉換後的音節基頻軌跡向量後三維轉換誤差，也就是代表基頻軌跡形狀之參數$[a_1, a_2, a_3]$的 NMSE。從表中可以看出，*MV* 轉換後的基頻軌跡形狀與目標語者基頻軌跡形狀差距最大，*M_1* 與 *M_1_adv* 則略優於 *GMM*，而 *M_2* 轉換後的基頻軌跡形狀最接近目標語者。值得一提的是，*M_1* 與 *M_1_adv* 只有在預估韻律狀態的方法上有不同，因此兩者之轉換基頻軌跡形狀並不會有所差異，只會在音節音高(pitch mean)有所變更。圖 4-2 為 F2 轉 F1 的另一範例，其線條表示如同圖 4-1，而第一與第二列分別為 *M_1* 與 *M_1_adv* 的轉換方法。

表 4-2：五種轉換方法對音節基頻軌跡形狀之 NMSE

	F1→M1	**M1→M2**	**M2→F2**	**F2→F1**
MV	0.0115	0.4365	0.0112	0.0529
M_1	0.0079	0.2655	0.0064	0.0344
M_1_adv	0.0079	0.2655	0.0064	0.0344
GMM	0.0079	0.2681	0.0067	0.0348
M_2	0.0075	0.2552	0.0052	0.0324

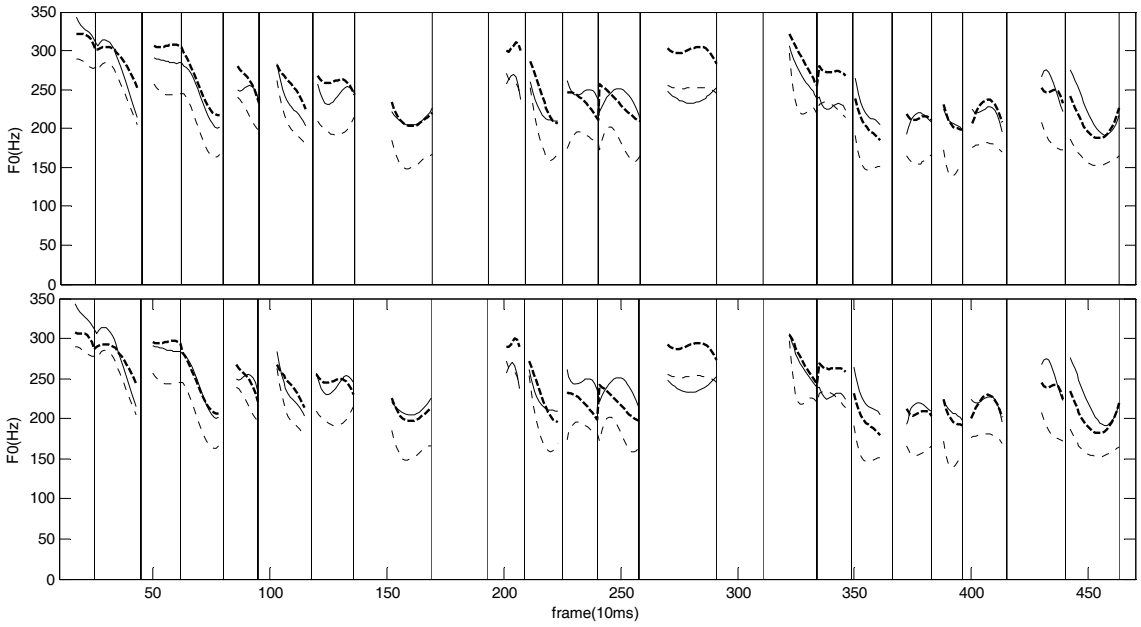

圖 4-2：F2→F1 基頻軌跡轉換圖，內容為「次日中午快下班時，他打電話說下午打牌打到七

點」

4.3 說話特性對基頻轉換影響之分析

在中文中，語者的說話特性會對基頻軌跡有很大的影響。直覺上會認為，在基頻轉換時，說話特性相近的語句會比說話特性差異大的語句有更好的轉換效能，因此本節接著分析各種轉換方式對語者說話特性的影響。此分析以 A-PLM 所提出之之中文語音韻律階層架構為基礎，如圖 2-3 所示，語句的韻律架構可視為由 SYL、PW、PPh 以及 BG/PG 四層所構成，並且以 *B*0、*B*1、*B*2-1、*B*2-2、*B*2-3、*B*3 以及 *B*4 作為邊界的區隔，故在此以停頓標記來定義說話特性；如果兩句相同的語句，有相似的停頓特性，則彼此的說話特性也應該很相似；相反的，如果來源與目標語者在同一個音節邊界上停頓標記差異很大，這意味著在此音節邊界上，來源與目標的前後音節基頻軌跡將會有截然不同的表現；舉例來說，在音節邊界上標記 *B*1，代表的是此邊界停頓很短且前後音節之連音效應影響嚴重；而如果標記的是 *B*4，則代表有較長的停頓，也會有明顯音高重置(pitch reset)的現象。

