國 立 交 通 大 學

電信工程學系

碩士論文

使用聲調模型輔助之基頻偵測器與國語連續語音聲調辨認

Pitch Detection with Tone Model and Tone Recognition in Mandarin Speech

研 究 生：李鴻彥

指導教授：陳信宏　博士

中華民國九十五年八月

使用聲調模型輔助之基頻偵測器

與國語連續語音聲調辨認

Pitch Detection with Tone Model

and Tone Recognition in Mandarin Speech

研 究 生：李鴻彥　　Student：Hong-Yan Lee

指導教授：陳信宏　　Advisor：Dr. Sin-Horng Chen

A Thesis
Submitted to Department of Communication Engineering
College of Electrical Engineering and Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in Electrical Engineering

August 2006
Hsinchu, Taiwan, Republic of China

中華民國九十五年八月

使用聲調模型輔助之基頻偵測器與國語連續語音聲調辨認

研究生：李鴻彥　　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

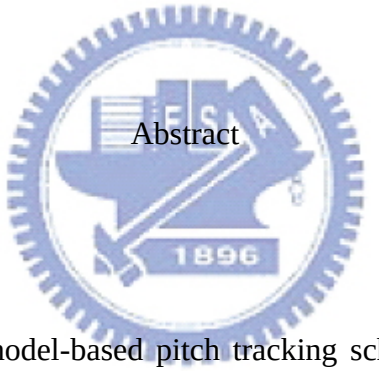
在本論文中，我們提出了使用聲調模型來輔助基頻軌跡的估測，並在估測軌跡的同時辨認每個音節的聲調。以往的基頻軌跡估測常發生半頻與倍頻錯誤，於是我們提出使用具有統計性的聲調模型輔助軌跡估測，希望可以利用模型的統計特性，使得半頻與倍頻錯誤的發生減少，在實驗中我們發現，使用模型輔助估測軌跡的方式，確實可以減少上述兩種錯誤。以往在聲調辨認方面，均是先求取基頻軌跡，然後利用估測後的基頻軌跡進行聲調辨認，而在本文中提出一種架構，此架構可以同時估測基頻軌跡與聲調辨認。

Pitch Detection with Tone Model and Tone Recognition in Mandarin Speech

Student : Hong-Yan Lee Advisor : Dr. Sin-Horng Chen

Institute of Communication Engineering

National Chiao Tung University



Abstract

In this thesis, a new model-based pitch tracking scheme is proposed. It tracks pitch and recognizes tones simultaneously using a statistical prosody model of Mandarin speech. With the guide of the prosody model, the pitch tracking can be more reliable so as to reduce both half pitch error and double pitch error. Experimental results showed that the gross pitch error (GPE) of pitch detection was reduced from 1.121% to 0.918% by using the proposed pitch estimation scheme. Both half and double pitch error rates were also reduced. Meanwhile, a tone recognition rate of 70% was achieved.

誌謝

感謝我的家人，爸爸、媽媽、大姊、二姊、大哥，尤其是大姊，在這兩年不時地在新竹照顧我，因為有你們的支持，我才能完成這篇論文，我愛你們。

謝謝我的指導教授陳信宏老師與王逸如老師，感謝兩位老師這兩年的諄諄教誨，是老師讓我知道自己在做人處事上缺少了什麼，雖然說在我畢業的時候有許多事情還沒學會，但我會抱著學習的心，在未來職場上繼續學習。

這兩年內認識了許多同學與朋友，實驗室扛霸子國興、小老闆愛將見偟、朋友超多的振豐、常在角落讀書的家勇、HTK 快樂時光的世帆、豐富職場經歷的 Paul、還有從高科大開始就是同學的東毅與世哲，因為有你們，在學習的路上才會如此有趣。謝謝活潑可愛的實驗室學弟與高科大學弟妹們，你們的祝福，讓我更加勇敢的面對未來的挑戰。我還要感謝在研究上幫了我大忙的程式達人振宇學長、愛看黑澀會的阿德學長、怕老婆的智合學長、鋼琴王子希群學長、正在當國小老師的輝哥學長，因為你們的知識與經驗，讓我在研究上克服了不少難關。

最後，僅將此篇論文獻給所有關心我的人，因為你們一路上的陪伴，讓我覺得旅程不再是一個人。

目錄

圖目錄

表目錄

第一章　緒論

1.1 研究動機

目前技術而言，人機介面不再只是排滿密密麻麻按鍵的鍵盤了，不僅僅是機械上、光學上或聲學上都有重大的突破，其中以聲學的方式最具人性化，所以其重要性是難以被取代的，而且不光是電腦，在手機或售票機上早已經出現使用語音輸入的方式將資訊傳送到裝置內，所以語音辨識這方面的科技早已經受到廣大的注意，於是，正確的辨識出語音資訊的要求也隨著技術的發展而越來越高，人機介面最終的目標不外乎是達到具人性化且具高可靠度的介面，目前語音辨識這方面的技術雖然已經發展相當成熟了，但其可靠度仍舊無法讓廣大的人們接受，所以目前語音辨識技術依舊朝著提供更高可靠度的方向發展。

語音辨識相關技術已經發展了五十年了，但是始終無法提供一個完善的系統給使用者，如何發展出一個良好的語音系統，在過去數十年間，一直是各大研究機構主要的探索方向之一，傳統中文語音辨識上，目前大部分都還是屬於音節(syllable)部分的辨識，而所使用的特徵參數莫過於 MFCC 參數，若能在語音資訊中抽取到有效的特徵參數，對於辨識必定能有大大的幫助，然而對於中文這類的 tonal language 而言，聲調是辨識這類語言語意再好不過的特徵參數了，然而辨識出正確的聲調卻又是一門技術，目前聲調辨識依舊是依賴基頻軌跡來當作辨識的依據，雖然說抽取基頻軌跡的相關技術非常的多，但依舊容易發生抽取出倍頻(Double pitch)或半頻(Half pitch)的情形，造成基頻軌跡不連續，但是基本這些錯誤其實是可以避免的，於是在這提出利用聲調統計資訊，輔助基頻軌跡建立，減少半頻與倍頻錯誤的發生。

1.2 研究方向

經由統計語料庫中基頻軌跡的聲調狀態與韻律狀態，建立聲調模型、韻律模型，並利用模型輔助基頻軌跡的建立，利用音框與音框之間基頻值的候選者，同時考慮當前音節的聲調狀態與韻律狀態的可能性，挑選最佳可能的基頻軌跡，並同時辨認當前音節的聲調，達到基頻軌跡建立與聲調辨認同步。

1.3 章節概要

本論文總共分為五個章節，各章節的編排與概要如下：

第一章　緒論：描述研究動機以及研究方向。

第二章　以抽取瞬時頻率方式求取基頻：介紹以抽取音框的瞬時頻率為基礎，建立以音框為單位可靠的基頻候選者。

第三章　中文聲調辨認與基頻軌跡建立：說明聲調模型與韻律模型的訓練過程，並且介紹如何利用模型輔助基頻軌跡建立，並同時完成聲調辨認。

第四章　實驗結果與分析：利用實驗說明微調參數的過程，並分析最後基頻軌跡改善的程度，與聲調辨識率。

第五章　結論與展望：對於本論文提出的方法與實驗結果做簡要的結論。

第二章　以抽取瞬時頻率方式求取基頻

目前可見求取基頻的技術都已經相當成熟，所以我們的研究目標在於高雜訊下亦能求取準確的基頻軌跡，瞬時頻率的觀念是在 1986 年 F.J. Charpentier[1]曾經提出過相關的研究，到最近 1996 年才由 Takao Kobayashi 與 Dhany Arifianto[2][3][4][5]繼續以瞬時頻率為基礎作研究，主要是因為訊號的瞬時頻率對於雜訊影響並不敏感的原理，在高雜訊下以求取訊號的瞬時頻率的方式來估測基頻，並證實利用瞬時頻率的方式可在高雜訊下求取出準確的基頻軌跡。

2.1 瞬時頻率

我們知道一個訊號通常包含著很多頻率成份，當我們想分析一個包含許多不同頻率成份的訊號時，我們必須想辦法將這些成份分開成單一頻率的弦波，通常我們會利用典型的 phase vocoder 的架構，如下圖：

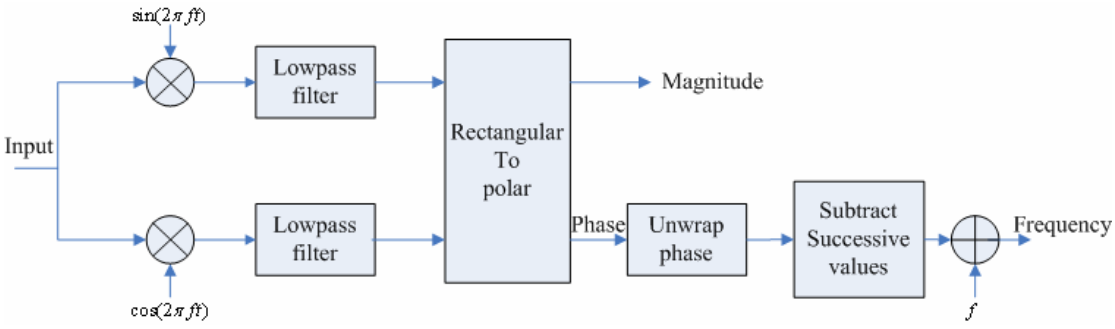


圖 2.1： Phase vocoder 架構圖

如圖利用外差的方式，將原始訊號乘上頻率為 f 的 sin 與 cos 訊號(or complex exponential)，頻譜部份原本在頻率 f 的成份會被移到頻率 0 Hz 的地方，經過低通濾波器之後，可以得到只有頻率為 f 的訊號成份，這個動作有如一 Filter

Bank(FB)般，讓我們可以只對訊號中的某一個頻率成份單獨進行分析。

假如，我們使用 window function $w(\tau)$ 對一訊號 $x(\tau)$ 在時間點 t 作 Short Time Fourier Transform (STFT)，可以如式 2.1 所示：

$$X(f,t)=\int_{-\infty}^{\infty} w(\tau-t)x(\tau)e^{-j2\pi f\tau}d\tau \tag{2.1}$$

觀察此 STFT 的積分式之後，發現此積分式可以表示成訊號 $x(t)$ 乘上一個含有 sine 跟 cosine 成份的 $e^{j2\pi ft}$，然後與 impulse response 為 $w(-t)$ 的 filter 作 convolution，如果 $w(-t)$ 是一個 lowpass function，那麼這動作與上述的 phase vocoder 的架構大致一樣，唯一不一樣的地方，就是 phase vocoder 將頻率為 f 的成份移到 0 Hz，而 STFT 並沒有，如果利用反超外差的方式(inverse heterodyne)，即可讓 FB 輸出訊號的頻率調回到原本的頻率 f，所以可以推導出 FB 輸出與 $X(f,t)$ 之間的關係為，

$$FB(f,t)=e^{j2\pi ft}X(f,t) \tag{2.2}$$

瞬時頻率的定義為，對音框其頻譜的相位作 t 微分，因此可知道 FB 在時間點 t 的瞬時頻率 λ (instantaneous frequency)為

$$\lambda(f,t)=\frac{1}{2\pi}\frac{\partial}{\partial t}\arg[FB(f,t)]=f+\frac{1}{2\pi}\frac{\partial}{\partial t}\arg[X(f,t)]\ , \tag{2.3}$$

單位為 Hz。

利用角度微分公式可以將 $\frac{\partial}{\partial t}\arg[X(f,t)]$ 化簡如下：

$$\frac{\partial}{\partial t}\arg[X(f,t)] = \frac{\mathrm{Re}[X]\mathrm{Im}[\frac{\partial X}{\partial t}] - \mathrm{Im}[X]\mathrm{Re}[\frac{\partial X}{\partial t}]}{|X|^2} \tag{2.4}$$

其中 $X = X(f,t)$，

$$\frac{\partial X(f,t)}{\partial t} = \int_{-\infty}^{\infty} -w'(\tau - t)x(\tau)e^{-j2\pi f(\tau - t)}d\tau \quad ， \quad w'(t) = \frac{dw(t)}{dt}$$

將瞬時頻率 $\lambda(f,t)$ 以 f 軸畫出，若訊號 $x(t)$ 含有週期的成份，則 $\lambda(f,t)$ 會有階梯狀的特性曲線出現，如下圖 2.2 中的藍點所示，上圖為 voiced 音框的頻譜，下圖則是此音框所求得的瞬時頻率，上下比對後，很明顯可以發現在此頻譜 harmonic 出現的地方對應到瞬時頻率上，同一個位置出現了階梯平台。

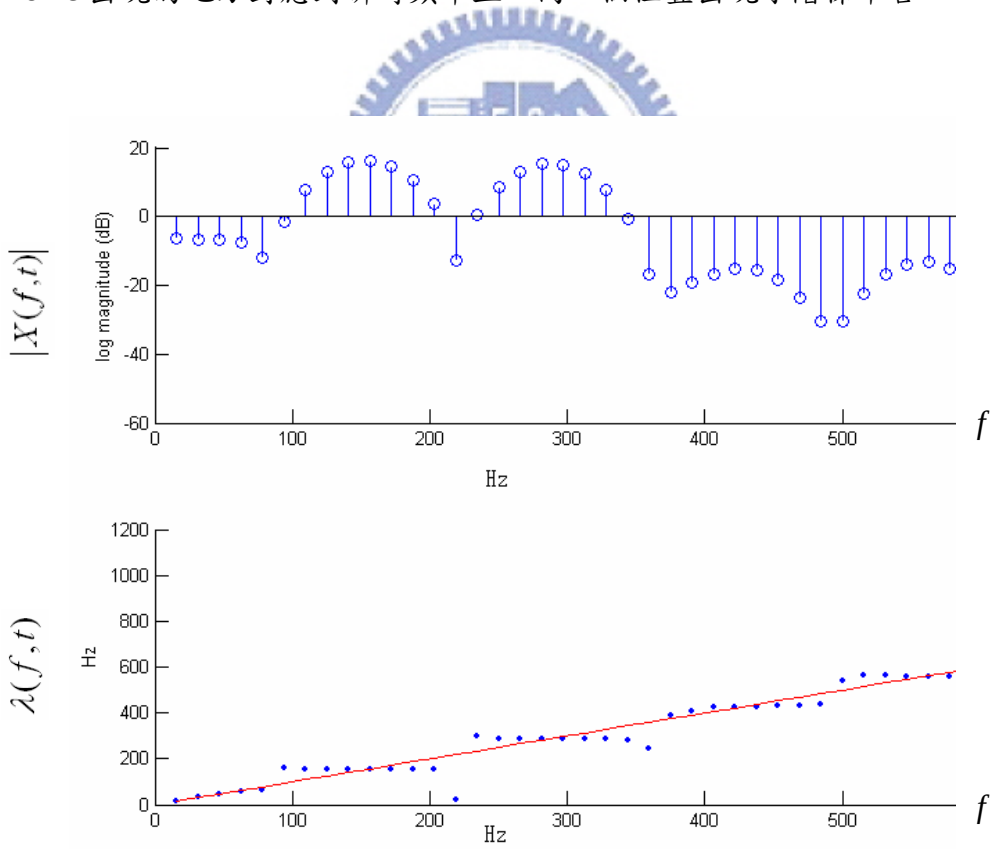


圖 2.2 An example of instantaneous frequency of voiced signal

接著觀察 unvoiced 音框所求得的瞬時頻率，如圖 2.3，上圖為此音框的頻譜，

從頻譜上來看，unvoiced 音框能量並沒有 voiced 音框大，而且並沒有很明顯的 harmonic 的成分存在，所以其瞬時頻率並不會出現如同 voiced 音框的階梯曲線(藍色點)，反而像是一條不規則的線。

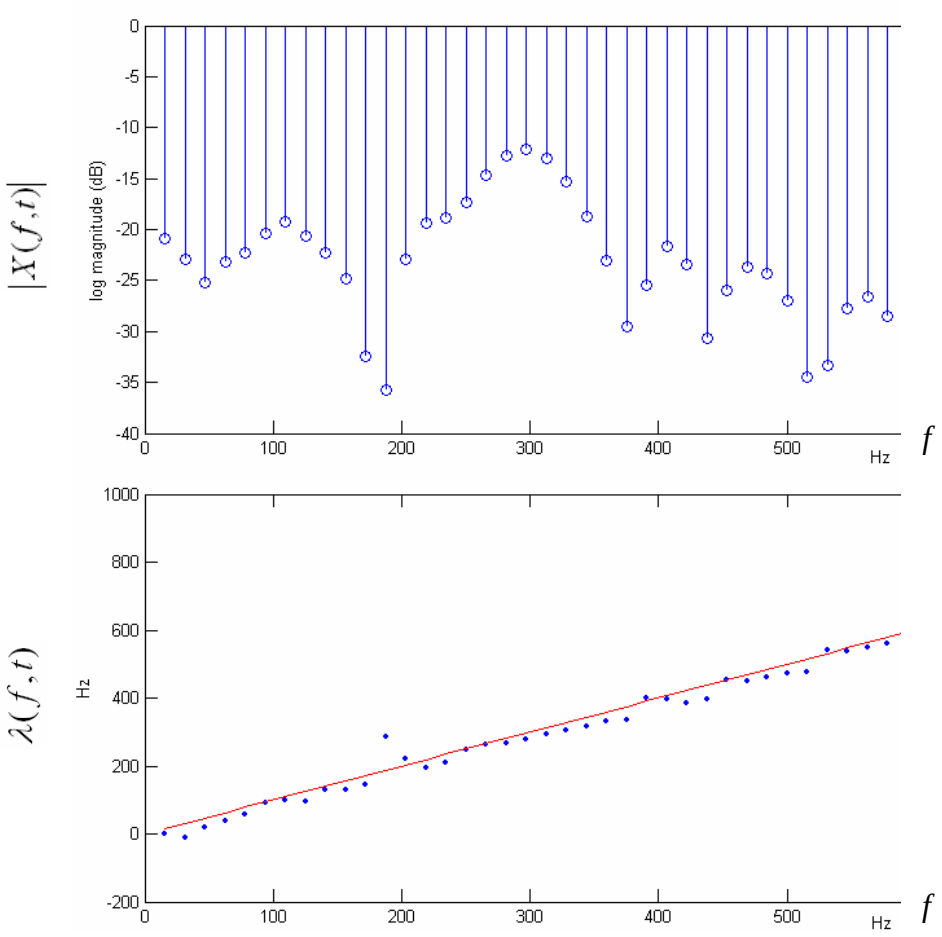


圖 2.3 An example of instantaneous frequency of unvoiced signal

從 voiced signal 與 unvoiced signal 兩個例子來看，voiced signal 的 IF(instantaneous frequency)與頻率軸呈現很明顯的階梯狀，而且階梯的平台出現在基頻與基頻的整數倍數上，而 unvoiced signal 的 IF 並沒有這種特性，反而呈現雜亂的分布，基於這兩者的差異，我們可以利用這個特徵，不僅僅能分辨出 voiced 或是 unvoiced，而且可以進而求到此音框的基頻頻率值。

2.2 瞬時頻譜

我們已經利用前一節的方法求得每個音框的瞬時頻率，瞬時頻譜(instantaneous frequency amplitude spectrum)即是將 $x(t)$ 經 STFT 後所得到的頻譜中的頻率軸(f axis)，利用瞬時頻率 $\lambda(f,t)$ 與 f 之間的轉換，將頻譜的頻率軸轉換成 IF 軸($\lambda(f,t)$ axis)，以一個 voiced signal 的實際例子來看，下圖 2.4 即是利用瞬時頻率轉換過後的瞬時頻譜圖，在轉換過後我們可以發現一個特性，所有能量都往基頻與基頻的整數倍數集中，如紅色線所圈選的地方，反之， unvoiced signal 則沒有這種特性，如圖 2.5 所示，能量在 IF 軸中的分布並沒有集中在基頻與基頻的整數倍數上，且平均能量均不如 voiced signal 高。