在此將六類的停頓標記分成以下三類：C_1={*B*0, *B*1, *B*2-3}，C_2={*B*2-1, *B*2-2}，以及 C_3={*B*3, *B*4}，分別代表 no break、minor break，與 major break。若 $B_n^{\mathrm{x}} = C_i$ 且 $B_n^{\mathrm{y}} = C_j$，其中 $i, j \in \{1,2,3\}$，則 B_n^{x} 與 B_n^{y} 的不一致情形可定義成以下三種：

(1) Type 1：$i = j$

(2) Type 2：$|i - j| = 1$

(3) Type 3：$|i - j| = 2$。

其中，Type 1、Type 2 與 Type 3 分別代表沒有不一致性、輕微的不一致性以及嚴重的不一致性。可以預期的是，因為來源與目標語者的說話特性相似，因此 Type 1 的情況並不會對轉換誤差有太大的影響；相反的，在 Type 2 與 Type 3 的情況下將會對轉換誤差有較大的影響。

表 4-3 顯示四組語者轉換組合其停頓標記不一致性的統計結果，從表中可看出，M1→M2 以及 M2→F2 與另外兩組語者相比，有較多 Type 2 與 Type 3 的情況，因此可以預期，M1→M2 與 M2→F2 這兩組轉換組合，其來源與目標語者的說話特性有較嚴重的不一致性。比較表 4-1 與 4-3，並分別比較相同性別與不同性別之轉換，可以發現在 M1→M2 與 F2→F1 的比較中，M1→M2 的 Type 2 與 Type 3 較多，使得其 MSE 也較大；同樣也可在 F1→M1 與 M2→F2 的比較上發現此一現象。

表 4-3：四組語者轉換組合之停頓標記不一致性統計結果(%)

	F1→M1	**M1→M2**	**M2→F2**	**F2→F1**
Type 1	83.4	76.7	75.9	82.9
Type 2	15.5	20.6	21.2	16
Type 3	1.1	2.7	2.9	1.1

若從另一個觀點來看，計算這四組轉換組合的音節基頻軌跡參數之相關係數(correlation coefficient)，其值越接近 0 則表示彼此線性相關性越小。表 4-4 為每組語者轉換各別維度之相關係數之值，由表中可知，M1→M2 以及 M2→F2 這兩組的相關係數值都比另外兩組語者轉換來得低，這也再次證實 M1→M2 以及 M2→F2 的說話特性有較嚴重的不一致性。

表 4-4：四組語者轉換組合之相關係數

	F1→M1	**M1→M2**	**M2→F2**	**F2→F1**
a_0	0.8	0.63	0.54	0.81
a_1	0.56	0.38	0.42	0.62
a_2	0.43	0.25	0.28	0.47
a_3	0.18	0.14	0.14	0.23

接著觀察來源與目標語者在不同 Type 情況下，其音節基頻軌跡參數的差異度。定義如下：

$$e_k^{ty,pr} = (yr_k^{ty,pr} - xr_k^{ty,pr})^2 \tag{4-2}$$

其中，$k=1\sim N_{ty,pr}$，$N_{ty,pr}$ 為在某一特定語者轉換組別 *pr* (pair)，不同之停頓標記不一致性情形 *ty* (Type)，其總共的音節邊界數目；而 $yr_k^{ty,pr}$ 與 $xr_k^{ty,pr}$ 分別為在 *pr* 以及 *ty* 的情形下，目標音節以及來源音節在第 *k* 個音節邊界之音高重置程度，如圖 4-3 所示。

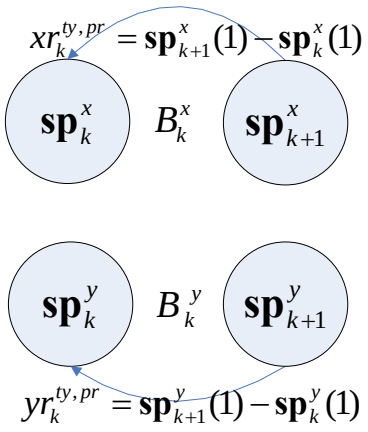

圖 4-3：說話特性差異示意圖

圖 4-4 顯示四組語者轉換組合在三種 Type 情況下，$e_k^{ty,pr}$ 的 95% confidence interval；從圖中可以發現，Type 3 的參數差異度最大，Type 2 次之，而 Type 1 差異度最小。

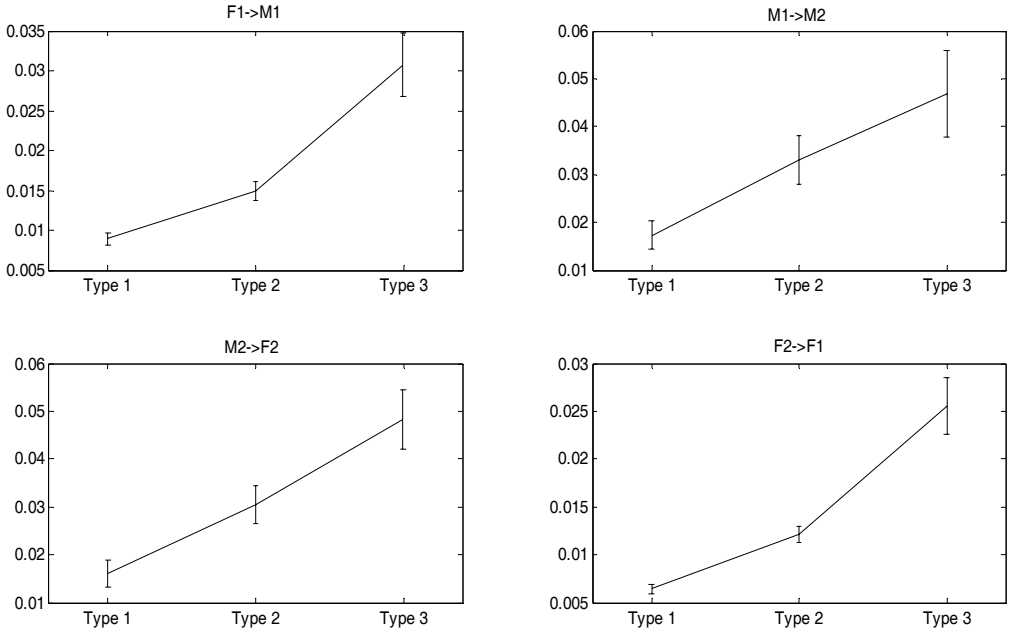

圖 4-4：四組語者轉換組合在三種 Type 情況下之 95% confidence interval

此結果顯示，停頓標記的不一致性越大，則來源與目標語者在音節間的音高重置程度差異也越大，因此說話特性的差異也越大。進一步將此圖對照表 4-3，可發現有較多 Type 2 與 Type 3 的轉換組合，其音高重置程度之差異也會隨著提高。

最後，藉由上述對於說話特性的定義，檢驗停頓標記不一致性對各基頻轉換方法的誤差影響。在語者轉換組別 *pr* ，且第 *k* 個音節邊界之停頓標記不一致類型為 *ty* 時，基頻轉換的誤差定義如下：

$$ce_k^{ty,pr} = \left\| \mathbf{y}_k^{ty,pr} - \hat{\mathbf{y}}_k^{ty,pr} \right\|^2 \tag{4-3}$$

其中，$k = 1 \sim N_{ty,pr}$，$\mathbf{y}_k^{ty,pr}$ 以及 $\hat{\mathbf{y}}_k^{ty,pr}$ 分別代表在語者轉換組別 *pr*，停頓標記不一致類型為 *ty* 時，第 *k* 個音節邊界右邊的目標以及轉換音節基頻軌跡參數，如圖 4-5 所示。