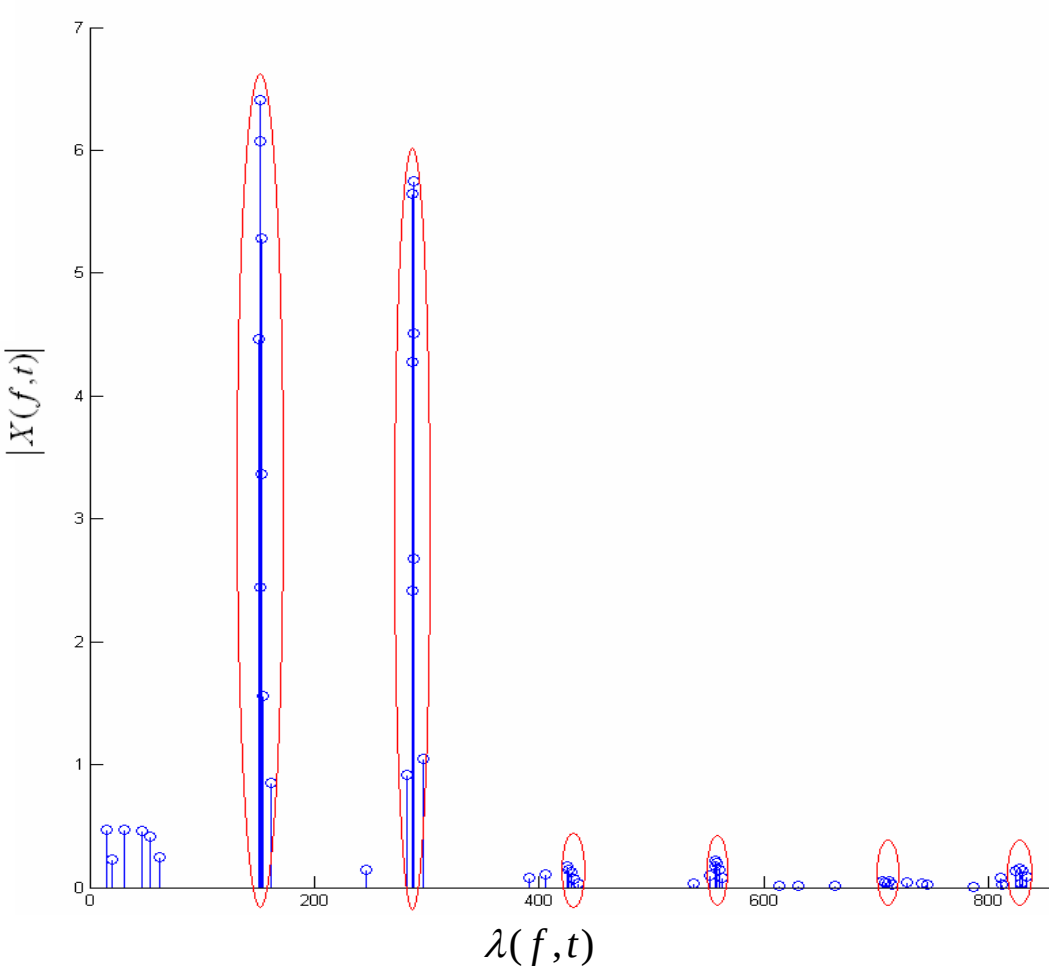


圖 2.4 An example of IFAS with voiced signal

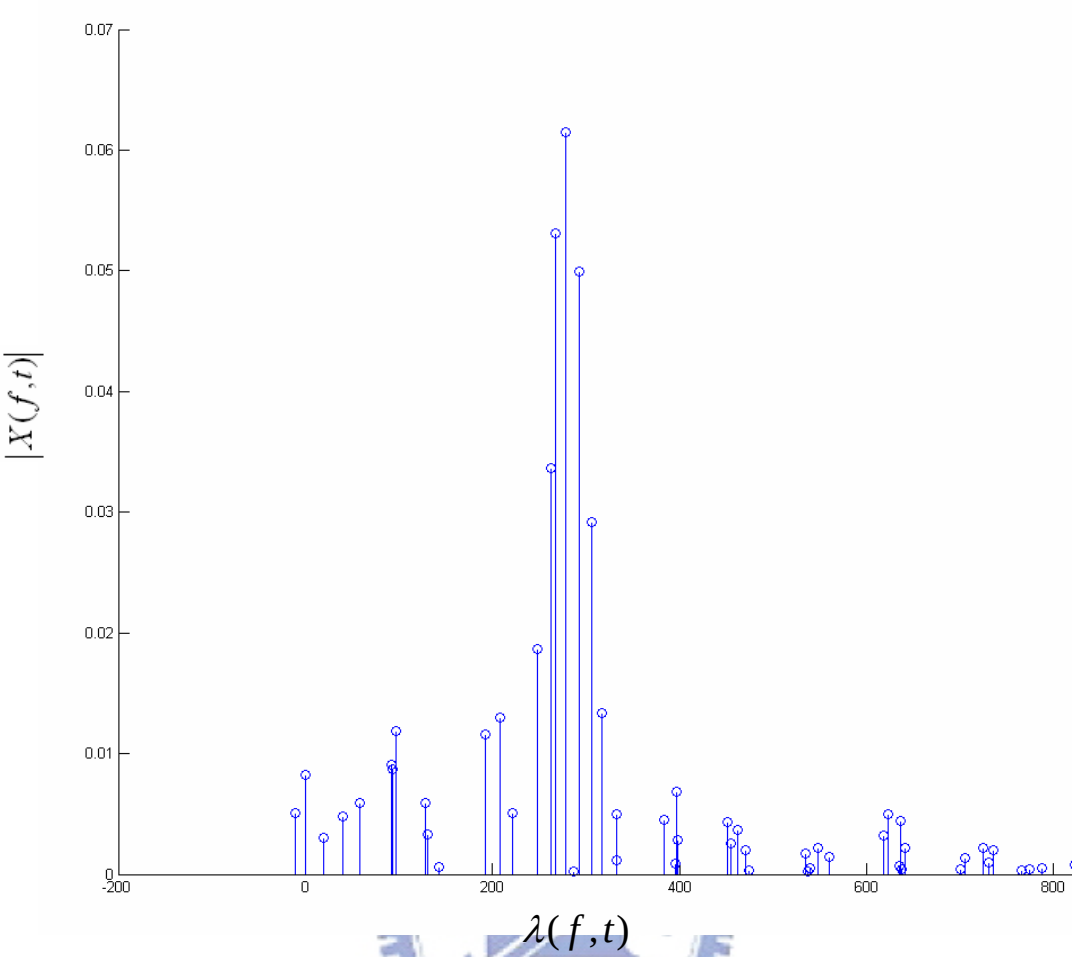


圖 2.5 An example of IFAS with unvoiced signal

2.3 利用瞬時頻譜產生基頻值候選者

由上節所述得知，如果聲音含有週期的特性，則瞬時頻率與頻率軸有階梯狀的轉換曲線，所以若將此音框的 amplitude spectrum 利用此轉換曲線轉換頻率軸後，可發現能量會往基頻與基頻的整數倍數上集中，可以由圖 2.4 例子明顯看出，於是我們可以利用這個特性，找出此音框的基頻候選者。

2.3.1 轉換函式

雖然我們可以在瞬時頻譜中很清楚找出此音框的基頻值，但我們仍舊需要一個轉換的動作，讓我們可以很簡單的判斷出基頻值與其他可能是基頻的值，於是我們定義一曲線函式如下：

$$\Lambda(\lambda,F)=\begin{cases}0, & ,2\pi\lambda/F<\pi\\ \frac{1}{2}(\cos(2\pi\lambda/F)+1), & 2\pi\lambda/F\geq\pi\end{cases} \tag{2.5}$$

單獨看式子中的$\frac{1}{2}(\cos(2\pi\lambda/F)+1)$可發現，此曲線函式有區域最大值出現在$\lambda/F=1,2,3...$的地方，相等於$F=\frac{\lambda}{n}$,$n=$positive integer 的地方，範例如下圖2.6，圖中設定λ值為180，畫出$F=60\sim400$所得到的$\Lambda(\lambda,F)$，可以發現，$\Lambda(\lambda,F)$在$F=\frac{180}{1}$、$\frac{180}{2}$、$\frac{180}{3}$...有最大值。

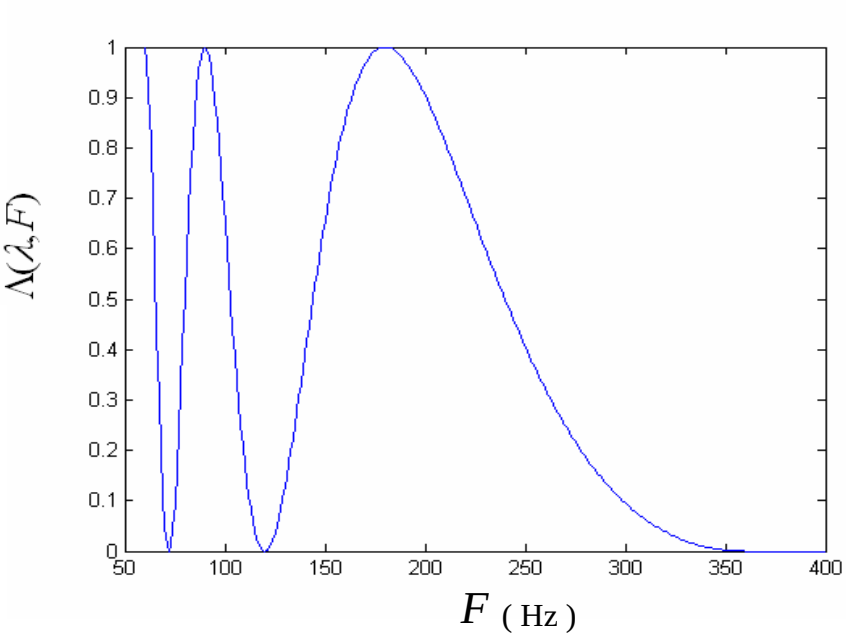


圖 2.6 曲線函式範例(一)，假設$\lambda=180$, $F=60\sim400$

接著用一個簡單的例子來解釋如何對瞬時頻譜進行我們所想要的轉換，假設現在有個瞬時頻譜圖只有在 λ_1=9， λ_2=10， λ_3=11 Hz 的地方有能量，先假定 magnitude 大小均為 1，如圖 2.7 藍線所示，每根能量均可以得到各自擁有的 $\Lambda(\lambda,F)$ 曲線($\Lambda(9,F)$ -綠線， $\Lambda(10,F)$ -紫線， $\Lambda(11,F)$ -紅線)，但各自的區域最大值出現的地方不相同，最後可以利用下面 2.6 式計算此三條曲線疊加後的結果，

$$\begin{aligned}\Lambda(\lambda_1,F)+\Lambda(\lambda_2,F)+\Lambda(\lambda_3,F)=&\frac{1}{2}(\cos(2\pi\times 9/F)+1)\ ,\ F\le 2\times 9\\&+\frac{1}{2}(\cos(2\pi\times 10/F)+1)\ ,\ F\le 2\times 10\\&+\frac{1}{2}(\cos(2\pi\times 11/F)+1)\ ,\ F\le 2\times 11\end{aligned}\qquad (2.6)$$

，疊加後的結果如圖 2.8 所示，

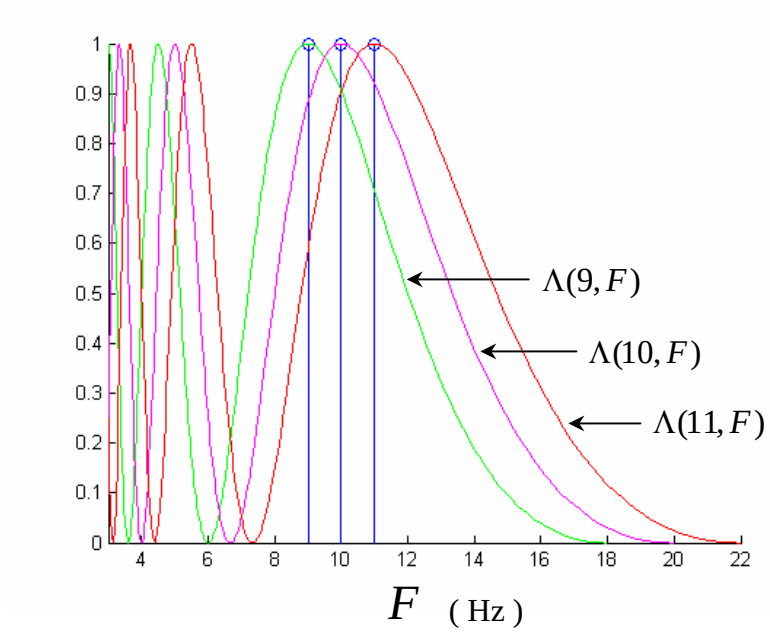


圖 2.7 曲線函式範例(二)， $F=3\sim 22$

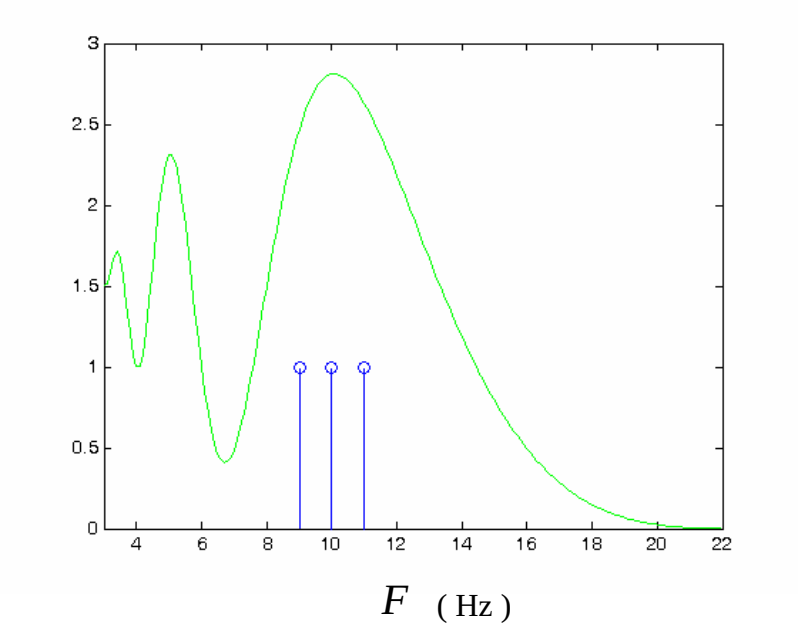


圖 2.8 $\Lambda(\lambda_1,F)+\Lambda(\lambda_2,F)+\Lambda(\lambda_3,F)$ 曲線， $F=3\sim 22$

如果以圖 2.8 產生後的結果來判斷，我們發現曲線的區域最大值出現在 $F=10,\frac{10}{2},\frac{10}{3}$ 的地方，如果說音框的瞬時頻譜圖剛好跟圖 2.8 一樣的話，那麼他的基頻值應該是 $F=10,\frac{10}{2},\frac{10}{3}$ 的地方，與曲線的區域最大值一樣，所以我們可以將上述的轉換方法，將所得到的瞬時頻譜轉換成最後的基頻判斷曲線，使我們更方便求取基頻候選者。

2.3.2 基頻判斷曲線

從上節的說明例子中，我們可以直接套用到實際求取的瞬時頻譜圖上，所以若我們將 $\lambda(f,t)$ 中 f 的值從 0 到 1kHz(經驗值)帶入式子中，便可得到 $\lambda_0=\lambda(0,t)$ 到 $\lambda_{1k}=\lambda(1k,t)$ ，進而可帶入式子 $\sum_{i=0}^{1k}\Lambda(\lambda_i,F)$ ，知道當瞬時頻率 $\lambda(f,t)$ 的值接近基頻時通常會出現很多值非常接近的點，形成一個階梯的形狀，如圖 2.9 圈選起來的地方所示，

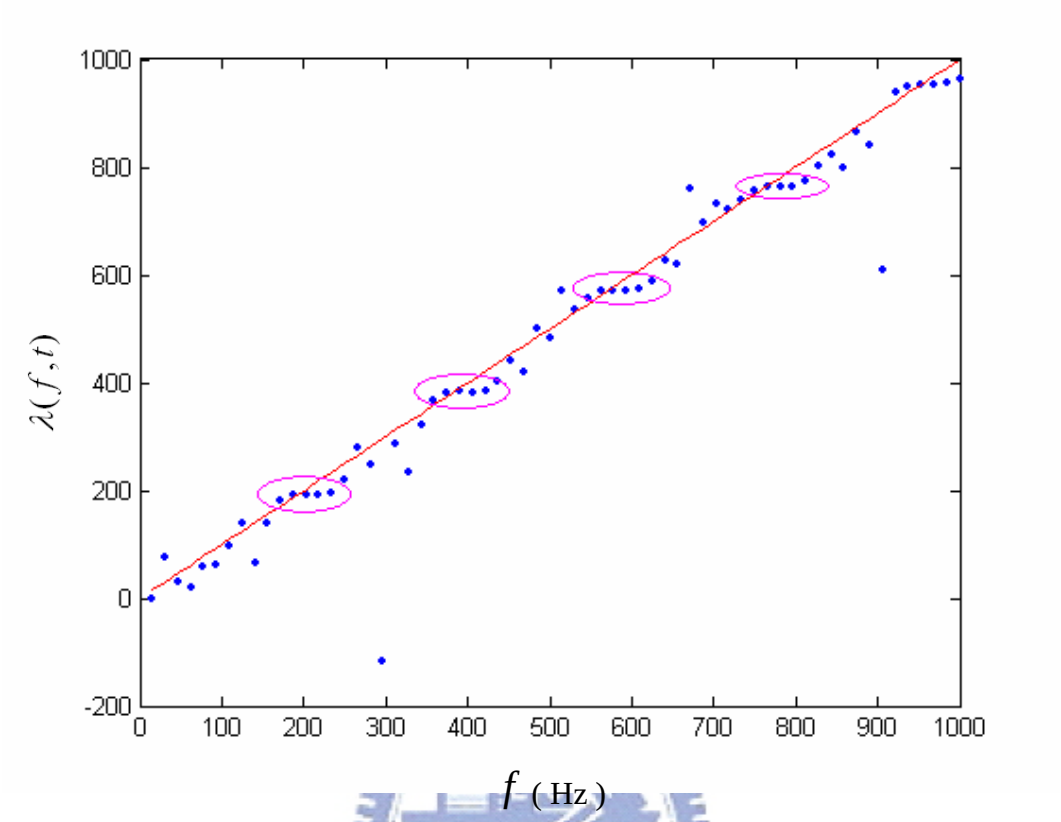


圖 2.9 voiced signal 瞬時頻率，基頻為 190Hz

所以利用此瞬時頻率所得到的瞬時頻譜會呈現出，能量集中在基頻與基頻的整數倍數上如圖 2.4 一般，而每一根能量又各自擁有一條曲線，最後疊加後的結果利用數學式可以寫出如下，

$$\eta'(F)=\sum_{i=0}^{1k}\Lambda(\lambda_i,F), \tag{2.7}$$

其中，$\lambda_0=\lambda(0,t),\ \lambda_{1k}=\lambda(1k,t)$

η' 為最後的基頻判斷曲線，利用上式積分，以同樣 voiced signal，基頻為 190Hz 的例子畫出 $\eta'(F)$，$F=60\sim400Hz$，最後如圖 2.10 所示

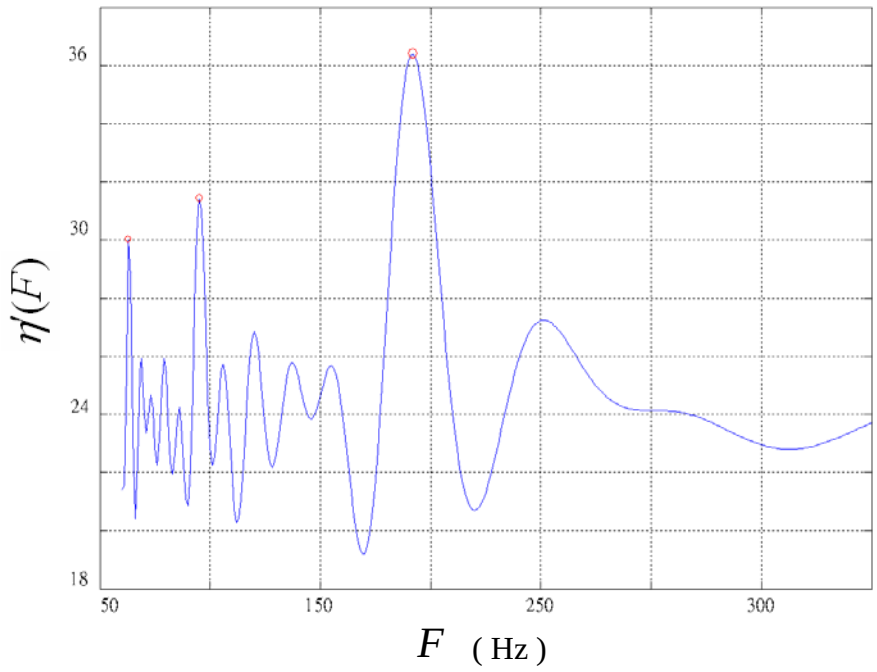


圖 2.10 voiced signal 基頻判斷曲線，基頻為 190Hz

可以從圖 2.10 很明顯看出，曲線區域最大值出現在 $F=\frac{F0}{n}, n=1,2,...(F0=190Hz)$，而且最高值出現在基頻 190Hz 的地方，接著我們看一個 unvoiced signal 的例子，如圖 2.11 所示，從曲線來看，我們無法明確在區域最大值上看出此音框的基頻值落在那個頻率上。

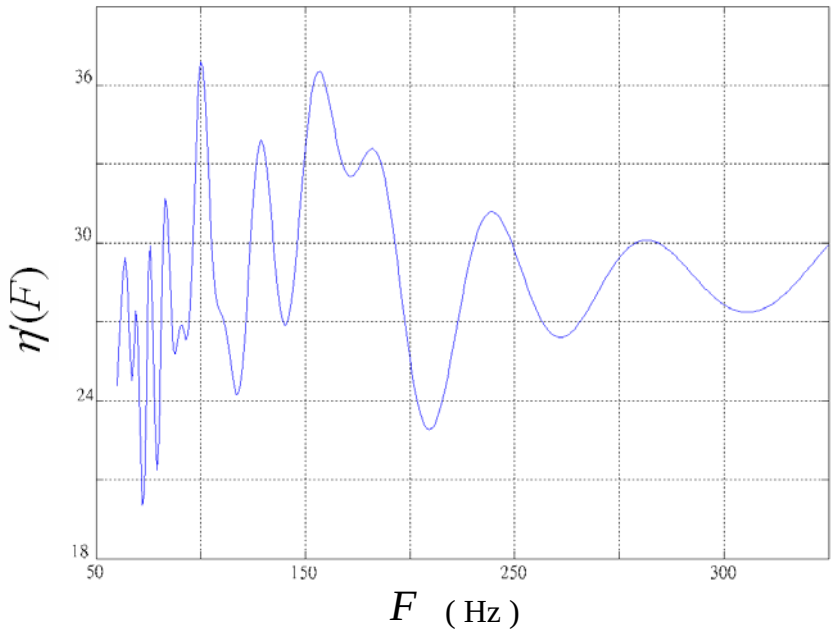


圖 2.11 unvoiced signal 基頻判斷曲線，$\eta'(F)$

雖然說我們可以利用尋找曲線$\eta'(F)$的區域最大值，來推得音框的基頻值，但是可以發現 voiced 與 unvoiced signal 兩曲線的區域最大值並沒有很大的差異，所以為了加強區別 voiced signal 與 unvoiced signal，我們利用兩者頻譜的差異性，觀察兩者頻譜如下圖 2.12