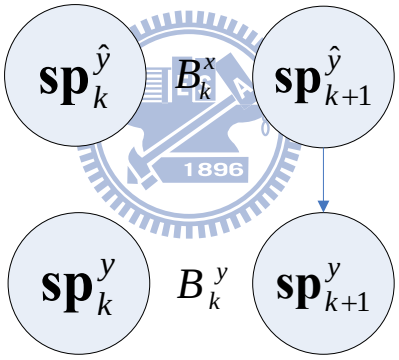

圖 4-5：基頻轉換誤差定義示意圖

由於 *M_1_adv* 與 *M_1* 轉換結果差異不大，故在此只比較 *M_2*、*M_1_adv*、*GMM* 以及 *MV* 的轉換效能。圖 4-6 顯示此四組轉換方法其轉換誤差累積分布函數(cumulative distribution function)，結果顯示 *M_2* 轉換方法在三種停頓標記不一致的情況下，對四組轉換組合都能夠有最小的基頻轉換誤差，*MV* 之轉換誤差為最大，而 *GMM* 的表現均優於 *M_1_adv*，此結果顯示不論語者說話特性差異為何，*M_2* 均能有優異的表現。

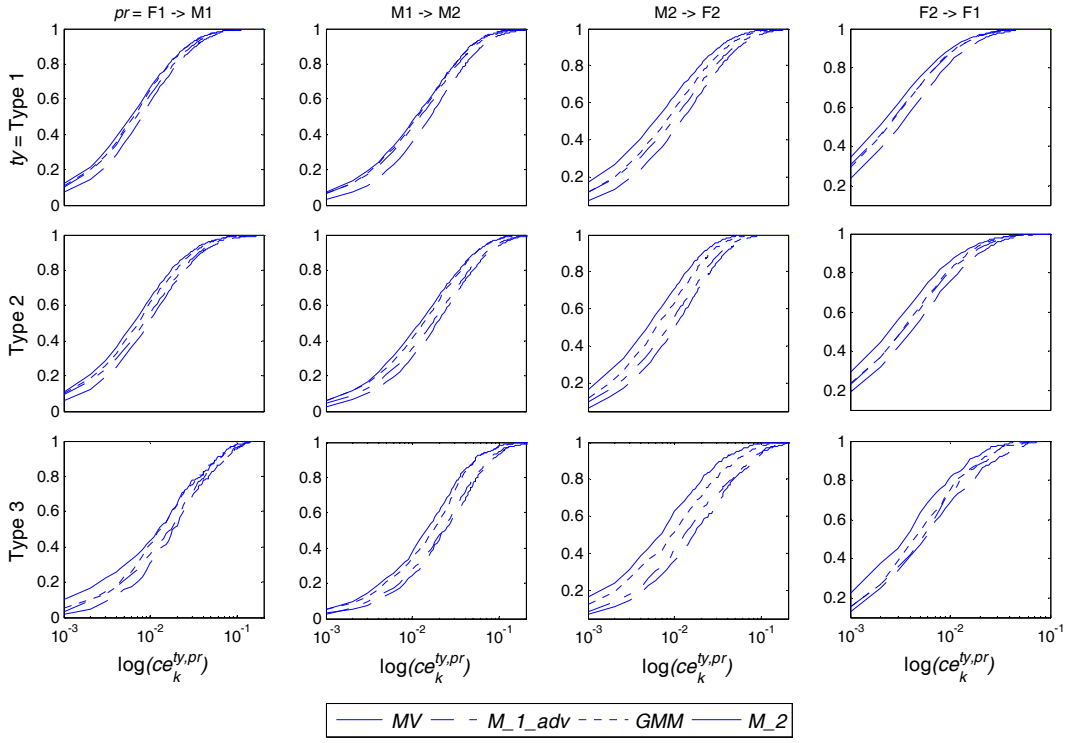

圖 4-6：四種轉換方法對於三種停頓標記不一致情形的轉換誤差 cdf

4.4 音節長度與能量轉換之客觀性評估

此小節同樣也以客觀評估方式，評量各方法對音長以及能量之轉換效能，評量方式以 NMSE 評估：

$$\varepsilon_{\text{norm mse}} = \frac{\frac{1}{N}\sum_{n=1}^{N}(\mathbf{y}_n - \hat{\mathbf{y}}_n)^2}{\frac{1}{N}\sum_{n=1}^{N}(\mathbf{y}_n - \mathbf{x}_n)^2} \tag{4-4}$$

其中 N 為測試語料所有的音節數目，$\mathbf{y}_n$、$\mathbf{x}_n$ 與 $\hat{\mathbf{y}}_n$ 分別代表目標、來源以及轉換後的音長與能量，其單位分別為秒(sec)以及分貝(dB)。以下比較 *MV*、*GMM*，以及 3-37 式為基礎的轉換方法(*M_1*)、以 3-38 與 3-39 式為基礎之轉換方法(*M_1_adv*)以及方法二(*M_2*)，另外，對於 *GMM* 轉換的 mixture 數設定為 16。音節長度與能量轉換結果分別如表 4-5 與 4-6 所示。

表 4-5：五種轉換方法對四組轉換組合的音節長度客觀評估結果

	F1→M1	**M1→M2**	**M2→F2**	**F2→F1**
MV	0.9949	0.9539	1.0567	0.6642
M_1	1.3597	1.7395	0.8069	0.5695
M_1_adv	1.0141	0.9176	0.8830	0.5837
GMM	0.8375	0.6876	0.7400	0.5344
M_2	0.7216	0.6620	0.6631	0.4144

首先，我們分析各方法在音節轉換之效能。由表 4-5 得知，*M_1* 在音節長度轉換的效能僅有兩個組別(M2→F2 及 F2→F1)優於傳統方法 *MV*，而 *M_1_adv* *僅有*在組別 F1→M1 略遜於傳統方法；然而，此方法在轉換組別(M2→F2 及 F2→F1)並無法改進 *M_1*。以平行語料為基礎的方法 *GMM*、*M_2* 則皆明顯優於其餘以非平行語料之轉換方法。而 *M_2* 在不同的轉換組別所呈現的轉換效能都優於其他轉換方法。

表 4-6：五種轉換方法對四組轉換組合的音節能量位階客觀評估結果

	F1→M1	**M1→M2**	**M2→F2**	**F2→F1**
MV	0.4920	0.9192	0.8176	0.8667
M_1	0.5292	0.6657	0.5851	0.5235
M_1_adv	0.5254	0.6037	0.5862	0.4897
GMM	0.3720	0.6593	0.6033	0.6768
M_2	0.3632	0.5409	0.4725	0.4509

接著，我們比較各方法在音節能量位階轉換之效能。如表 4-6 所示，*M_1* 、*M_1_adv* 除了在轉換組別 F1→M1 略遜於傳統方法外，在其餘轉換組別則皆優於 *MV* 。此外，*M_1_adv* 除了在 M2→F2 略遜於 *M_1*，在其它組別中，皆可觀察到 *M_1_adv* 改進了 *M_1*。值得注意的是，*GMM* 方法在 M1→M2、M2→F2、F2→F1 之轉換效果比 *M_1_adv* 來得不理想。與基頻軌跡轉換以及音節音長轉換結果相同，*M_2* 轉換效果皆優於其他方法。此一結果也證實，本研究所提出的音節長度與能量轉換，即類比基頻軌跡轉換的方法，也能呈現出不錯的效能。

4.5 主觀性評估

在本小節做了相似度以及喜好度兩項主觀性實驗。為了要單獨評估基頻轉換方法之效能，在此將轉換後的基頻軌跡與目標聲音原本之頻譜、音節長度，以及能量使用STRAIGHT[23]合成器產生轉換聲音。本實驗測試人員為 10 人；首先在測試語句中挑選 20 句較長之語句，並將其中 10 句用於相似度實驗，剩下 10 句用於喜好度實驗；在這兩個實驗語料中平均音節數目分別為 36.4 以及 41.5 個音節；相似度實驗之最長語句音節數為 45 字，最短語句音節數為 32 字，而在喜好度實驗則分別為 61 字與 32 字。在此實驗中，以 *M_2*、*M_1_adv*、*GMM* 以及 *MV* 評估了 M2→F2 的結果。在喜好度實驗，將這四種轉換方法所產生的聲音，以隨機的順序讓測試人員聽，每位測試人員被要求必須從四種轉換方法中，挑選聽起來最自然的語句，而其所對應到的方法就可得到 1 分；如果測試人員在一句測試語句中難以抉擇，則可以選擇兩種聽起來最自然的語句作為答案，而其所對應到的方法將可各得 0.5 分。在相似度測試中，將未知方法的轉換聲音以及其對應的合成目標聲音讓測試人員聽，而測試人員必須從四種轉換方法中選擇與目標聲音最相似之答案，如果測試人員在一句測試語句中難以抉擇，則測試人員將可以選擇兩種聽起來最相似的聲音作為答案，其記分方式如同喜好度實驗。

表 4-7：M2→F2 主觀性評估結果

	MV	***GMM***	***M_1_adv***	***M_2***
相似度(%)	13	16.5	26	44.5
喜好度(%)	11.5	18.5	15.5	54.5