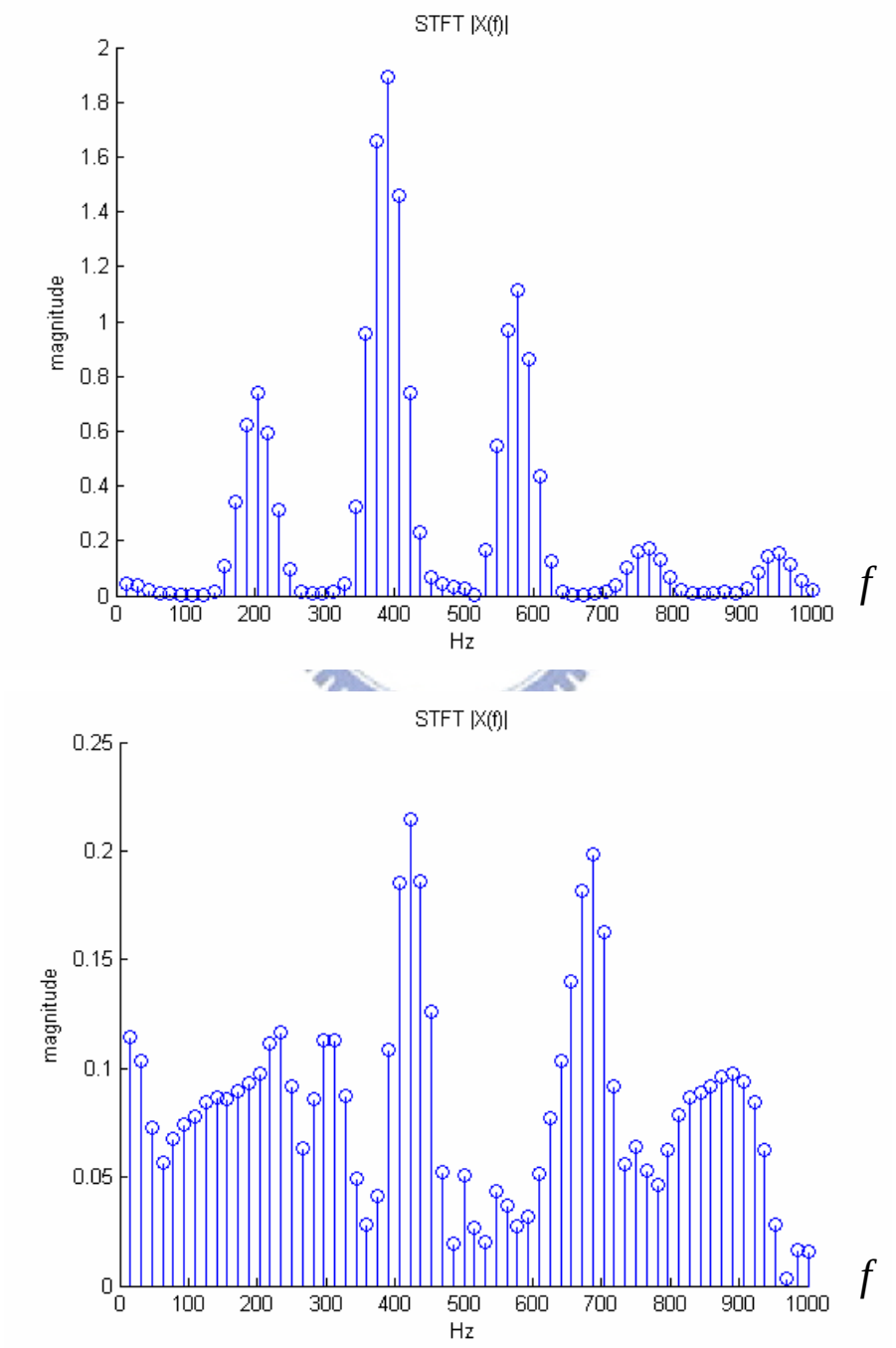


圖 2.12 voiced signal 頻譜圖(上圖)與 unvoiced signal 頻譜圖(下圖)

依觀察可發現兩者之間頻譜的 magnitude 值差異甚大，voiced signal 通常能量集中在 harmonic 上，而 unvoiced signal 能量則是均勻分布，而且在 voiced signal 的能量可以突顯出基頻與基頻的整數倍所在位置，所以我們可以利用 STFT$|X(f)|$ 的 magnitude 加強曲線$\eta'(F)$，使得曲線$\eta'(F)$能更加明確分別 voiced 與 unvoiced signal，並提升判斷基頻值的準確性，於是改寫曲線$\eta'(F)$如下：

$$\eta(F)=\sum_{i=0}^{1k}|X(f_i)|\Lambda(\lambda_i,F) \tag{2.8}$$

其中，$\lambda_0=\lambda(f_0,t)$，$f_0=0Hz$，$f_{1k}=1kHz$。

從$\Lambda(\lambda,F)$式子中我們可以知道，每一個λ值都可以找到對應的曲線，而且此曲線的最大值分別在$\frac{\lambda}{n}$上，而我們知道λ呈現階梯狀，階梯平台上的λ值都非常接近基頻的整數倍數，所以若將類似圖 2.6 一樣的曲線$\Lambda(\lambda,F)$不斷的疊加上去，而且我們可以直覺的知道區域最大值會出現在階梯平台對應到的λ值，而這些λ值通常都是在基頻的整數倍數上，乘上$|X(f)|$加強了當λ的值是落在平台上的曲線，將這些區域最大值拉高，因此我們可以搜尋判斷式$\eta(F)$，$F=60\sim400Hz$，在基頻的整數倍數上有區域最大值，而且頻率越低，$\eta(F)$的值會越高，因此會引發出容易找到 half-pitch，所以我們在$\eta(F)$式子前面加上一個權重$\alpha^{-\beta/F}$，其中α與β均為正整數(經驗值為α=10，β=8)，如下面式子 2.9 所示：

$$\eta(F)=\alpha^{-\beta/F}\sum_{i=0}^{1k}|X(f_i)|\Lambda(\lambda_i,F)\ ， \tag{2.9}$$

加上此權重後，會將頻率較低的部分的壓低，相等於讓高頻部分擁有更高的 priority，

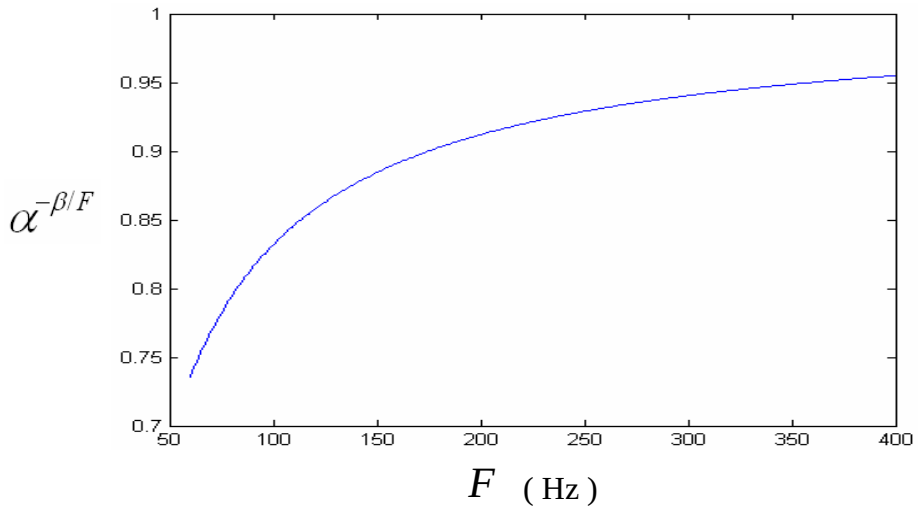
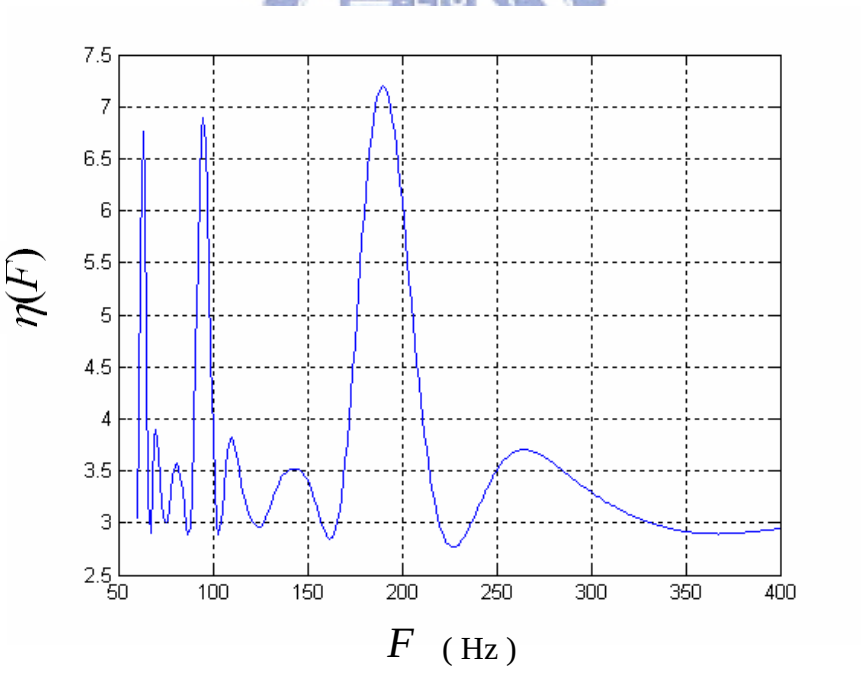


圖 2.13 $\alpha^{-\beta/F}$ 函式，$\alpha=10,\ \beta=8$

後我們可以在每個音框求得到一條$\eta(F)$曲線，尋找這條曲線的最大值，便可以得到此音框的 pitch value (即為最大值所對應到的F)。



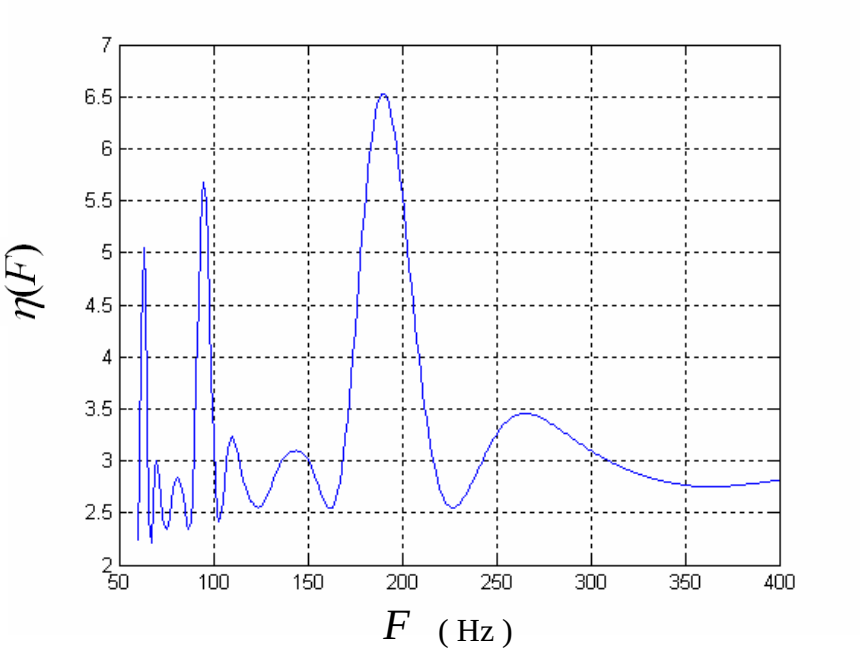


圖 2.14 基頻判斷曲線$\eta(F)$(上圖未加權重，下圖則有，基頻為 190Hz)

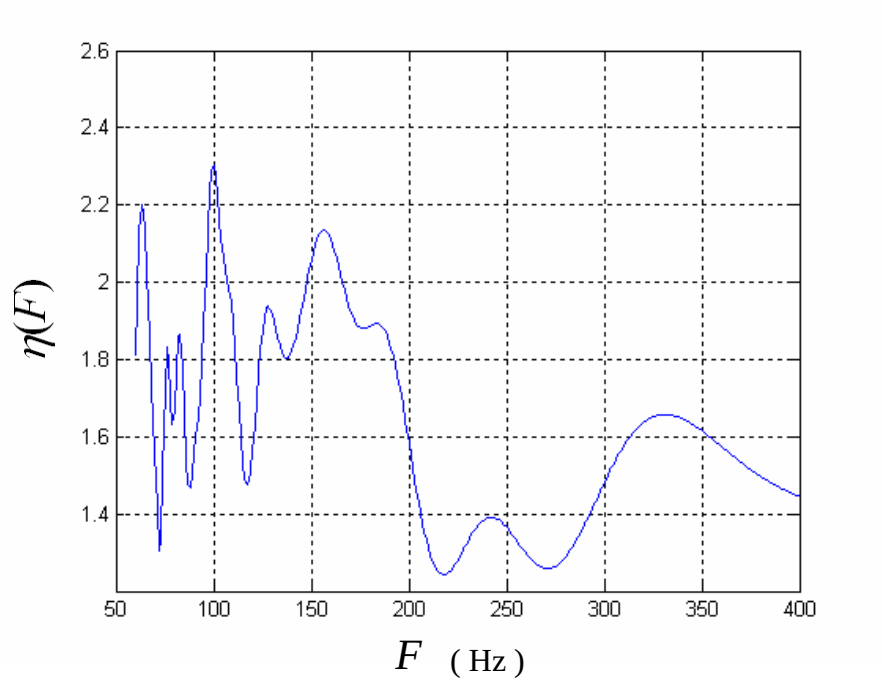


圖 2.15 Unvoiced signal 之基頻判斷曲線$\eta(F)$

從圖 2.14 與圖 2.15 可以發現 voiced signal 與 unvoiced signal 所得到的曲線中的區域最大值有了明顯的差異，於是我們將每一個 frame 所求得的最大$\eta(F)$對於 voiced 音框與 unvoiced 音框的作分佈情形觀察，結果如下圖 2.16：

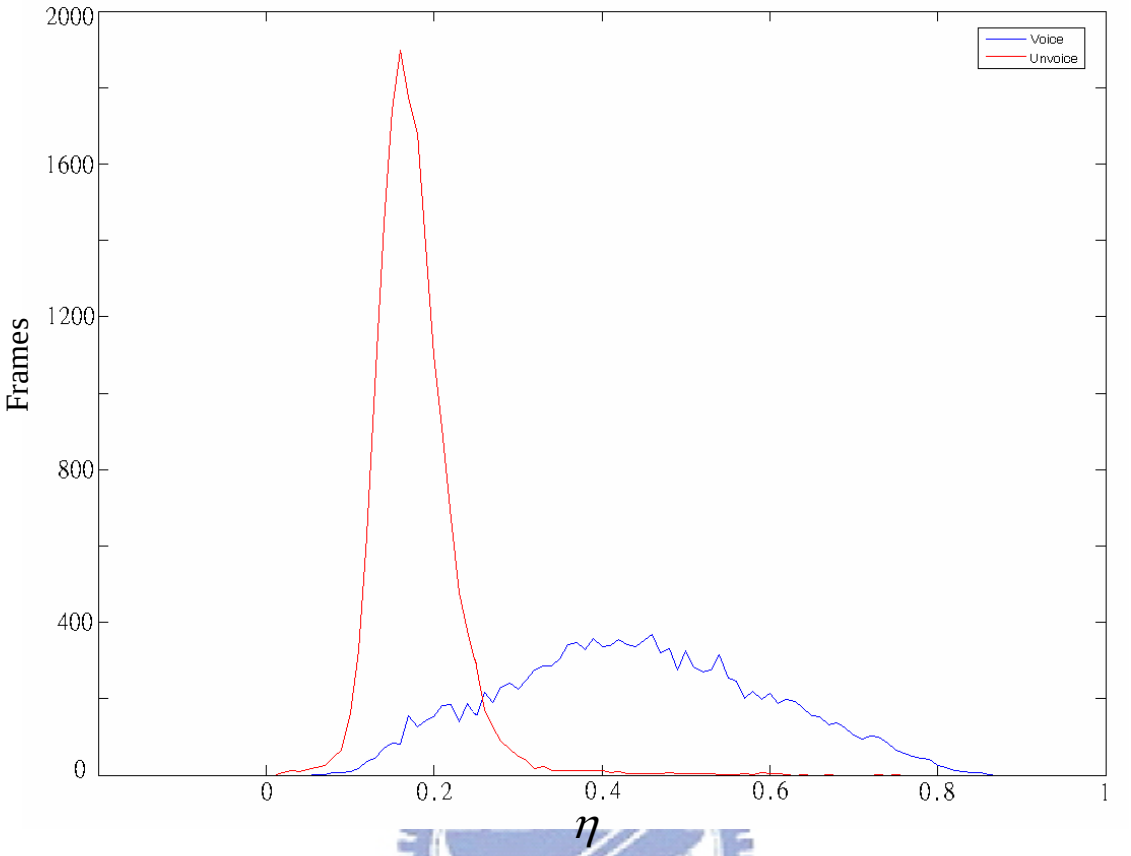


圖 2.16 η 分佈圖

從圖 2.16 發現 voiced 音框(藍線)與 unvoiced 音框(紅線)之間在臨界值附近並沒有很明顯的分界，造成 voiced 與 unvoiced 之間的判定錯誤甚高，其原因是因為判斷曲線利用了頻譜的能量大小來加強曲線上的每一點，所以如果今天有一個訊號，其能量大但是並沒有週期特性，所產生出來的曲線區域最大值與 voiced signal 所得到的差異不大，反之，有週期特性但能量小的音框通常得到的值很容易比 unvoiced 小，為了更有效分辨 voiced 音框與 unvoiced 音框，我們可以將頻譜作能量正規化，於是，雖然能量大但是因為沒有週期，所以頻譜上的能量分佈非常平均，經過正規化後，unvoiced 音框的判斷曲線 $\eta(F)$ 會被整體壓低，而 voiced frame 的判斷曲線會因為正規化而整體提升，最後我們再作一次經過能量正規化的 $\eta(F)$ 值分佈，結果如下圖 2.17：

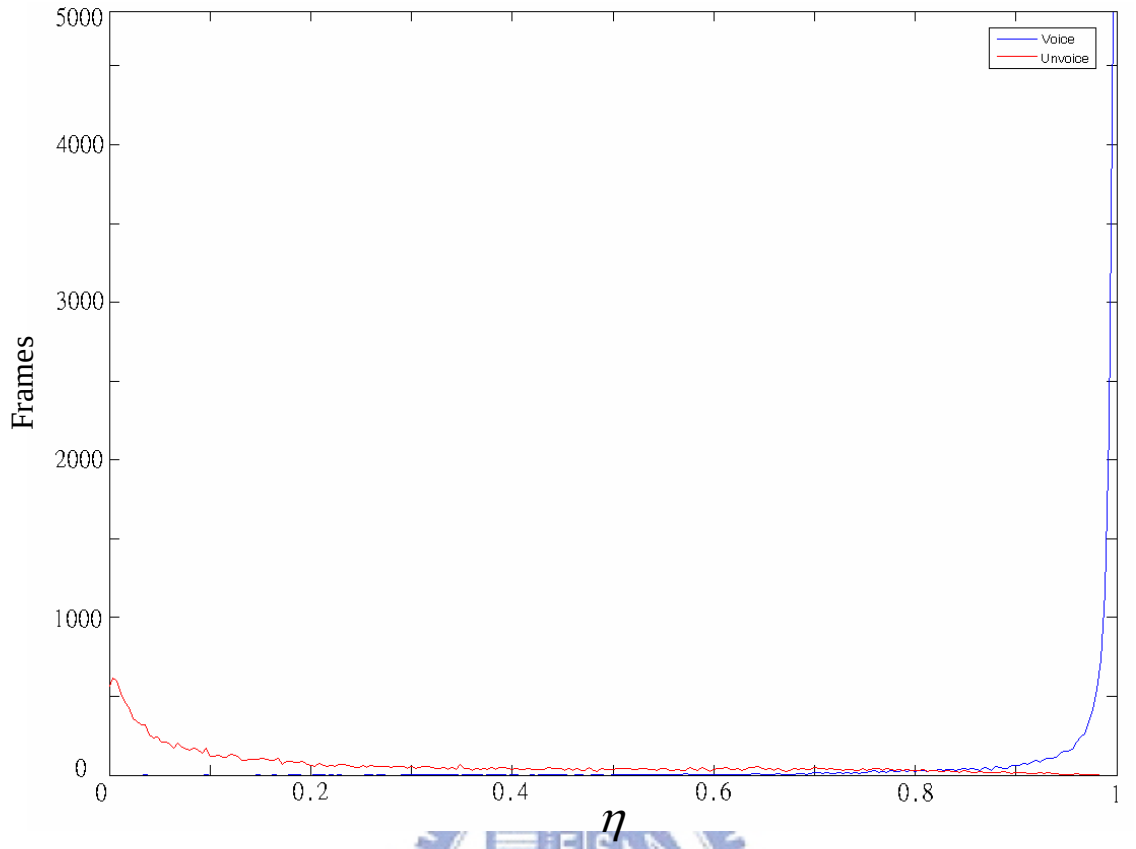


圖 2.17 經正規化後η分佈圖

很明顯的，經過能量正規化後，voiced 音框(藍線)與 unvoiced 音框(紅線)可以從$\eta(F)$中更容易分辨出來，使得 voiced 與 unvoiced 之間的錯誤大大的降低，這個結果對於往後作 V/U 判斷是個很可靠的特徵參數。

2.3.3 從基頻判斷曲線中產生基頻值候選者

瞬時頻譜經過轉換函式的轉換後，接著利用正規化後的頻譜加強了基頻與基頻的整數倍數部分，最後我們可以得到一條大小值落於 0~1 之間的基頻判斷曲線$\eta(F)$，接著可以利用尋找判斷曲線的區域最大值可以產生出基頻值候選者，依照$\eta(F_i)$值的大小分別可以找到基頻值候選者F_i，$i=1\sim5$，如下圖 2.18 所示，由於$\eta(F)$的大小來自於音框能量與 harmonic 能量的比例，所以其值越大，表示

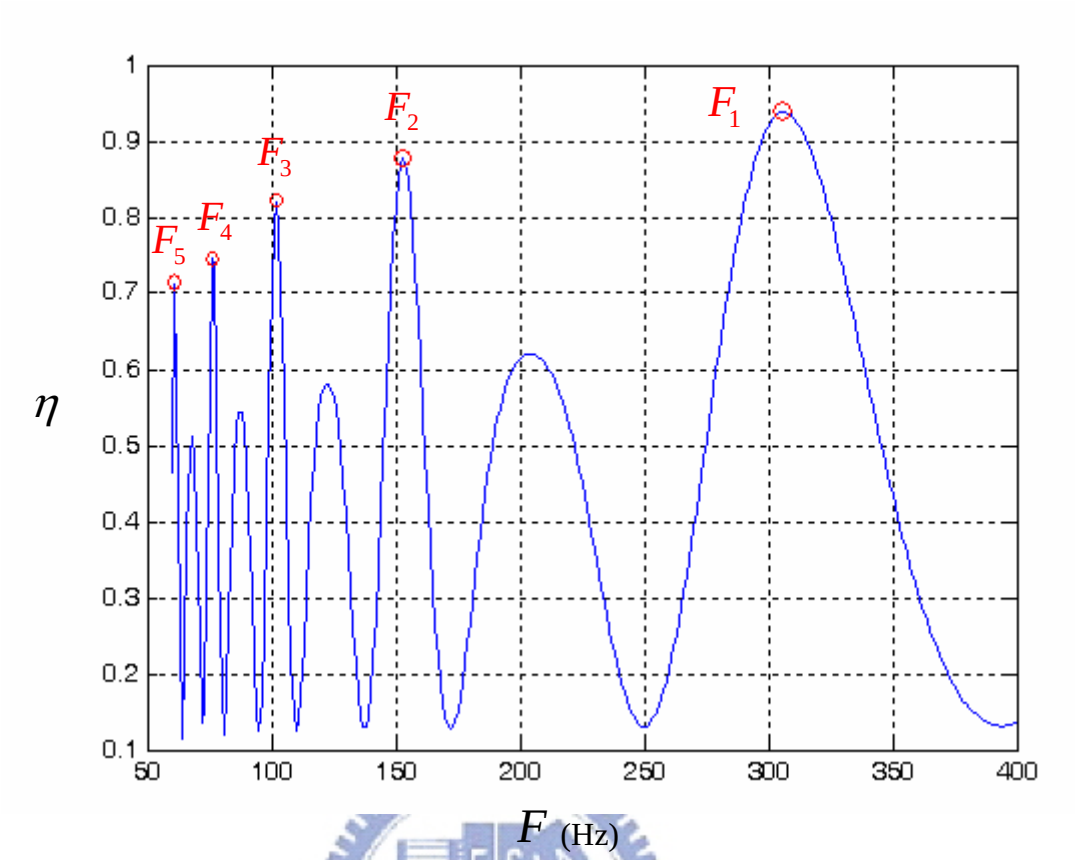


圖 2.18 基頻候選者 F_i 產生方式

此音框為 voiced 的可靠性越高。

接著我們為了瞭解對於所選出來的基頻候選者的正確性與否，於是我們對手標後的基頻參考值進行比對，同時，我們拿目前常見的 Auto-correlation 傳統方式所求得的基頻候選者同時比對，比對的範圍只有當參考基頻為 voiced 的地方，其餘地方則不進行比對，而比對方式為，若此音框其中一個基頻候選者與手標參考基頻值差異小於參考基頻的±5% 內，則視此 frame 視為“正確”，並且紀錄最接近參考基頻值是落於哪一個基頻候選者，比對結果如表 2.1 所示：

表 2.1 各方法基頻候選者與參考基頻值比較(total voiced frame=13355)

瞬時頻率方法		自相關函式方法	
F_1	12,666	F_1	10,941
F_2	179	F_2	1,204
F_3	43	F_3	560
F_4	13	F_4	151
F_5	16	F_5	52
Total	12,917	Total	12,908

(F_1=基頻候選者第一順位，依此類推…)

我們可以從上表發現，自相關函式法雖然在總數上與瞬時頻率方法差不多，但其正確基頻值並非集中在候選者第一名，這表示加上追蹤與平滑後，很容易出現 double-pitch 與 half-pitch 的情形，由上面實驗結果，我們可以確信瞬時頻率的方法所求得基頻候選者有非常好的可靠度與準確度。

第三章　中文聲調辨認與基頻軌跡建立

由本文前一章節所敘述的方式，以一個音框為時間單位產生五個基頻候選者，接著在這候選者中，利用前後音框候選者之間的關係與前後音框基頻值變化...等等，選擇一條最佳路徑建立基頻軌跡估測，雖然利用瞬時頻譜為基礎的方法抽取出的基頻軌跡具有相當可靠的程度，但是仍舊會發生半頻基頻(Half pitch)或是倍頻基頻(Double pitch)的錯誤，於是本節希望能利用以統計為基礎的中文韻律模型與聲調模型[8]，因為具有 tone shape 與 tone、prosody transition 的統計特性，所以能有效輔助基頻軌跡的建立，使得上述兩種錯誤的發生率減少，並同時完成中文聲調辨認。

對於一般中文語音而言，每一個中文字都是由一個音節所構成，而每一個音節結構部分，又可以分成 411 基本音節與聲調兩大部分，本文接下來就是要對聲調部分的辨認技術結合基頻軌跡建立加以研究與探討。

在基頻軌跡估測方面，利用每個音框所抽取出來的候選者與其對應之$\eta(F)$值，接著以每一個音節為單位，利用 Viterbi tracking 方式產生出前五名最佳路徑，並將這五名路徑視為基頻軌跡候選者。在聲調辨認方面，利用統計的聲調與韻律模型，對於每一段音節的基頻軌跡候選者比對，配合聲調轉移機率、韻律轉移機率，搜尋最大可能性的基頻軌跡，同時決定此基頻軌跡候選者所屬的聲調與韻律組合。

3.1 國語聲調的特性

中文語音中，每一個音節都具有一種聲調，在眾多發音裡面，總共包括了五

種不同的聲調，一般我們分為一聲、兩聲、三聲、四聲與五聲。這裡指的聲調，就是指我們在發音的時候，隨者時間的變化下，頻率會有不同的高低起伏變化而產生出不同聲調。如果從基頻軌跡來觀察，我們可以發現在一般的單字音，我們所發出的聲調，其基週軌跡之標準形式如圖 3.1 所示，各自具有其獨特的基頻軌跡分佈。在這圖中並沒有標示出第五聲(一般稱為輕聲)的基頻軌跡，這是因為通常第五聲的基頻軌跡並不像其它四種聲調一樣具有規則性的基頻軌跡；在其出現的同時，常常是一不規則軌跡，且能量與音節長度也較其它聲調低、短。

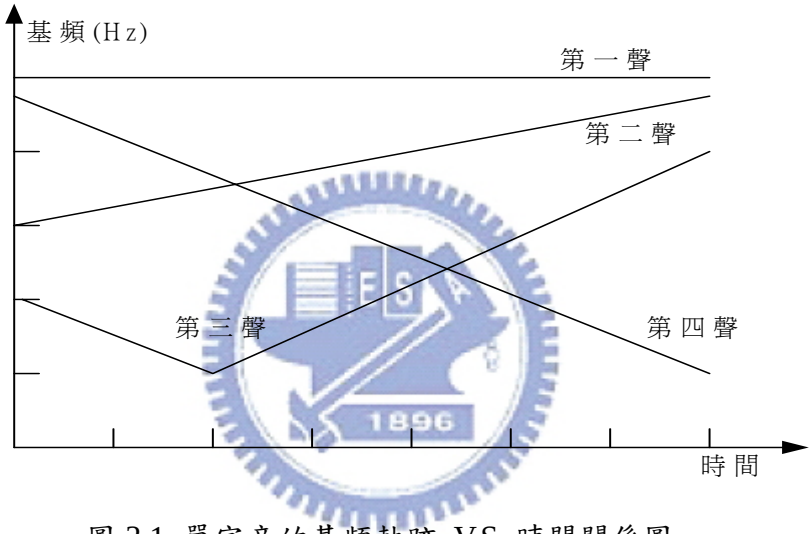


圖 3.1 單字音的基頻軌跡 V.S 時間關係圖

從頻率分佈來看不同的聲調，也可以發現一聲的整體平均值較其它聲調來得高，接下來從高到低依序約為二聲、四聲、三聲，不過這也只是就同一個語者所發出的頻率而準。因為一般而言，女生所發出的頻率都比男生來的高，所以女生所發出第三聲的頻率，也常比男生發出第一聲的頻率還要高，故只是一個相對性的比較。

對於三聲在基頻軌跡上的表現，其實並非每一個三聲都會出現完整的先降後升(falling-raising)的軌跡，於是我們對三聲的基頻軌跡再加以分類，第一類(圖 3.2)為只降不升(falling-dipping)與四聲的基頻軌跡相當類似，第二類(圖 3.3)為只升不

降(raising)，與二聲類似，通常發生於變調(tone-sandhi)的狀況，第三類(圖 3.4)則是完整的先降後升。

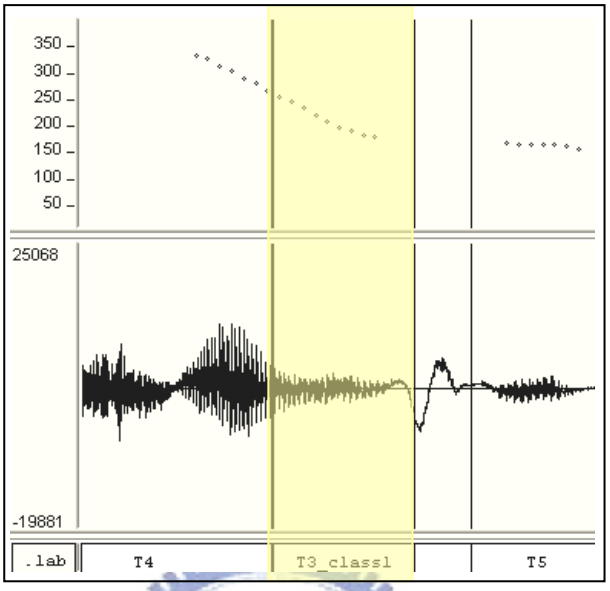


圖 3.2 中文聲調三聲類別一

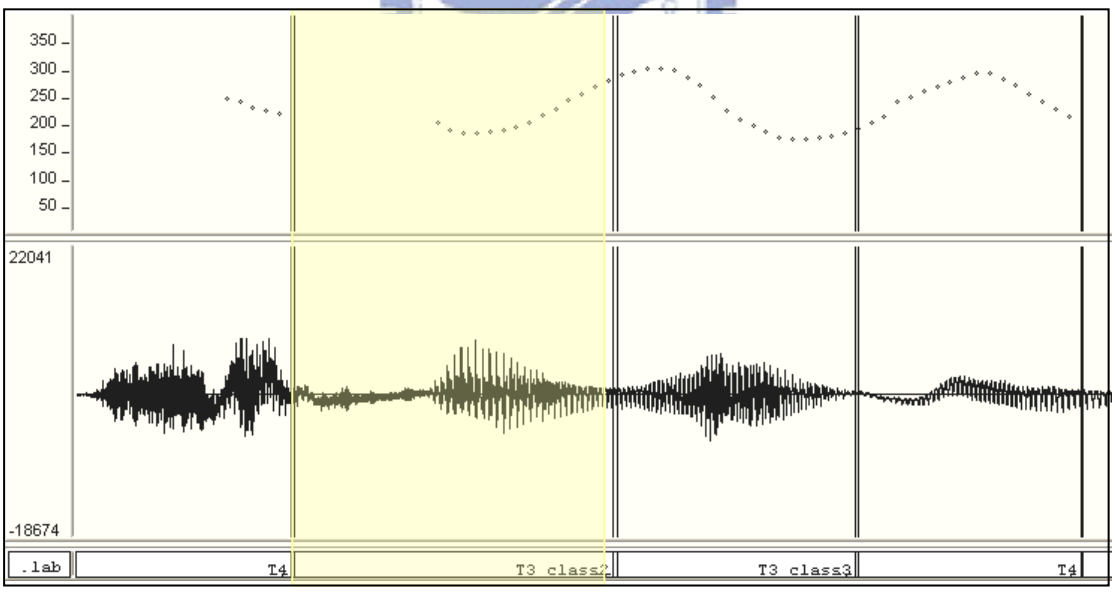


圖 3.3 中文聲調三聲類別二

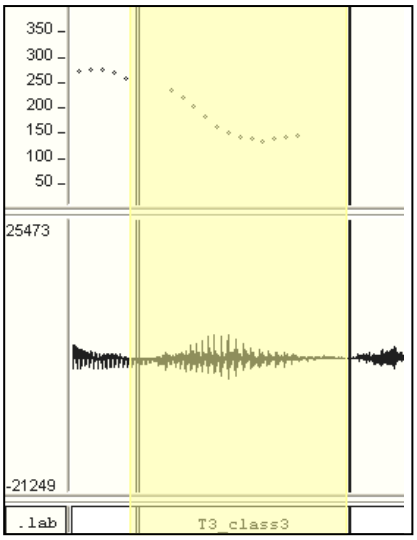


圖 3.4 中文聲調三聲類別三

儘管我們能夠知道一般單字音當中，聲調的基頻軌跡分佈多如我們所預期的樣子，而現實中，連續語音則會受到前後音之聲調、語音之韻律模式，甚至語意等其它因素的影響，都是造成聲調的辨識困難度大為提高的原因之一。

3.1 聲調模型與韻律模型建立

我們假設影響基頻軌跡形成的主要因素有三種，一種為音節的聲調，一種為此音節的韻律狀態，一個為語者，而且這三個影響因素是具有加成性的，並且可以表示成如下式子：

$$\mathbf{x}_{k,n} = \mathbf{y}_{k,n} + \mathbf{PT}_{t_{k,n}} + \mathbf{PP}_{p_{k,n}} + \boldsymbol{\mu}\mathbf{p} \tag{3.1}$$

其中 $\mathbf{x}_{k,n}$ 為利用正交展開式將所觀察到音檔 k 中第 n 個音節的基頻軌跡展開所得到的四維係數，$\mathbf{y}_{k,n}$ 則是 normalized (i.e., residual) pitch contours，$\mathbf{PT}_{t_{k,n}}$ 是影響因

素中的聲調(tone) $t_{k,n} \in \{1,2\cdots 8, L1\cdots, L8, 1R\cdots, 8R, L1R\cdots, L8R\}$，其中 $t_{k,n}=3$ 表示中文語音聲調中的“三聲”，基頻軌跡呈現 falling-dipping，$t_{k,n}=6$ 為呈現 tone-sandhi 表現的“三聲”，$t_{k,n}=7$ 亦是“三聲”，其基頻軌跡呈現 falling-rising，$t_{k,n}=5$ 為基頻值較低的 neutral tone，$t_{k,n}=8$ 則為基頻值較高的 neutral tone，在 tone n 前面加上一個符號“L”表示說 syllable n 與前一個 syllable 的基頻軌跡是連接的(兩段基頻軌跡間隔小於 3 個音框)，同理，若後面加上符號“R”則表示 syllable n 與後面 syllable 的基頻軌跡是連接的，如果 tone n 前後並無任何符號，表示 tone n 的基頻軌跡跟任何 syllable 並無連接(兩段基頻軌跡間隔大於 3 個音框)，$\mathbf{PP}_{p_{k,n}}$ 則是韻律狀態(prosodic state) $p_{k,n} \in \{1,2,....P=16\}$ 的影響因素，$\boldsymbol{\mu}\mathbf{p}$ 為不同語者的影響因素。

由於我們將“三聲”細分成三類，但實際上我們並不知道究竟哪一個“三聲”應該是屬於哪一類，於是我們利用一些準則去分類，並且在每一次更新 tone model 的時候，再一次重新分類，使得“三聲”更加有效的分成三類。首先，將語料庫中所有 tone-sandhi 的組合收集起來，例如 3-3 與 3-3-3 等組合，並標成 6-3 與 6-6-3，最後將其餘剩餘的“三聲”利用 VQ 分成兩類，分別標成 tone3 與 tone7，如此便可完成“三聲”的初始分類，同樣的方式，利用 VQ 將“五聲”分成兩類，pitch mean 較低的一群標成 tone5，較高的一群標成 tone8，既可得到“五聲”的初始分類。

如上，每一段基頻軌跡都含有聲調、韻律因素與語者因素，扣除三個主要因素之後，剩餘項 $\mathbf{y}_{k,n}$ 可以 model 成一個四維的高斯分佈 $N(\mathbf{y}_{k,n};\mathbf{0},\mathbf{Rp})$ (or equivalently $\mathbf{x}_{k,n}$ is modeled by $N(\mathbf{x}_{k,n};\mathbf{PT}_{t_{k,n}}+\mathbf{PP}_{p_{k,n}}+\boldsymbol{\mu}\mathbf{p},\mathbf{Rp})$)，$\mathbf{Rp}$ 為基頻軌跡的 covariance matrix。

3.2 聲調模型與韻律模型之訓練

在上一節我們已經將基頻軌跡的聲調與韻律狀態建立出對應的模型，並將中文五個聲調更細分成八個聲調，並且考慮到前後音節基頻軌跡是否有連接，接下來便是要如何初始化且接著訓練這些模型，在這一節中我們採用了 Maximum Likelihood (ML) criterion，利用遞迴的方式重複訓練模型，使得這些模型可以得到最佳化。由於每個因素的更新順序是非常重要的，以本文三個因素而言，語者因素為第一優先，接著為聲調，其次為韻律狀態。

下圖為聲調模型的訓練流程，由於實驗所使用的語料庫為單一語者所錄製的，所以 speaker mean 我們可以經過一次計算就可求得，接著分別計算 tone 與 prosodic pattern 的初始值，每次遞迴時都會重新標 tone 3、6、7 與 tone 5、8，以及再一次更新 tone 與 prosodic pattern，並計算 tone 與 prosodic state 的 transition probability，直到 likelihood function 收斂為止，接著，我將在下一節詳盡介紹訓練過程的每一個步驟。

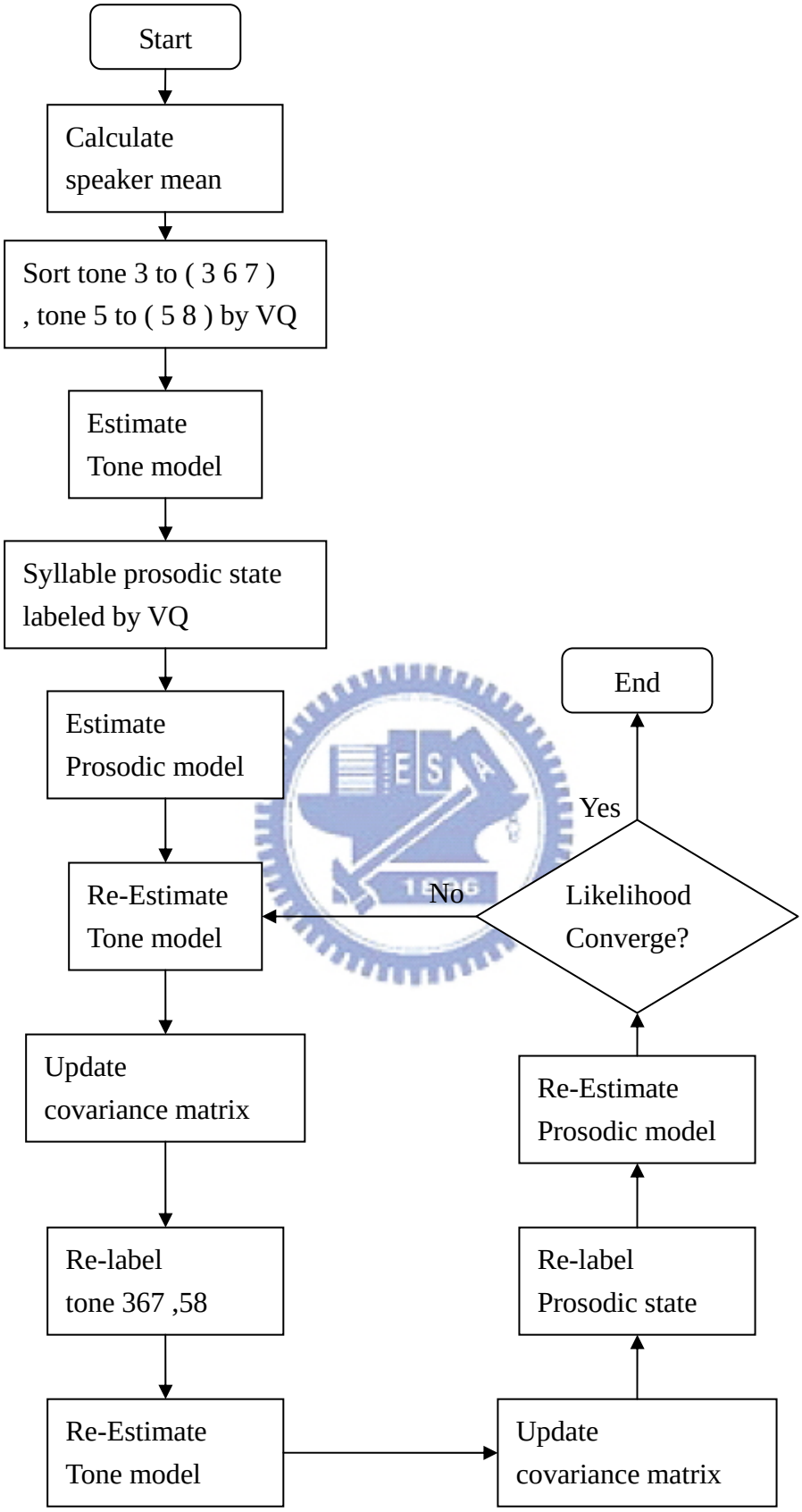


圖 3.5 聲調與韻律模型訓練流程圖

3.2.1 初始化模型

- 步驟(1)，首先解決的是語者因素，利用所有音節的基頻軌跡取平均，可以輕易的得到 speaker mean ($\boldsymbol{\mu p}$)，計算式如下：

$$\boldsymbol{\mu p} = \frac{1}{N_t}\sum_{k=1}^{K}\sum_{n=1}^{N_k}\mathbf{x}_{k,n} \tag{3.2}$$