表 4-7 為評估之實驗結果，從表中可以發現，*M_2* 不論是在相似度或是喜好度實驗中，皆有最好的效果，而 *MV* 的效果為最差；*M_1_adv* 在相似度實驗中表現優於 *GMM*，而在喜好度實驗中卻剛好相反。

第五章 結論與未來展望

本論文提出以韻律模型為基礎的中文韻律轉換方法，不同於傳統的韻律轉換，我們以韻律模型描述韻律資訊，並將韻律參數拆解成各個影響因素，建立來源以及目標韻律模型各影響因素間的對應關係。實驗結果證實，在客觀與主觀評估中，以平行語料為基礎的轉換方法，*M_2* 之效果明顯優於以高斯混合模型為基礎的轉換；而以非平行語料為基礎的轉換方法之比較，*M_1* 亦優於高斯正規化轉換。我們也更進一步的定義語者的說話特性，發現在各種語者間說話特性不一致的情況下，本論文所提出之方法仍優於傳統轉換方法。

語者說話的停頓時間長短，亦是韻律上的一個重要參數，因此未來有必要將來源語者的說話停頓特性，一併以韻律模型為基礎做轉換，以期能有更完整的韻律轉換架構。而要精確的以韻律模型描繪語者的韻律資訊，必須要有大量的訓練語料，這在聲音轉換技術上是很不利的。因此未來若能夠利用來源韻律模型對少量的目標語料做語者調適，並建立目標韻律模型，將能使此方法更具實用性。

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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] M. Abe, S. Nakamura, K. Shikano and H. Kuwabara, “Voice conversion through vector Quantization,”in *Proc.of ICASSP*, New York, NY, USA, pp. 655–658, Apr. 1988.

[9] T. Toda, A.W. Black, and K. Tokuda, “Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory,” *IEEE Trans. Audio, Speech and Language*

Processing, Vol. 15, No. 8, pp. 2222–2235, Nov. 2007.

[10] 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.

[11] 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,

[12] T. En-Najjary, O. Rosec, T. Chonavel, "A Voice Conversion Method Based on Joint Pitch and Spectral Envelope Transformation," *Proc. of Interspeech*, pp.1225–1228, Oct. 2004.

[13] Z. Hanzlicek and J. Matousek, "F0 Transformation within the Voice Conversion Framework," *Proc. of Interspeech*, pp.1961–1964, Aug. 2007.

[14] C. H. Lee, C. C. Hsia, C. H. Wu and M. C. Lin, "Regression-Based Clustering for Hierarchical Pitch Conversion." in *Proc. of ICASSP*, pp. 3593–3596, Taipei, Taiwan, April 2009.

[15] O. Turk, "New Methods for Voice Conversion," Master Degree Thesis of Science. Bogazici University, 2003.

[16] B. Gillet, and S. King, "Transforming F0 Contours", Proceedings of Eurospeech 2003, pp. 101–104.

[17] Z. Inanoglu, "Transforming Pitch in a Voice Conversion Framework", Master thesis, St. Edmund's College, University of Cambridge, 2003.

[18] G. Y. Zuo, Y. Chen, X. G. Ruan and W. J. Liu, "Learning Mandarin Tone Mapping Codebook for Voice Conversion," *Proc. of ICMLC*, pp. 4824–4828, Guangzhou, China, Aug. 2005.

[19] Y. Xu and Q. E. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech Commun.*, vol. 33, pp. 319–337, 2001.

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

[21] C. Y. Chiang, S. H. Chen, H. M. Yu, and Y. R. Wang, “Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech,” *J. Acoust. Soc. Am.*, Vol. **125**, No. **2**, pp. 1164–1183(2009).

[22] C. Y. Chiang,“Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech,” Department of Communication Engineering, NCTU, Dissertation for Doctor of Philosophy, March 2009.

[23] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne≙, “Restructuring Speech Representations Using a Pitch Adaptive Time-Frequency-Based F0 Extraction: Possible Role of a Repetitive Structure in Sounds,” Speech Communication, vol.27, no.3-4, pp.187–207, Apr.1999.

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

[25] 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).

[26] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P.Woodland, The HTK Book Version 3.0. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[27] K. Sj¨olander and J. Beskow. Wavesurfer - an open source speech tool. In Proceedings of the International Conference on Spoken Language Processing, 2000.

[28] The CKIP on-line word segmentation system. Available: http://ckipsvr.iis.sinica.edu.tw/