其中 N_k = 在句子 k 中的所有音節數

K = 語料庫中所有的句子數

N_t = 語料庫中所有音節數

- 步驟(2)，將語料庫中的“三聲”與“五聲”分別分類成 tone 3、6、7 與 tone 5、8。

- 步驟(3)，利用所求的 $\boldsymbol{\mu p}$，便可以對所有音節的基頻軌跡移除語者因素，利用剩下來的餘項 $\mathbf{x}_{k,n} - \boldsymbol{\mu p}$，與“三聲”、“五聲”分類後的結果，建立聲調初始模型，計算式如下：

$$\mathbf{PT}_t = \frac{\sum_{k=1}^{K}\sum_{n=1}^{N_k}(\mathbf{x}_{k,n} - \boldsymbol{\mu p})\times\delta(t_{k,n} = t)}{\sum_{k=1}^{K}\sum_{n=1}^{N_k}\delta(t_{k,n} = t)} \tag{3.3}$$

其中 $t_{k,n} \in \{1, 2\cdots 8, L1\cdots, L8, 1R\cdots, 8R, L1R\cdots, L8R\}$

- 步驟(4) 經過步驟(1)與步驟(3)後，我們有了初始化的聲調模型 $\mathbf{PT}_t$ 與 speaker mean $\boldsymbol{\mu p}$，移除了這兩個主要因素後，那麼剩餘便只剩下韻律因素了，所以利用 VQ 分群的方式，將最後的餘項 $\mathbf{x}_{k,n} - \mathbf{PT}_{t_{k,n}} - \boldsymbol{\mu p}$ 分成

16 群，也就是 16 個韻律狀態，最後對分完群後的所有基頻軌跡標記上所屬的韻律狀態 $\mathbf{PP}_{p_{k,n}}$，並且利用如下數學式計算出 $\mathbf{PP}_{p_{k,n}}$，直覺上，如果初始韻律狀態能有效的分出 16 群，對於之後能大大的減少訓練次數。

$$\mathbf{PP}_p = \frac{\sum_{k=1}^{K}\sum_{n=1}^{N_k}(\mathbf{x}_{k,n} - \mathbf{PT}_{t_{k,n}} - \boldsymbol{\mu}\mathbf{p}) \times \delta(p_{k,n} = p)}{\sum_{k=1}^{K}\sum_{n=1}^{N_k}\delta(p_{k,n} = p)} \tag{3.4}$$

其中， $p \in \{1, 2, ...P = 16\}$

上述步驟(1)~(4)介紹了語者模型、聲調模型與韻律模型的初始化方法，為了使得模型更佳準確，我們使用了 ML 訓練方式，接著下一節我們會介紹，整個訓練過程，與模型更新順序與方式。

3.2.2 模型訓練流程

在整個訓練過程最重要的就是模型的更新順序，其更新的順序依次是語者模型、聲調模型、韻律模型，因為所使用的語料庫為單一語者的語料庫(請參考附錄一)，所以關於語者模型只需要第一次的初始化後，就不需要再去更新了，而更新順序影響最嚴重的是聲調模型與韻律模型，所以必須特別注意這兩者更新順序不可以顛倒。

在每一次更新完聲調模型與韻律模型後，利用 ML criterion，我們可以計算出每一次的 likelihood function L (詳細的定義在步驟(10))，重複更新並且觀察 L，當 L 達到收斂之後，便可以停止更新模型的動作。

- 步驟(5)，得到更新後的韻律模型 $\mathbf{PP}_p$，再次更新聲調模型 $\mathbf{PT}_t$，其計算式重寫如下所示：

$$\mathbf{PT}_t = \frac{\sum_{k=1}^{K}\sum_{n=1}^{N_k}(\mathbf{x}_{k,n} - \mathbf{PP}_{p_{k,n}} - \boldsymbol{\mu}\mathbf{p}) \times \delta(t_{k,n} = t)}{\sum_{k=1}^{K}\sum_{n=1}^{N_k}\delta(t_{k,n} = t)} \tag{3.5}$$

其中，$\mathbf{PP}_{p_{k,n}}$ = 在句子 k 中第 n 個音節的韻律狀態模型

- 步驟(6)，由於我們希望我們利用聲調模型與韻律模型重建的基頻軌跡(reconstruct pitch contour; $\boldsymbol{\mu}\mathbf{p} + \mathbf{PT}_{t_{k,n}} + \mathbf{PP}_{p_{k,n}}$)能與觀察到的基頻軌跡(observer pitch contour; $\mathbf{x}_{k,n}$)非常相似，所以在重新標記韻律狀態的時候，為了使得標記為正確的韻律狀態機會提升，所以我們利用兩者之間的誤差 $\mathbf{y}_{k,n}$，而這個誤差可視為一個常態分佈；$N(\mathbf{0}, \mathbf{Rp})$，其 mean 為 0，covariance matrix 為 $\mathbf{Rp}$，計算式如下所示：

$$\mathbf{Rp} = \frac{\sum_{k=1}^{K}\sum_{n=1}^{N_k}(\mathbf{x}_{k,n} - \boldsymbol{\mu}\mathbf{p} - \mathbf{PT}_{t_{k,n}} - \mathbf{PP}_{p_{k,n}}) \times (\mathbf{x}_{k,n} - \boldsymbol{\mu}\mathbf{p} - \mathbf{PT}_{t_{k,n}} - \mathbf{PP}_{p_{k,n}})^T}{N_t} \tag{3.6}$$

- 步驟(7)，利用步驟(6)求得的 covariance matrix $\mathbf{Rp}$，依照 ML criterion 我們可以進行重新標記 tone 3、6、7 與 tone 5、8 的動作，並且再一次更新聲調模型與 pitch covariance matrix $\mathbf{Rp}$。

- 步驟(8)，利用上述步驟求得的 covariance matrix **Rp**，同樣依照 ML criterion 可以進行重新標記韻律狀態的動作，主要目的是讓每段基頻軌跡的韻律狀態能得到最佳的可能性，判斷式如下：

$$p_{k,n}^{*} = \arg\max_{p_{k,n}} \left\{ \sum_{k=1}^{K} \sum_{n=1}^{N_k} \log N(\mathbf{x}_{k,n}; \boldsymbol{\mu}\mathbf{p} + \mathbf{PT}_{t_{k,n}} + \mathbf{PP}_{p_{k,n}}, \mathbf{Rp} \mid t_{k,n}, p_{k,n}, \boldsymbol{\mu}\mathbf{p}) \right\} \quad (3.7)$$

- 步驟(9)，利用重新標記韻律狀態後的結果，重新計算與更新韻律模型，更新方式與步驟(4)更新聲調模型相相同，但原本一開始利用 VQ 所分出來的 16 個韻律狀態群組，經過步驟(6)之後，會將每一群距離中心比較的離散部分轉移到更接近的群組，因此韻律模型更加接近最佳化，同時亦能幫助聲調模型的準確性。

- 步驟(10)，重複步驟(5)～(9)，直到 likelihood function L 收斂為止，likelihood function L 定義如下：

$$\begin{aligned} L &= \log[P(\mathbf{x} \mid \mathbf{p}, \mathbf{t}, \boldsymbol{\mu}\mathbf{p})] \\ &\approx \sum_{k=1}^{K} \sum_{n=1}^{N_k} \log P(\mathbf{x}_{k,n} \mid p_{k,n}, t_{k,n}, \boldsymbol{\mu}\mathbf{p}) \\ &= \sum_{k=1}^{K} \sum_{n=1}^{N_k} \log N(\mathbf{x}_{k,n}; \mathbf{PT}_{t_{k,n}} + \mathbf{PP}_{p_{k,n}} + \boldsymbol{\mu}\mathbf{p}, \mathbf{Rp}_{t_{k,n}, p_{k,n}}) \end{aligned} \quad (3.8)$$

在步驟(10)的時候我們不斷檢查 L 的值與模型更新次數之間的變化關係，並且記錄下來，最後得到的結果如下圖所示：

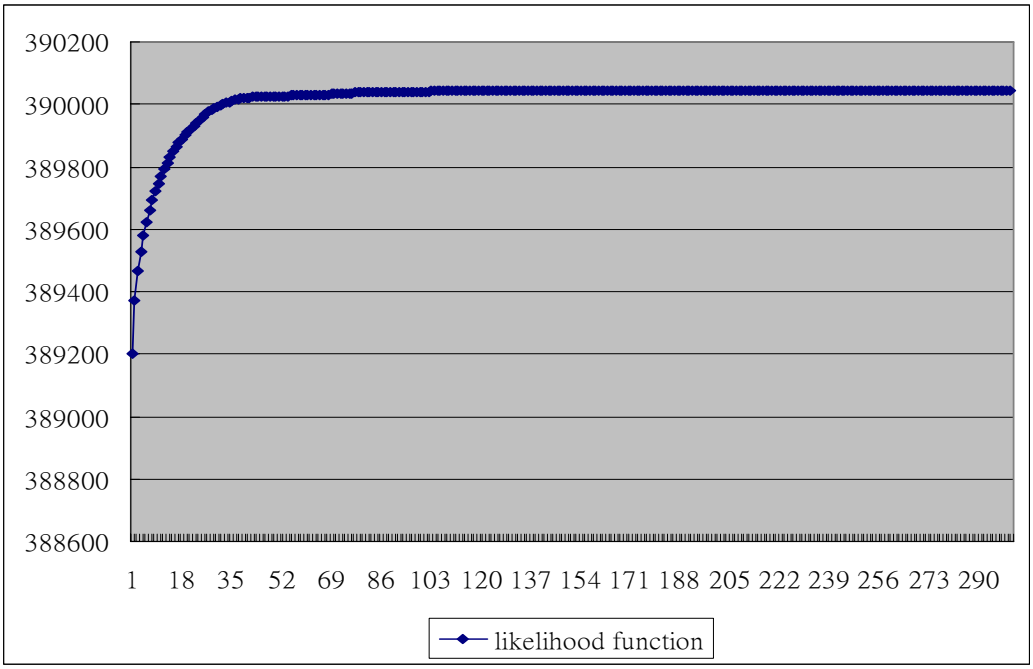


圖 3.6 likelihood function

從圖 3.6 我們可以推斷，重複更新模型達到約 150 次左右時，likelihood function 的值就不再出現大幅的變化，即是達到收斂的程度，於是便可以停止模型的更新動作，接著我們觀察訓練後的聲調模型與韻律模型是否如預期一樣，圖 3.7 為 speaker mean (**μp**)，圖 3.8(a)～圖 3.8(h)為中文聲調模型，其中 tone 3、tone 6 與 tone 7 皆屬於三聲，tone 5 與 tone 8 屬於 neutral tone，圖 3.9(a)～圖 3.9(d)為韻律狀態 1～韻律狀態 16 所呈現的韻律模型，下列圖 3.7～圖 3.9 的縱軸均為頻率值(log frequency)，橫軸為時間軸(frames)。

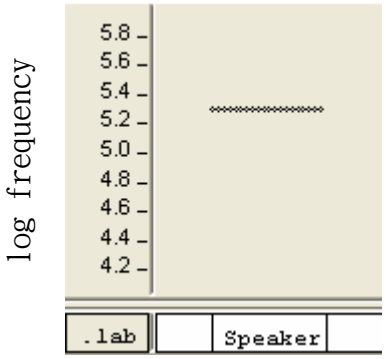


圖 3.7 speaker mean (**μp**)

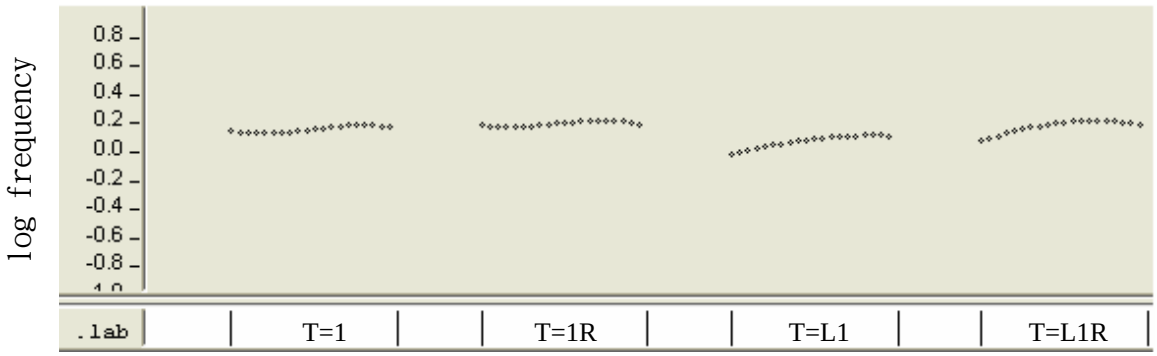


圖 3.8(a) tone 1 的聲調模型

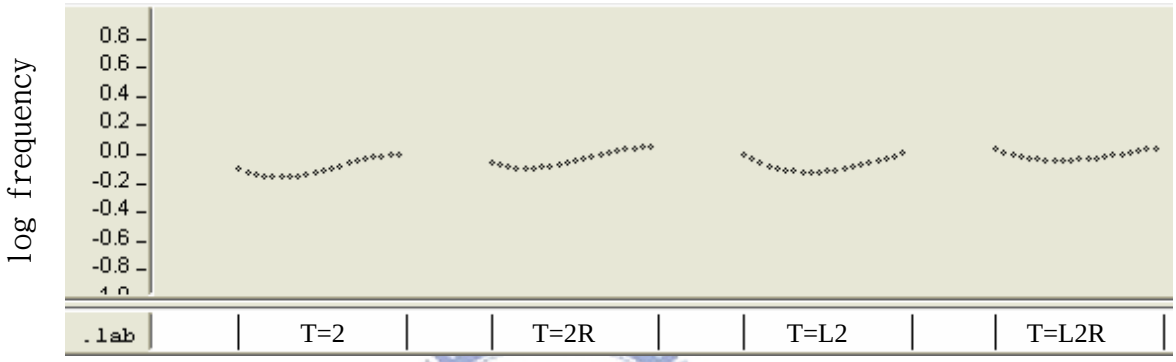


圖 3.8(b) tone 2 的聲調模型

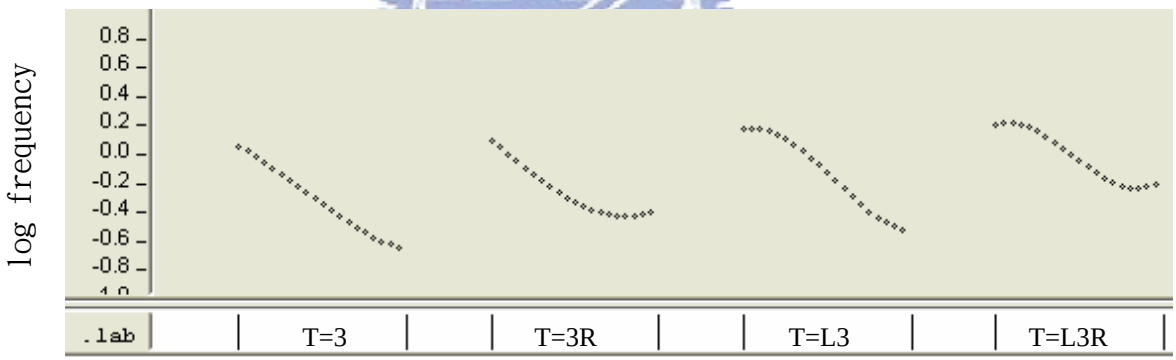


圖 3.8(c) tone 3 的聲調模型 (基頻軌跡呈現 falling-dipping)

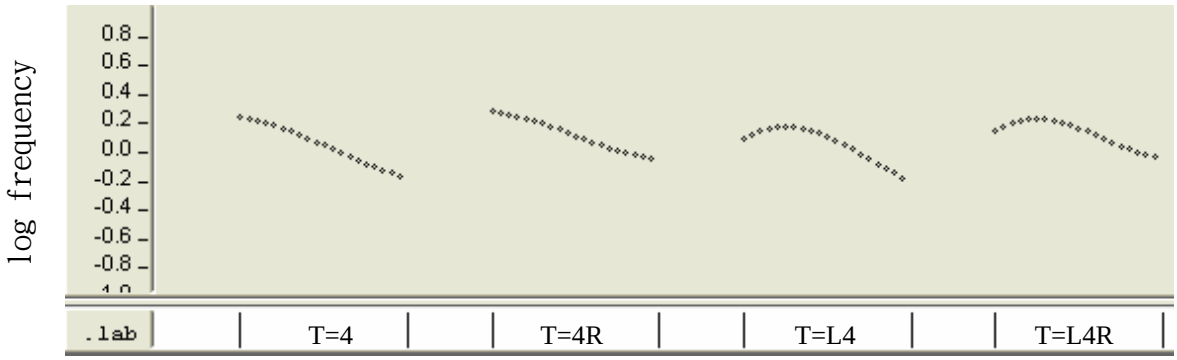


圖 3.8(d) tone 4 的聲調模型

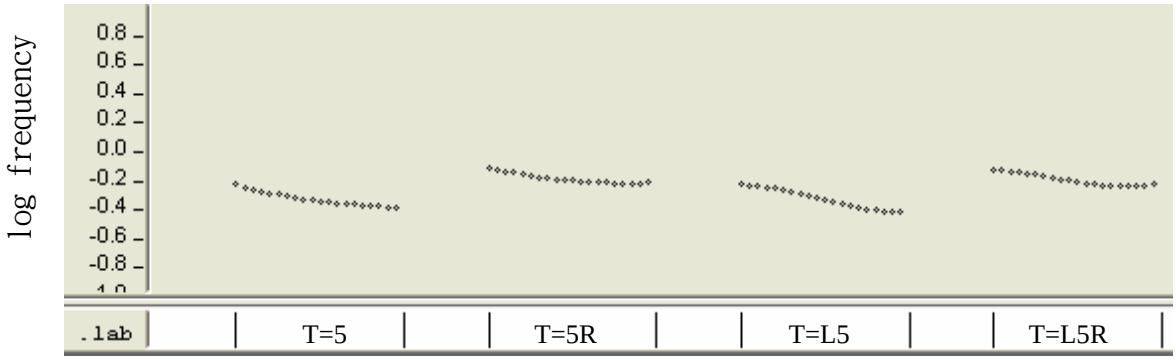


圖 3.8(e) tone 5 的聲調模型(基頻值較低的 neutral tone)

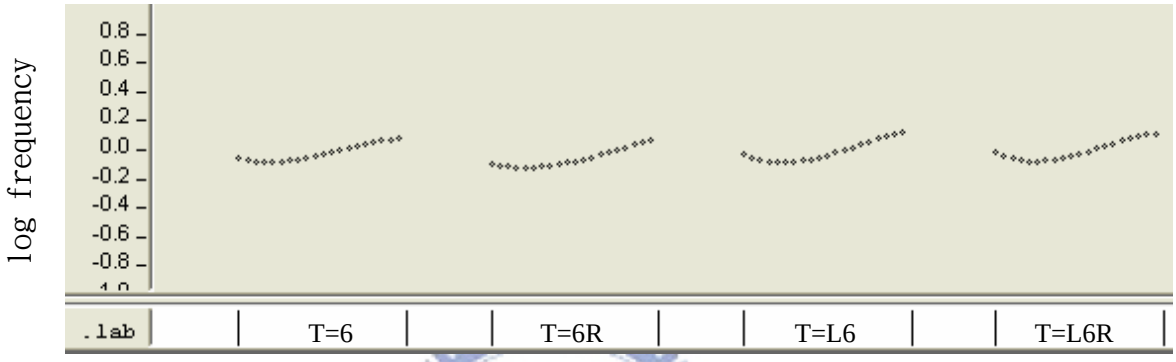


圖 3.8(f) tone 6 的聲調模型(經 tone-sandhi 影響的 tone 3)

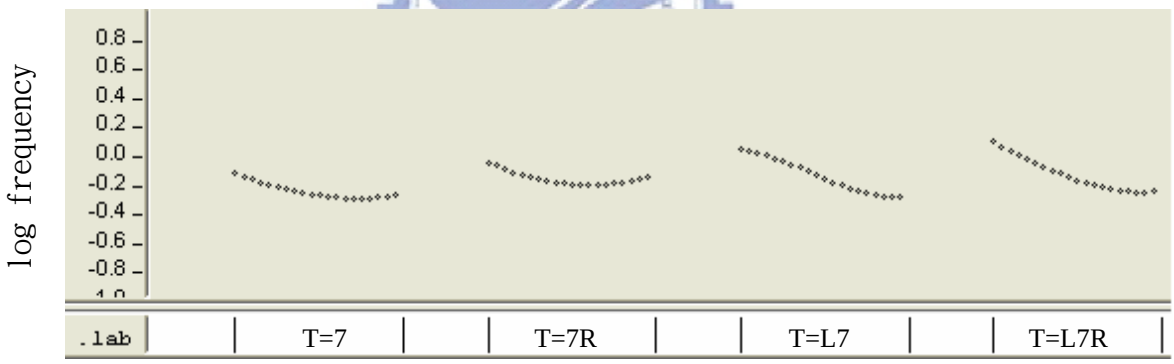


圖 3.8(g) tone 7 的聲調模型(基頻軌跡呈現 falling-rising 的 tone 3)

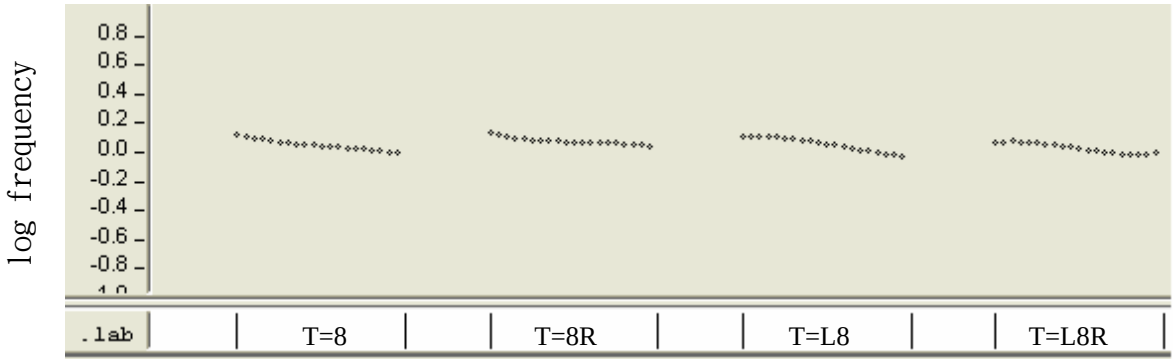


圖 3.8(h) tone 8 的聲調模型(基頻值較高的 neutral tone)

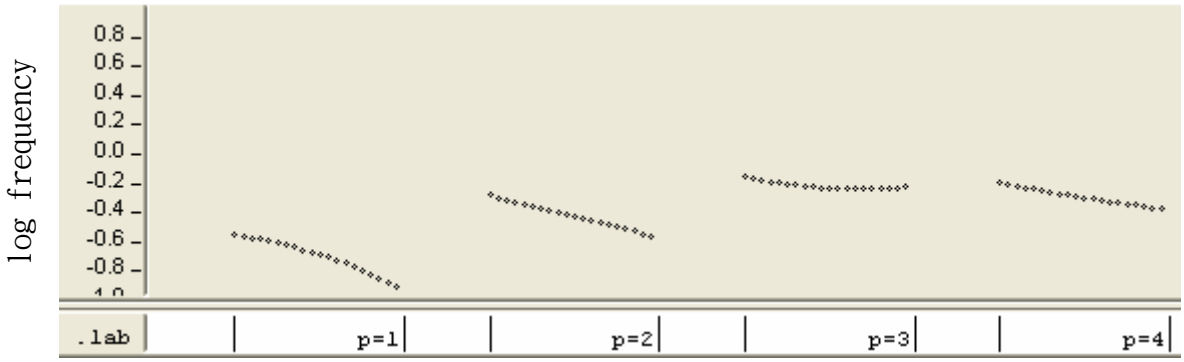


圖 3.9(a) 韻律狀態 1～4 的韻律模型

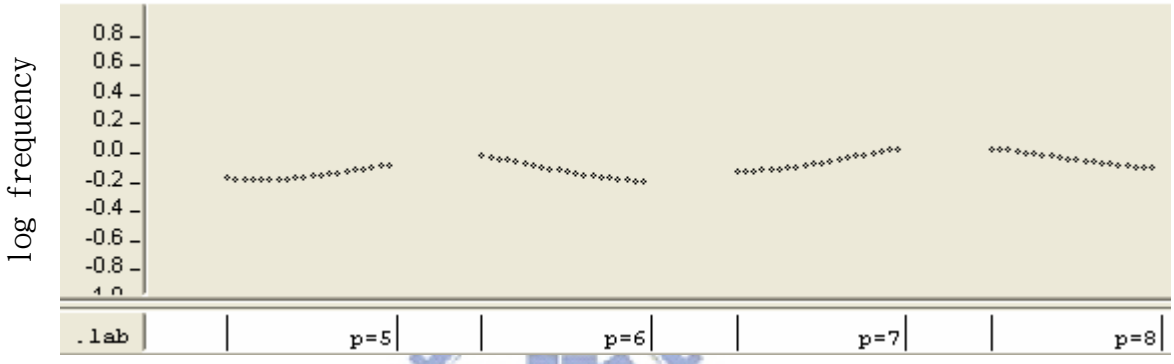


圖 3.9(b) 韻律狀態 5～8 的韻律模型

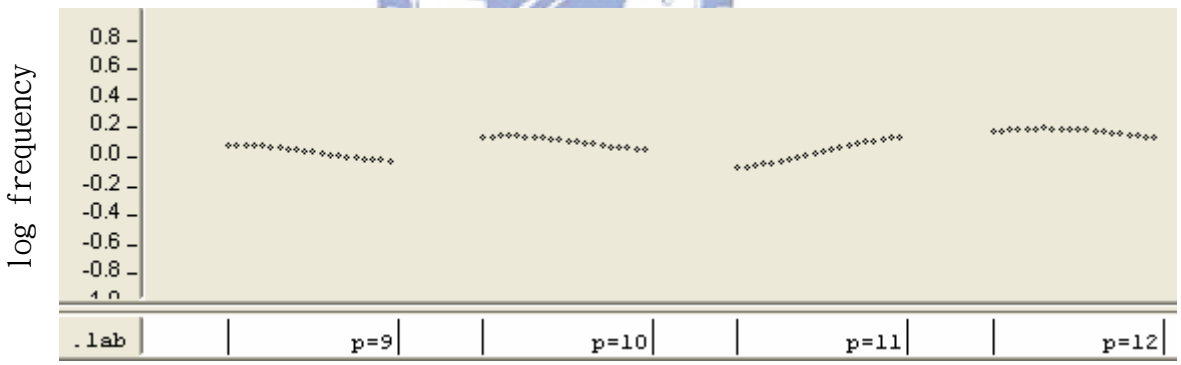


圖 3.9(c) 韻律狀態 9～12 的韻律模型

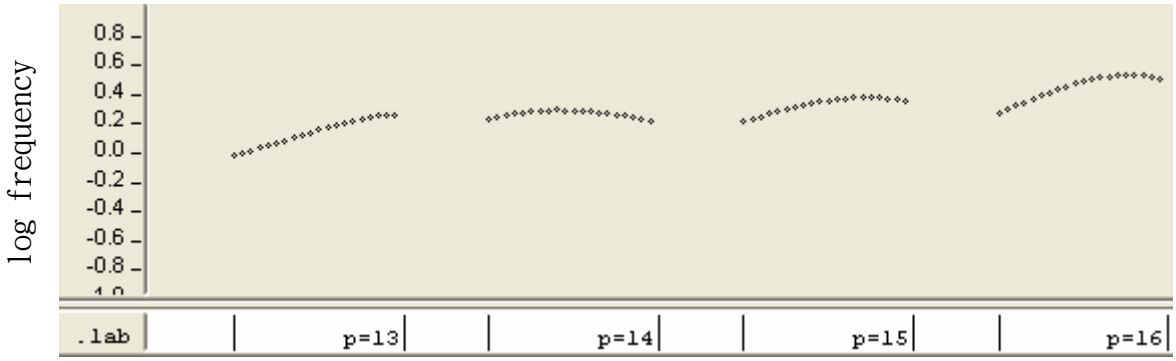


圖 3.9(d) 韻律狀態 13～16 的韻律模型

由上列的圖我們可以看出，基本上聲調模型很容易受到左右是否有連接其他聲調而影響，尤其在 tone 3 與 tone 4 可以很明顯看出影響處，對於韻律模型而言，每個韻律狀態與狀態之間最大的差異處，很明顯的在於基頻平均值，於是我們統計訓練後的韻律模型的在整個語料庫的分佈狀況，分佈如圖 3.10 所示，與每個韻律模型的平均基頻值的分佈，圖 3.11。

從圖 3.10 中我們可以發現除了少數機個韻律狀態外，其餘韻律狀態呈現均勻分佈的情形，而從圖 3.11 中可以看出，除了幾個數量較少的韻律狀態外(例如韻律狀態 1)，其餘每韻律狀態的基頻平均值分佈大致上亦呈現均勻分佈的情形，這表示說經過訓練過後的聲調模型加上韻律模型的樣版(pattern)，其能涵蓋的頻率範圍與解析度足夠模擬語料庫中大部分音節的基頻軌跡。

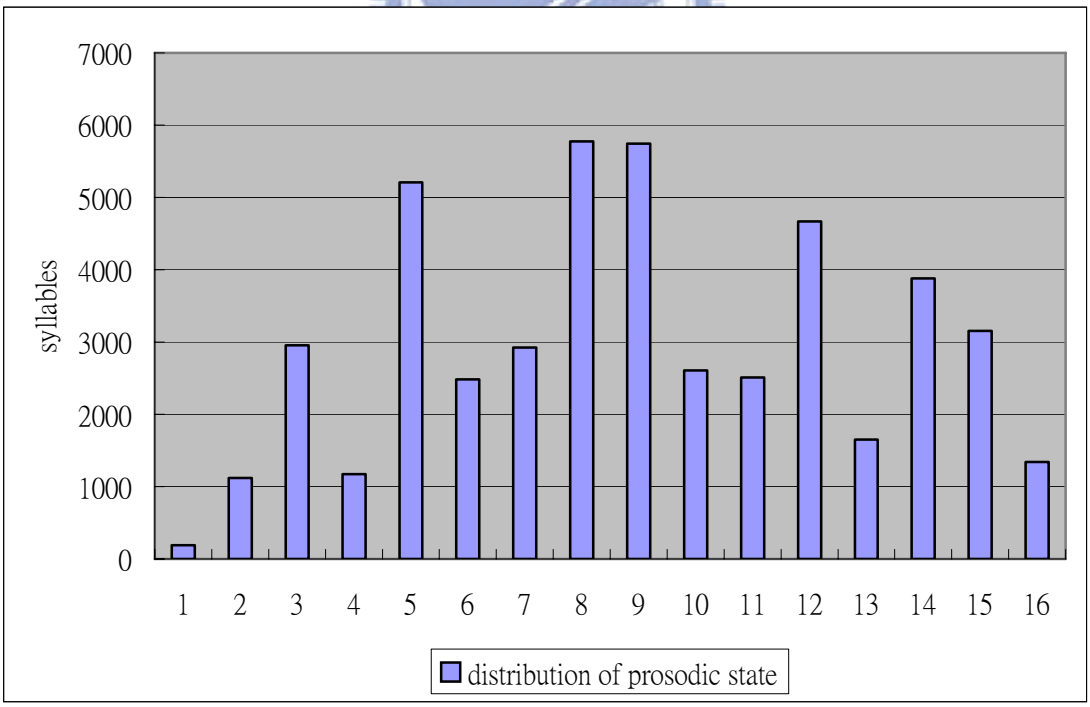


圖 3.10 韻律狀態分佈圖

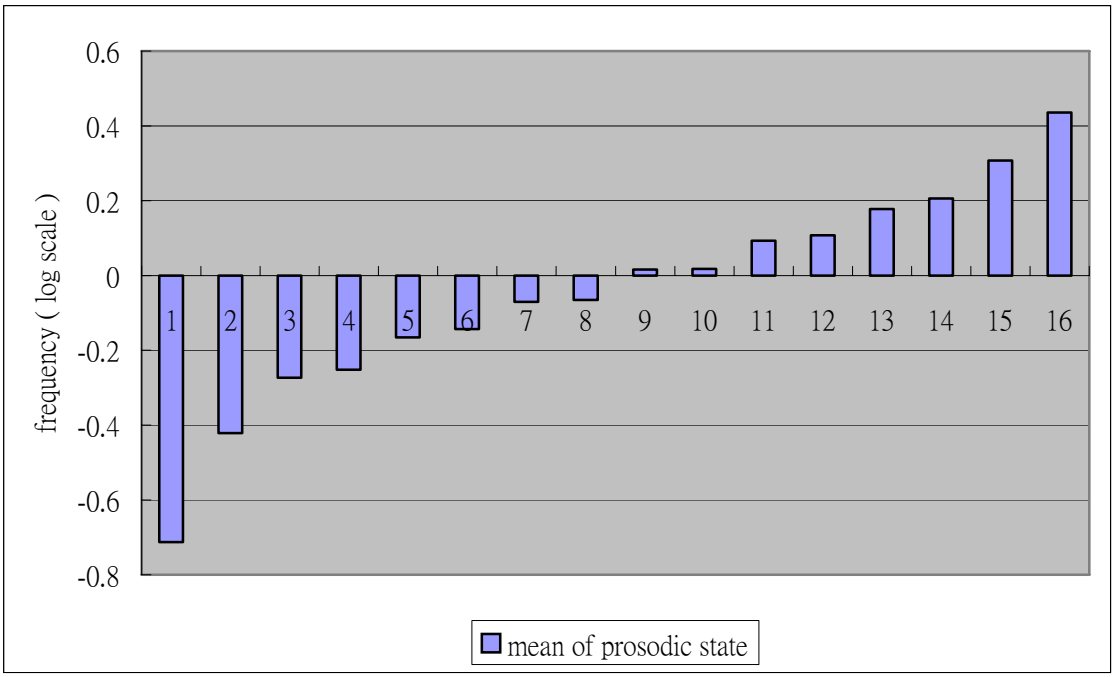


圖 3.11 各韻律狀態的基頻平均值分佈

3.3 整合聲調辨認與基頻軌跡建立

以往的聲調辨認都是基於先做基頻軌跡建立，然後再將抽取後的基頻軌跡進行聲調辨認，本論文提出一個整合架構，此架構可同時進行基頻軌跡建立與聲調辨認，由於在基頻軌跡建立時，常常發生 Double pitch 與 Half pitch 的錯誤，如果能加上聲調的模型同時輔助基頻軌跡的建立，相信上述的錯誤能大大減少，增加基頻軌跡的準確度，進而使聲調辨認率提升。

3.3.1 以音節為單位建立音節基頻軌跡候選者

由於如果以音框為單位進行基頻軌跡的建立的話，不僅僅計算量太大，而且某些基頻候選者的軌跡組合是不合理得，例如相鄰兩個音框基頻值變化過大，若

我們在同時進行基頻軌跡抽取與聲調辨認前，基於某些基本準則，例如基頻值變化量、候選者可靠度(η值;詳見本文 2.3.3 節)等等，先對基頻軌跡候選這進行第一步挑選，可以減少最後基頻軌跡建立錯誤的機率，所以在這提出以音節為單位建立基頻軌跡候選者，主要在每一個音節中，利用基頻候選者的基頻值與可靠度(η值)，利用維特比搜尋(Viterbi Search)挑選出前五名最佳基頻軌跡候選者。

下圖為基頻軌跡候選者搜尋示意圖，假設某個音節擁有 M 個 voiced 音框，從音節的起頭到結尾中的每個音框都可以挑選出五個 voiced 基頻候選值($F_1 \sim F_5$)，音框與音框之間間隔為 10 ms，其中 voiced 基頻候選值的選取方式為上一章所敘述，每次搜尋一共有五條路徑存在，一直搜尋到音節的最後一個音框為止。

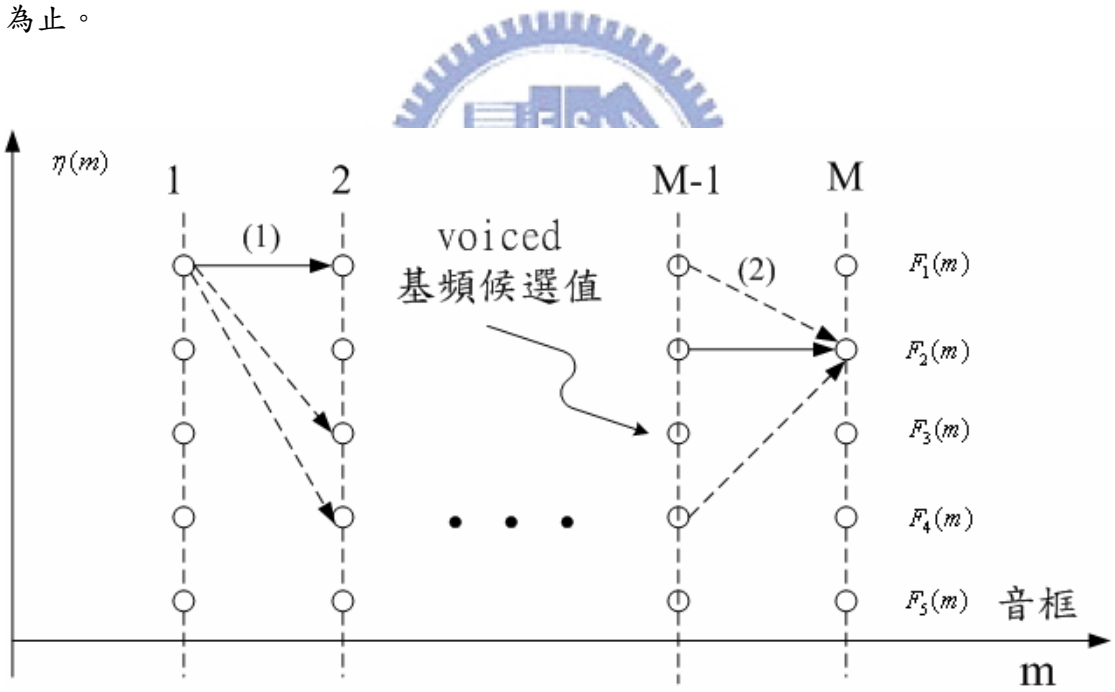


圖 3.12　基頻軌跡候選者搜尋示意圖

路徑決定方式為選擇一條最小誤差(cost，誤差)的路徑，每條路徑的誤差從第一個音框開始累積，直到音節結束，觀察各路徑累積的誤差值，選擇一條 cost 值最小的路徑，現在音框 m 所擁有的候選者數為 5 個，$\eta_{m,j}$ 為音框 m 中第 j 個候

選者的η值，於是我們可以定義本地誤差(local cost，本地誤差)如下：

$$d_{m,j} = 1 - \eta_{m,j} \tag{3.9}$$

接著我們定義音框與音框之間的轉移誤差(transition cost，轉移誤差)，假設當前音框為候選者j，與先前音框的候選者k之間的轉移誤差定義為(圖 2.19 中的路徑(1)與(2))：

$$\delta_{m,j,k} = FREQ\_WT \times \left| \ln(\frac{F_{m,j}}{F_{m-1,k}}) \right| \tag{3.10}$$

其中 $1 \le j \le 5, 1 \le k \le 5$，$FREQ\_WT$ 為正值(此設定為 3，請參考[9])

最後我們可以定義在音框 m 時計算累積誤差的遞迴式如下：

$$D_{m,j} = d_{m,j} + \min_{k \in 5}\{D_{m-1,k} + \delta_{m,j,k}\}, \quad 1 \le j \le 5 \tag{3.11}$$

初始值設定為，$D_{0,j} = 0, \quad 1 \le j \le 5$

在每個音框中的每一候選者均有一個指標，指向著前一音框中累積誤差最小路徑(minimize $D_{m,j}$)的候選者，當累積誤差計算到音節結束時，便可以決定出五條誤差最小的最佳路徑，此五條路徑即是我們所求得的基頻軌跡候選者。

最後的搜尋結果如下圖範例所示，圖 3.13 中，中間視窗為尚未做維特比搜尋的基頻值候選者，上視窗則為基頻候選者經過 voiced 區段性的維特比搜尋後所產生的基頻軌跡候選者，假設原本音節內共有 M 個 voiced 音框，每個音框共有五個基頻值候選者，則其所有可能的路徑一共有 5^M 條，經過初步的維特比搜尋後，可以把路徑減少成 5 條，如此便大大的減少不必要的基頻軌跡候選者與計算量。

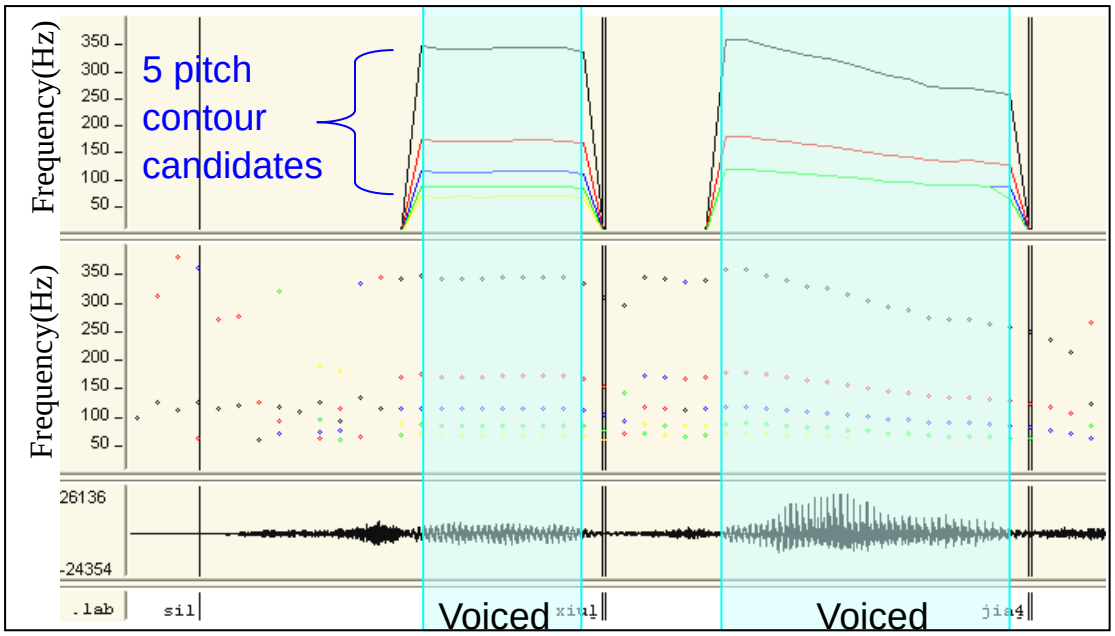


圖 3.13 以音節為單位建立基頻軌跡候選者範例

3.3.2 聲調辨認與基頻軌跡搜尋

在上一節中提到以音節為單位的基頻軌跡搜尋，主要是為了減少計算量以及提供可靠的基頻軌跡候選者，接著這一節將延續上一節所產生的結果，並以音節為單位對整個句子進行基頻軌跡的建立與連續音節聲調辨認，方塊圖如下所示，在進行聲調辨認與基頻軌跡搜尋之前，必須先擁有以音節為單位的基頻軌跡候選者，以及已經訓練好的聲調模型與韻律模型，還有音節邊界(syllable boundary)與有聲邊界(voiced boundary)，接著，以音節為單位，而每個音節擁有 512 個聲調與韻律組合，利用維特比搜尋方式，尋找出句子的最佳基頻軌跡，與每個音節的聲調辨認結果。

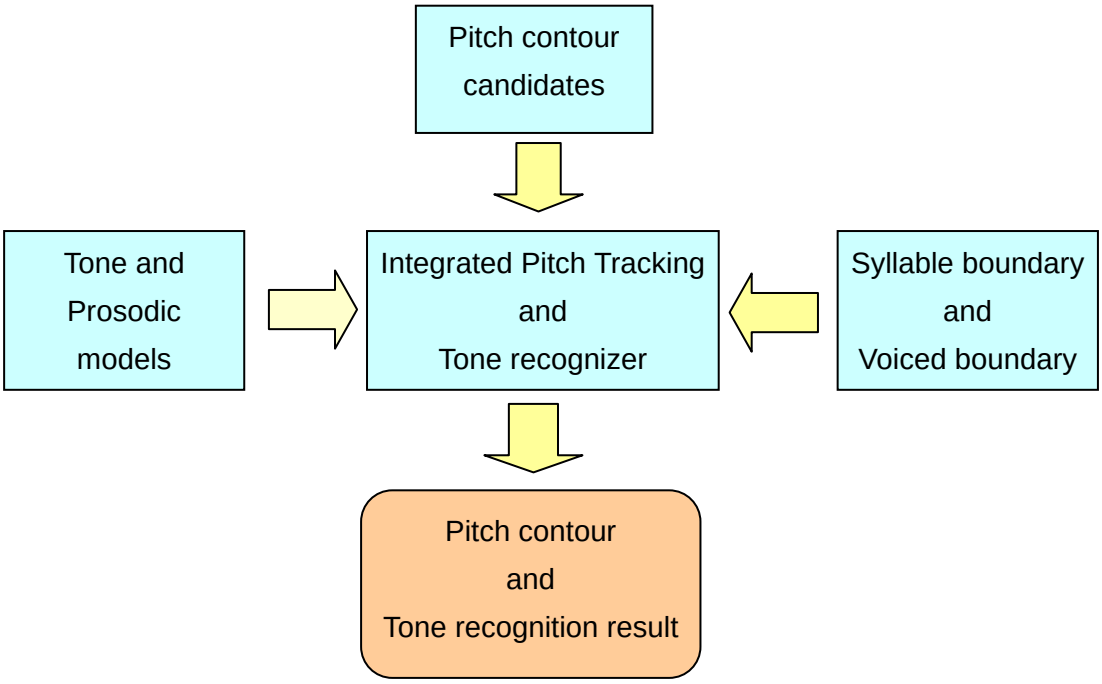


圖 3.14 整合基頻軌跡搜尋與聲調辨認系統方塊圖

完整的基頻軌跡與聲調辨認之維特比搜尋示意圖如下所示，以一段句子的開始與結束為搜尋邊界，而示意圖的橫軸為句子的音節數(假設共有 N 個音節)，縱軸為此音節聲調與韻律狀態組合，所以縱軸為 32 個聲調狀態 $T \in \{1,2\cdots 8, L1\cdots, L8, 1R\cdots, 8R, L1R\cdots, L8R\}$ 乘上 16 個韻律狀態 $P \in \{1,2\cdots 16\}$，共 512 種組合的可能性。

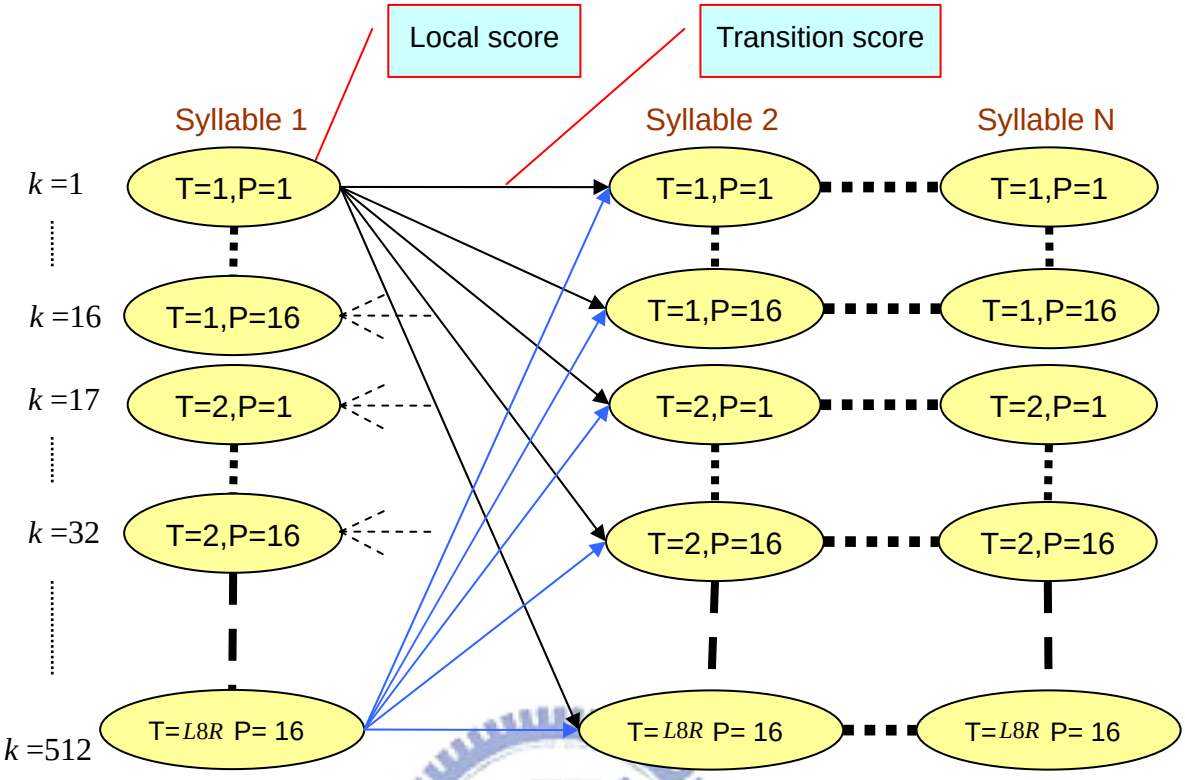


圖 3.15 基頻軌跡與聲調辨認維特比搜尋示意圖

以搜尋一條最高分數路徑為目標，影響每條路徑分數的高低共有兩個因素，本地分數(S_{Local};Local score)與轉移分數(S_{Trans};Transition score)，而本地分數中包含了：

1.基頻軌跡分數(S_{IFAS};IFAS Score)

2.聲調模型分數(S_{model};Tone Modeling Score)

轉移分數內包含了：

1.聲調轉移機率(TTP;Tone Transition Probability)

2.韻律狀態轉移機率(PTP;Prosodic Transition Probability)

首先我們定義第 n 個音節，第 k 種組合的本地分數 $S_{Local}(n,k)$ 如下：

$$S_{Local}(n,k)=\arg\max_{c}\{W_{\text{model}}\times S_{\text{model}}(n,k,c)+W_{IFAS}\times S_{IFAS}(n,k,c)\} \qquad (3.12)$$

其中$1 \le n \le N, 1 \le k \le 512$，$W_{\text{model}}$與$W_{IFAS}$分別為聲調模型分數與基頻軌跡分數的權重(weight)，均為正值，

$$S_{\text{model}}(n,k,c) = \log N(O_{n,c}; \mathbf{PT}_{t(n,k)} + \mathbf{PP}_{p(n,k)} + \boldsymbol{\mu}\mathbf{p}, \mathbf{Rp}) \tag{3.13}$$

$$S_{IFAS}(n,k,c) = \frac{\sum_{m=1}^{M} \log(\eta_{m,c})}{M} \quad ,1 \le m \le M \tag{3.14}$$

欲決定第n音節，第k組合的分數$S_{Local}(n,k)$時，搜尋音節n的所有候選基頻軌跡，選擇基頻軌跡c，其正交展開式為$O_{n,c}$與可靠度$\sum_{m=1}^{M} \log(\eta_{m,c})$，使得式子(3.12)獲得最大值，並利用指標$q(n,k)$紀錄所選擇的基頻軌跡$c$，式子(3.13)中$t(n,k)$與$p(n,k)$分別為第$n$音節第$k$組合的聲調與韻律狀態，當式(3.12)得到最大值時，表示候選基頻軌跡c與此組合的聲調狀態、韻律狀態非常相似，而且同時具有相當的可靠度，式子 3.14 中M為音節n所擁有 voiced 音框的個數，其餘變數如**PT**、**PP**、**μp**、**Rp**定義皆與上一章相同。

接著我們定義由k_{n-1}聲調韻律組合轉移到k_n組合的轉移分數$S_{Trans}(k_{n-1}, k_n)$定義如下式子(3.15)所示，：

$$S_{Trans}(k_{n-1}, k_n) = \log TTP_{t(n-1,k_{n-1}),t(n,k_n)} + \log PTP_{p(n-1,k_{n-1}),p(n,k_n)} \tag{3.15}$$

其中，$2 \le n \le N$，$1 \le k_{n-1}, k_n \le 512$

$TTP_{t(n-1,k_{n-1}),t(n,k_n)}$為音節$n-1$中第$k_{n-1}$組合的聲調狀態，$t(n-1,k_{n-1})$，轉移到音節$n$的第$k_n$組合的聲調狀態，$t(n,k_n)$，之轉移機率，而$PTP_{p(n-1,k_{n-1}),p(n,k_n)}$為音節$n-1$中第$k_{n-1}$組合的韻律狀態，$p(n-1,k_{n-1})$，轉移到音節$n$的第$k_n$組合的韻律狀態，$p(n,k_n)$，之轉移機率，其中所使用到的聲調、韻律轉移機率($TTP$、$PTP$)與聲

調、韻律狀態初始機率($P_{t(1,k_1)}$、$P_{p(1,k_1)}$)均為語料庫中所統計而得。

最後，我們可以定義在第n個音節時計算累積分數的遞迴式如下：

$$SD(n,k_n)=W_{Local}\times S_{Local}(n,k_n)+\min_{k_{n-1}\in 512}\{SD(n-1,k_{n-1})+W_{Trans}\times S_{Trans}(k_{n-1},k_n)\} \quad (3.16)$$

初始值，$SD(1,k_1)=W_{Local}\times S_{Local}(1,k_1)+W_{Trans}\times(\log P_{t(1,k_1)}+\log P_{p(1,k_1)})$ (3.17)

其中，$2\le n\le N$，$1\le k_{n-1},k_n\le 512$，W_{Local}與W_{Trans}分別為本地分數與轉移分數的權重(weight)，兩者均為正值。

於是，當我們在句子最後一個音節選擇了最高分數的一條路徑的同時，亦對每個音節選擇了一條最佳的基頻軌跡候選，$q(n,k)$，與所選擇到的基頻軌跡候選之聲調辨認結果，$t(n,k_n)$。

第四章　實驗結果與分析

本章將介紹實驗所使用的語料庫，說明其錄製對象、內容與錄製情形，與基頻軌跡與聲調辨認細節參數的調整順序與實驗結果比較，並將比較結果整理成表格。

4.1 使用語料

目前所使用之語料庫是由一位專業的女性廣播人員所錄製，錄音內容為提供一篇文字稿請語者照著文字稿流利唸出，在經由麥克風所錄製而成，其文字稿的文字部分來自於「中央研究院中文文句結構樹資料庫 1.1 版」(Sinica Treebank Version 1.1)[10]，從中央研究院詞庫小組之「中央研究院現代漢語語料庫」得來，其錄音相關設定如下表所示。一共為 380 個音檔的乾淨語料，共 52,148 個音節，其中使用 342 個音檔作為訓練語料，共有 47,380 音節，外部測試選用了剩餘 38 個音檔，共 4,768 音節。

表 4.1 錄音相關設定

錄音軟體	Cool Edit Pro 直接錄成聲音檔案
麥克風	單一指向性 (uni-directional)
錄音場所	普通房間
錄音情境	依照所選出文稿唸出
取樣頻率(sampling rate)	20 kHz，down sampling 至 16k
發音速度	每秒約 4.6 個音節
取樣大小	16 bits (位元)
聲道	單聲道(mono)
檔案格式	PCM
能量（句平均）	平均 60.81dB，最小 52.18dB，最大 66.48dB

為了檢驗整合式的基頻軌跡建立結果，我們利用常用的 ESPS(Entropic Corp.)

軟體中基頻軌跡求取程式對整個語料庫求取基頻軌跡，接著利用人工逐一檢查基頻軌跡的方式，修正 ESPS 求取之基頻軌跡有問題的地方，最後修正完的基頻軌跡便可視為基頻比對的參考基頻(reference pitch)，接著統計整個語料庫中參考基頻所含有 voiced 音框與 unvoiced 音框的數量，整理成下表 4.2 所示：

表 4.2 語料庫有聲音框數統計表(單位為 frame)

	訓練語料	外部測試語料
Voiced	635,896	64,655
Unvoiced	647,302	64,915
總計	1,283,198	129,570
語料庫總計	1,412,768	

4.2 參數微調

在上一章的最後一節中，我們在式子中定義了幾個權重(W_{model} 、W_{IFAS} 、W_{Local} 、W_{Trans})，為了使得基頻軌跡建立能有更佳的結果，我們進行權重之間的微調動作，主要是因為，不同的分數擁有各自的值域變化，為了使得每個分數能擁有一致的值域，所以必須壓低或是相對拉高某一分數。過程中我們一邊觀察基頻的 Gross Pitch Error (GPE)值[7]，單位為百分比(%)，一邊適當的調整某一個權重，GPE 的定義如下：

$$GPE = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{E_k}{E_{\max}}\right)^{1/2}\left|\frac{f_k - \hat{f}_k}{\hat{f}_k}\right| \tag{4.1}$$

其中，K 為參考基頻中有聲(voiced)的音框個數，E_k 為句子中第 k 個音框的能量(short-time energy)，$E_{\max}$ 為整個句子最大的音框能量，f_k 為句子中第 k 個音框，基頻軌跡偵測器所估測的基頻值，$\hat{f}_k$ 為句子中第 k 個音框，參考基頻的基頻值。

式子(4.1)中利用音框能量跟最大音框能量之間的比值，作為此音框重要性的依據，如果當前音框的能量與句子中最大音框能量相當接近時，$\frac{E_k}{E_{\max}}$ 此項就會接近於 1，那麼這個音框的基頻值與參考基頻值之間的差異對於最後的 GPE 值更是有影響力，反之則否。

觀察(3.12)式，

$$S_{Local}(n,k)=\arg\max_{c}\{W_{\text{model}}\times S_{\text{model}}(n,k,c)+W_{IFAS}\times S_{IFAS}(n,k,c)\}$$

與(3.16)式，

$$SD(n,k_n)=W_{Local}\times S_{Local}(n,k_n)+\min_{k_{n-1}\in 512}\{SD(n-1,k_{n-1})+W_{Trans}\times S_{Trans}(k_{n-1},k_n)\}$$

式子中四個權重之間的調整依據來自於每一個數值的變動範圍(dynamic range)，而調整的順序以影響結果的重要性為優先，以本地分數與轉移分數而言，本地分數對於最後結果的影響力大於轉移分數，所以先對本地分數內部的兩個分數之間的權重做調整，即是 W_{model} 與 W_{IFAS} 。

所以我們接著分析 $S_{\text{model}}(n,k,c)$ 、 $S_{IFAS}(n,k,c)$ 兩變數的數值變動範圍，發現 $S_{\text{model}}(n,k,c)$ 的範圍介於10 ～ -700左右，而 $S_{IFAS}(n,k,c)$ 的範圍介於0 ～ -7左右，很明顯的，$S_{\text{model}}(n,k,c)$ 數值變動範圍大上許多，所以希望藉由調整權重使得兩者擁有大致相同的數值範圍，於是我們調整 W_{model} 與 W_{IFAS} 之間的比例，在一段合理的範圍內嘗試性的實驗，並記錄每次實驗結果與參考基頻比較後的GPE值，實驗紀錄如下圖4.1所示，橫軸為每次嘗試的比值($\frac{W_{IFAS}}{W_{\text{model}}}$)，縱軸為實驗結果的GPE，從圖中我們可以看出，最佳的權重比大約是在1：160左右，接著，我們再以更細微的間隔在1：160附近搜尋，結果紀錄在圖4.2中，最後我們可以發現大約在1：159的時候可以得到最低的GPE值(0.949%)。

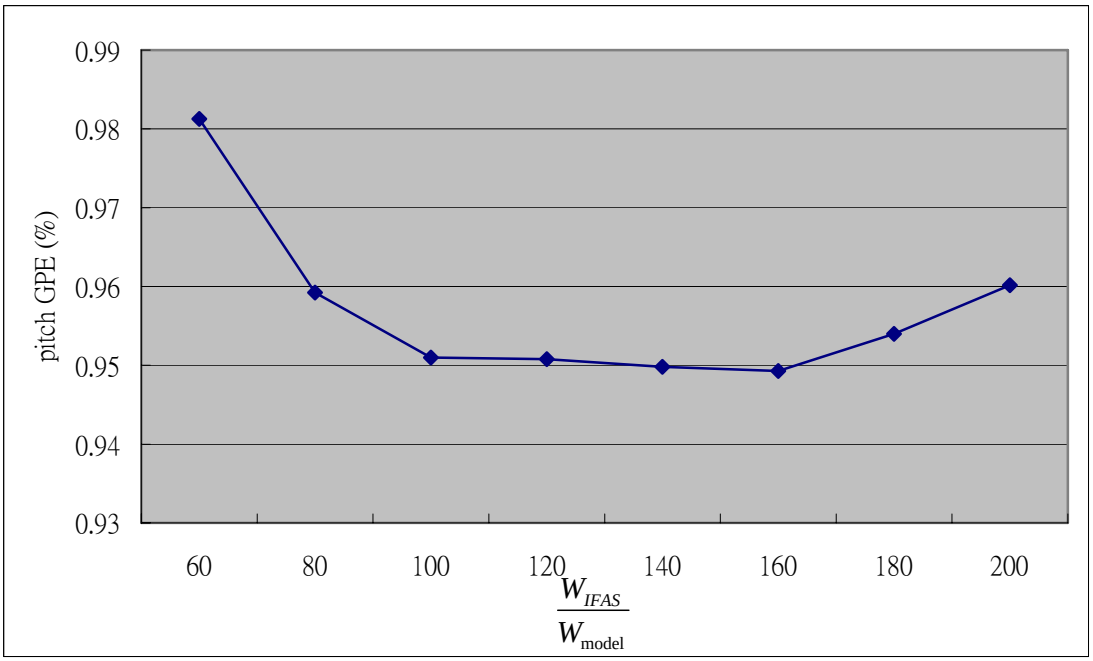


圖 4.1 權重值調整紀錄1

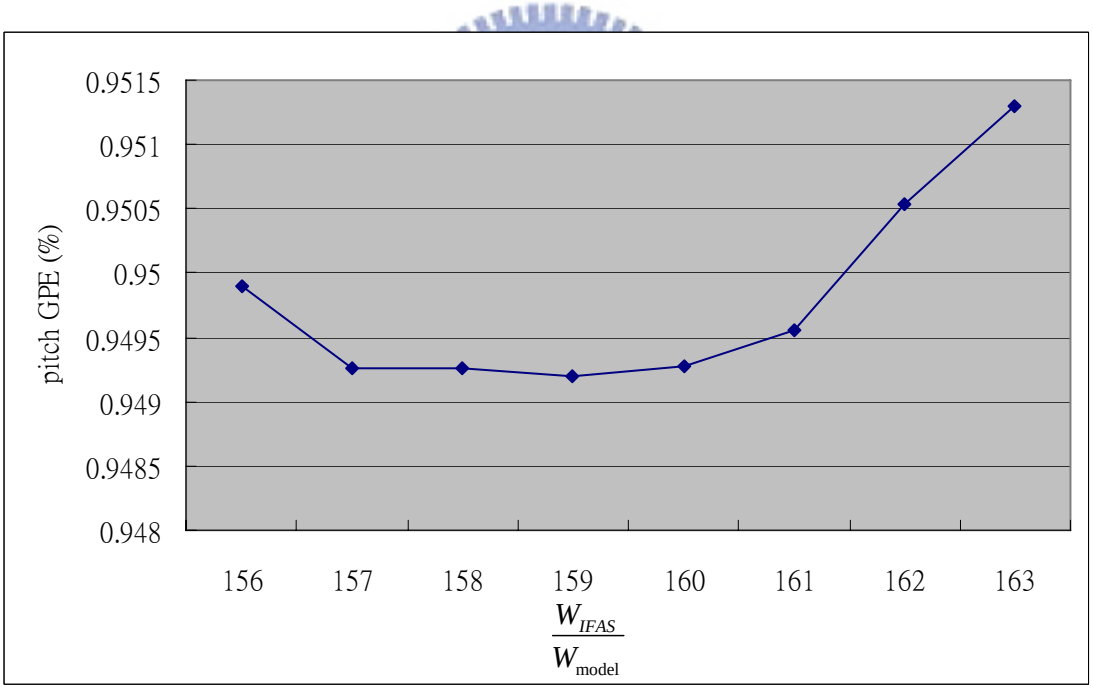


圖 4.2 權重值調整紀錄2

在調完W_{model}與W_{IFAS}之間的比值後，接著我們開始嘗試調整W_{Local}與W_{Trans}之間的比值，同樣的，一開始利用大範圍的嘗試，接著在可能性較大的區域用更細微的間隔搜尋，最後在大約時13得到最低GPE值，實驗結果記錄如下圖：

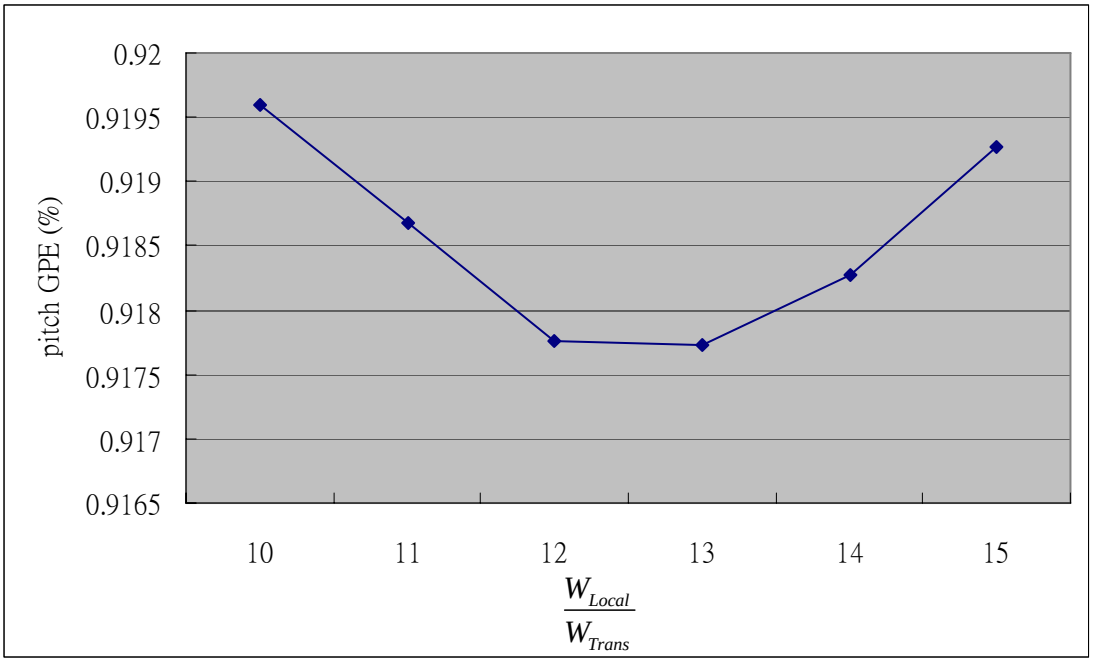


圖 4.3 權重值調整紀錄3

經由上述我們可以找到，當$\frac{W_{Local}}{W_{Trans}}$=13與$\frac{W_{IFAS}}{W_{\text{model}}}$=159的時候，基頻軌跡搜尋可以得到最低的GPE值約0.918%，比較基本型基頻偵測器(只單靠候選者的基頻值與相對應的可靠度η值，不包含聲調模型與韻律模型的輔助)估測出來的基頻軌跡，其GPE值為1.121%，可以證實，擁有聲調、運律模型以及轉移機率為輔助建立的基頻軌跡比單純只靠音框可靠度還要準確，最後整理成下表， 表中Up bound是指從基頻候選者中挑選一條與參考基頻最相似的基頻軌跡所得到的GPE。

表4.3 基頻偵測器比較表

	使用模型輔助之 基頻偵測器	基本型基頻偵測器
Voiced/unvoiced boundary	包含	包含
基頻值與可靠度	包含	包含
聲調與韻律模型與轉移機率	包含	不包含
GPE (%)	0.918	1.121
Up bound (%)	0.711	

4.3音節間及音框間之基頻值平均值的比較

4.3.1 音節間基頻平均值的比較(R_{mean_ratio})

除了利用 GPE 來觀察經過權重調整後的基頻軌跡建立之準確性外，本節將以每一個獨立的音節為單位，在此音節的區間裡將所有的基頻值求其平均值，代表此音節的基頻參考值，同時與人工標音的部分求其比值，觀察比值的分佈便可以瞭解基頻值以音節為單位的準確性為何。

以參考基頻值為正確值($f_{correct}$)，測試比較的基頻頻率為測試頻率(f_{test})，同時以音節位置為依據。在獨立音節(single-syllable)的區間內，將頻率不為零的音框(假設共包含 N 個 frame)相加作平均，作為正確音節的基頻平均值($f_{correct\_mean}$)與測試音節的基頻平均值(f_{test_mean})。數學式表示如下：

$$f_{correct\_mean}=\frac{\sum_{i=b,f_{correct}(i)\neq 0}^{e} f_{correct}(i)}{\sum_{i=b,f_{correct}(i)\neq 0}^{e} 1},\ f_{test\_mean}=\frac{\sum_{i=b,f_{test}(i)\neq 0}^{e} f_{test}(i)}{\sum_{i=b,f_{test}(i)\neq 0}^{e} 1} \tag{4.2}$$

其中b與e分別為獨立音節的起始音框與結束音框。我們定義兩音節基頻平均值的比為R_{mean_ratio}，則數學表示式為

$$R_{mean\_ratio}=\frac{f_{test\_mean}}{f_{correct\_mean}} \tag{4.3}$$

將R_{mean_ratio}從 0.5~1 之間取其倒數作為比值相同的範圍作統計，得結果如下表 4.4，並且引用文獻[11]中使用的統計法基頻軌跡偵測器所統計的資料，統計的方法與文獻中相同，並整理在表 4.5 內。

表 4.4 音節間平均基頻值的比較表

	使用模型輔助之 基頻偵測器	基本型基頻偵測器
總音節個數	46,861	
比值範圍	範圍內音節個數百分比	範圍內音節個數百分比
0.9~1.1	97.9 %	97.0 %
0.8~1.2	98.7 %	97.6 %
0.7~1.4	98.9 %	97.8 %
0.6~1.7	99.1 %	98.0 %
0.5~2.0	99.1 %	98.1 %

表 4.5 參考文獻之音節間平均基頻值的比較表

	統計法之基頻軌跡偵測器 (調整模型後)	ESPS
測試比較音節數	36,233	36,751
比值範圍	範圍內音節個數百分比	範圍內音節個數百分比
0.9~1.1	96.2 %	94.8 %
0.8~1.2	97.0 %	96.3 %
0.7~1.4	97.7 %	97.1 %
0.6~1.7	98.2 %	97.7 %
0.5~2.0	99.2 %	98.7 %

除了所有音節的基頻值比值外，我們另針對基本型與模型輔助基頻偵測器所估測之不同處做比較，以音節為單位，若同一音節內基本型與模型輔助型偵測器所估測出來的基頻軌跡不一致，則兩者分別與參考基頻作音節基頻平均值比較，最後得到的結果整理在表 4.6 中。

從表 4.6 中很明顯可以發現，模型輔助型有效地改善了基本型基頻偵測器的準確度，大約修正了 20%左右基本型基頻偵測器所估測錯誤的音節，而基本型基頻偵測約有 25%左右的相異音節是落在比值的範圍外，而模型輔助型只有約 5%左右落在範圍外。

表 4.6 基本型與模型輔助型音節間相異比較

	使用模型輔助之 基頻偵測器	基本型基頻偵測器
相異音節個數	3,004	
比值範圍	範圍內音節個數百分比	範圍內音節個數百分比
0.9~1.1	83.4 %	67.5 %
0.8~1.2	91.6 %	70.1%
0.7~1.4	93.6 %	72.9 %
0.6~1.7	95.2 %	76.5 %
0.5~2.0	95.3 %	76.9 %

4.3.2 音框間基頻值的比較(R_{ratio})

比較過每個音節區間內基頻平均值的分佈後，如果將每個比較的單位再更細分為每個音框來做比較，則可以更清楚的瞭解求得的基頻軌跡除了在音節部分的分佈外，每個音框的基頻值是不是也具有相當的正確性。

比較方法如同前項音節間比較一般，以參考語句基頻值為正確值($f_{correct}$)，測試比較的基週期頻率為測試頻率(f_{test})，同時以音節位置為依據。在為獨立音節的區間內，以每個音框為單位下，當兩者都有求出基頻值($f_{test} \neq 0$ 且 $f_{correct} \neq 0$)的情況下，才作比較，我們定義兩音框的基頻值比為 R_{ratio}，則數學表示式為

$$R_{ratio} = \frac{f_{test}}{f_{correct}} \quad (4.4)$$

將 R_{ratio} 從 0.5~1 之間取其倒數作為比值相同的範圍作統計，得結果如下表 4.7，並引用文獻[11]中使用相同統計方法的統計資料整理在表 4.8 中一起比較。

表 4.7 音框間基頻值的比較表

	使用模型輔助之 基頻偵測器	基本型基頻偵測器
總音框個數	616,337	
比值範圍	範圍內音框個數百分比	範圍內音框個數百分比
0.9~1.1	97.6 %	97.0 %
0.8~1.2	98.4 %	97.7 %
0.7~1.4	98.7 %	98.1 %
0.6~1.7	99.1 %	98.4 %
0.5~2.0	99.1 %	98.5 %

表 4.8 參考文獻之音框間基頻值的比較表

	統計法之基頻軌跡偵測器 (調整模型後)	ESPS
測試比較音框數	516,788	593,440
比值範圍	範圍內音框個數百分比	範圍內音框個數百分比
0.9~1.1	97.1 %	95.7 %
0.8~1.2	97.5 %	96.7 %
0.7~1.4	97.8 %	97.2 %
0.6~1.7	98.0 %	97.6 %
0.5~2.0	99.0 %	98.7 %

由上一節以音節為單位比較與此節以音框為單位之比較表中，可以發現使用聲調、韻律模型輔助之基頻抽取所得到的基頻軌跡，比基本型與統計型基頻軌跡偵測器還更準確，同時，倍頻與半頻的錯誤也有改善，接著，與上一節相同，比較基本型與模型輔助型之間相異的音框，結果如下表所示。

在表 4.9 中可以發現，使用模型輔助之基頻偵測器準確的音框數比基本型約多了 30%的相異音框數，而且在範圍外的音框個數上，模型輔助型比基本型少了約 34%左右，可見模型輔助型在半頻與倍頻錯誤上發生率比基本型還要少上許多。

表 4.9 基本型與模型輔助型音框間相異比較

	使用模型輔助之 基頻偵測器	基本型基頻偵測器
相異框個數	13,036	
比值範圍	範圍內音框個數百分比	範圍內音框個數百分比
0.9~1.1	55.1 %	25.1 %
0.8~1.2	64.9 %	33.0 %
0.7~1.4	71.7 %	40.1 %
0.6~1.7	82.2 %	47.3 %
0.5~2.0	84.2 %	48.6 %

4.4 連續語音聲調辨認

在第三章我們有提到如何利用聲調模型與韻律模型同時進行基頻軌跡建立與聲調辨認，主要是依據搜尋聲調模型與韻律模型的組合，比對基頻軌跡的圖樣(pattern)，挑選每個音節最相似的一組聲調韻律組合，便可得到句子內各個音節最相似的聲調模型。

使用的語料共 52,148 個音節，訓練語料有 47,380 音節，外部測試則有 4,768 音節，其中第一聲至第五聲之聲調音節數分佈如表 4.10。

表 4.10 第一聲至第五聲之音節分佈統計

	一聲	二聲	三聲	四聲	五聲	總音節數
訓練語料	33848	11510	8532	16143	2473	47380
外部測試語料	934	1033	937	1613	251	4768

根據文獻[12]的研究，中文連續語音在連續出現第三聲時(3-3 或 3-3-3)，會有將第三聲發成第二聲變調(tone sandhi)的狀況，所以我們試圖對語料庫中的所有的三聲分類成三類，將其中類歸為屬於發生變調的三聲(標成 tone6)，希望可以將這些變調成二聲的音節從三聲中挑出。

實驗一：

使用參考基頻進行聲調辨認，得到的辨識率如下表，表 4.11 的 tone 1 為“一聲”，tone 2 為“二聲”，tone 3、6、7 為“三聲”，tone 4 為“四聲”，tone 5、8 為“五聲”，詳細請見本文第三章：

表 4.11 參考基頻之聲調辨認率(內部測試，8 tone)

輸入聲調	聲調辨認正確率(%)							
	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Tone 6	Tone 7	Tone 8
Tone 1	74.96	13.53	0.09	10.25	0.16	0.47	0.54	0.00
Tone 2	11.70	76.89	0.25	2.37	1.21	3.77	3.78	0.03
Tone 3	0.44	1.03	57.32	24.78	1.74	0.04	14.61	0.04
Tone 4	10.54	2.71	3.24	80.84	0.28	0.12	2.28	0.00
Tone 5	8.58	36.87	3.38	14.99	9.79	1.15	25.19	0.07
Tone 6	23.46	57.32	0.00	3.47	1.49	9.77	4.50	0.00
Tone 7	6.50	14.00	2.54	25.73	4.04	0.03	47.13	0.03
Tone 8	39.01	22.58	1.41	30.04	0.40	1.01	5.24	0.30
					平均正確率		68.82 %	

表 4.12 參考基頻之聲調辨認率(內部測試)

輸入聲調	聲調辨認正確率(%)				
	一聲	二聲	三聲	四聲	五聲
一聲	74.96	13.53	1.10	10.25	0.16
二聲	11.70	76.89	7.80	2.37	1.23
三聲	9.12	21.44	47.07	19.66	2.72
四聲	10.54	2.71	5.64	80.84	0.28
五聲	20.78	31.14	20.87	21.03	6.19
			正確率		68.82 %

這個實驗是基於假設基頻軌跡抽取出與參考基頻一致的話，那麼得到的聲調辨識率，在表 4.11 中我們發現 Tone 6 與 Tone 2 之間的混淆非常嚴重，尤其是 Tone 6 被混淆成 Tone 2，從本文第三章所呈現的聲調模型的軌跡形狀亦可發現，其實 Tone 6 與 Tone 2 的軌跡形狀非常相似，所以可以解釋為什麼 Tone 6 與 Tone

2 容易混淆的原因，如果我們將 Tone 6 歸為與 Tone 2 一樣是“二聲”的話，辨認率可以提升至 72.21%左右。

實驗二：

利用聲調模型與韻律模型輔助基頻估測並同時聲調辨認，詳細請見本文第三章敘述，聲調辨識率整理如下表：

表 4.13 模型輔助基頻偵測器之聲調辨識率(內部測試)

輸入聲調	聲調辨認正確率(%)				
	一聲	二聲	三聲	四聲	五聲
一聲	73.08	14.03	1.55	10.82	0.52
二聲	10.92	79.11	5.85	3.09	1.03
三聲	8.15	24.51	41.50	22.95	2.89
四聲	9.53	3.05	4.71	82.37	0.34
五聲	17.51	35.30	18.64	22.81	5.74
			正確率		68.51 %

表 4.14 模型輔助基頻偵測器之聲調辨識率(外部測試)

輸入聲調	聲調辨認正確率(%)				
	一聲	二聲	三聲	四聲	五聲
一聲	77.52	12.53	1.18	8.57	0.21
二聲	13.94	79.67	3.29	2.32	0.77
三聲	10.67	24.65	40.55	22.09	2.03
四聲	12.28	2.98	4.59	79.85	0.31
五聲	25.90	34.66	10.76	23.51	5.18
			正確率		67.70 %

由上面結果顯示，其實外部測試得到的結果與內部測試差不多，其原因可能是因為語料庫的外部測試與內部測試均是同一位語者的緣故。從表 4.3 我們可得知，所估測出的基頻軌跡其實與參考基頻相當接近了，所以可以推測出估測基頻的聲調辨識率非常接近實驗一的結果，同樣的，如果我們將 Tone 6 歸為“二聲”的話，辨認率可以提升至 71.5%左右。

實驗三：

使用基本型基頻偵測器所估測的基頻進行聲調辨認，整理結果如下表：

表 4.15 基本型基頻偵測器之聲調辨識率(內部測試)

輸入聲調	聲調辨認正確率(%)				
	一聲	二聲	三聲	四聲	五聲
一聲	72.16	14.93	1.63	10.56	0.72
二聲	11.09	78.57	6.30	2.75	1.29
三聲	9.11	23.35	41.91	21.64	4.00
四聲	9.78	2.69	5.66	81.08	0.79
五聲	18.44	34.82	18.96	21.39	6.39
			正確率		67.88 %

表 4.16 基本型基頻偵測器之聲調辨識率(外部測試)

輸入聲調	聲調辨認正確率(%)				
	一聲	二聲	三聲	四聲	五聲
一聲	75.37	12.63	1.39	9.21	1.39
二聲	14.62	75.22	5.71	2.81	1.65
三聲	11.95	23.59	36.71	24.87	2.88
四聲	12.34	3.22	4.96	79.05	0.43
五聲	25.10	32.27	13.94	21.51	7.17
			正確率		65.39 %

這個實驗是為了比較基本型基頻軌跡偵測器與模型輔助基頻軌跡偵測器之間的差異，雖然說最後的結果並沒有太大的改善，但是結果證實了模型輔助型基頻偵測器不僅可以使基頻軌跡更加準確，同時也可以使得聲調辨認率上升。

實驗四：

假設語料庫中每個音節的韻律狀態為已知，使用模型輔助基頻偵測器偵測基頻軌跡與聲調辨認，整理實驗結果如下表所示：

表 4.17 已知與未知韻律狀態音框間基頻值的比較表

	模型輔助基頻偵測器 (韻律狀態已知)	模型輔助基頻偵測器 (韻律狀態未知)
總音節個數	616,337	
比值範圍	音節平均基頻值比	音節平均基頻值比
0.9~1.1	97.98 %	97.9 %
0.8~1.2	98.80 %	98.7 %
0.7~1.4	99.12 %	98.9 %
0.6~1.7	99.54 %	99.1 %
0.5~2.0	99.61 %	99.1 %
GPE	0.761 %	0.918 %

表 4.18 已知韻律狀態之模型輔助基頻偵測器聲調辨識率(內部測試)

輸入聲調	聲調辨認正確率(%)				
	一聲	二聲	三聲	四聲	五聲
一聲	95.71	0.03	0.13	3.84	0.29
二聲	0.22	95.48	3.60	0.38	0.32
三聲	1.29	22.53	71.23	2.45	2.51
四聲	2.70	0.27	0.56	95.91	0.56
五聲	11.65	7.04	8.49	19.61	53.21
			正確率		89.1 %

從上表 4.17 中可以發現，若韻律狀態為已知時，基頻軌跡的偵測準確度幾乎快到達上限(Up bound)，而在聲調辨識率上更可以看出韻律訊息的重要性，聲調辨識率從七成提高至九成，如果將 tone 6 視為“二聲”的話，辨識率更是達到了 91.93%左右，但儘管如此，“五聲”的辨識率依舊不到六成。

在前三次實驗的結果我們可以發現“一聲”、“二聲”與“四聲”都有不錯的辨認率，主要是因為這幾類的聲調在基頻軌跡的表現上比較穩定而且明顯，所以較容易辨識，而剩餘的“五聲”與“三聲”辨識率較低，尤其是“五聲”，主要原因是因為這兩個聲調容易與前後音節耦合所影響，而經過訓練後的聲調模型仍舊無法 model 大部分的耦合影響，所以導致辨認率不高。

在前三次的實驗結果中，我們可以發現，有使用聲調與韻律模型輔助的基頻偵測器在基頻軌跡偵測與聲調辨認上，都可以比基本型基頻偵測器的基頻準確度與聲調辨認率少許的提升，而在實驗四中發現，已知韻律狀態的模型輔助基頻偵測器可以得到幾乎達到上限的基頻準確度，與將近九成的聲調辨認率。

第五章　結論與展望

5.1 結論

在本論文中，我們證明利用聲調模型與韻律模型輔助的方式不僅僅可以減少基頻軌跡產生半頻或倍頻錯誤的發生，同時也可以改善聲調辨認率，但不方便的是必須事先將各分數之間的權重調整至適當的比例。

關於連續語音聲調的辨認上，在前三次實驗中，五聲的辨識率仍舊不高，主要是因為只用正交展開式的四維參數並無法有效的把五聲與其他聲調鑑別出來，在統計上，五聲的音節長通常比其他聲調短，而在實驗中我們並沒有使用到音節長這個特徵來辨認聲調，再加上三聲與五聲容易因為左右音節連音的關係使得基頻軌跡表現上變化，所以單純只靠基頻軌跡與左右是否有與別的音節連接的特徵，無法有效的辨識出聲調。

最後一個實驗裡，我們假設了每個音節的韻律狀態為已知，發現基頻軌跡的估測可以說幾乎達到最佳，而在聲調辨認也達到了九成，可見若能有效的辨認出音節的韻律狀態的話，對於基頻估測與聲調辨認上都能獲得極大的幫助。

5.2 未來之展望

在本論文中，基頻軌跡建立是假設音節位置與 U/V 的邊界為已知的，其實，將來可以利用更多的特徵去辨認音節位置或是 U/V，使得這個基頻估測與聲調辨認器更佳完善，而在聲調辨認方面，如果可以再加入更多的特徵參數來幫助辨認的話，相信可以大大改善五聲與三聲的辨認率。

參考文獻

[1] F.J Charpentier ， “Pitch detection using the short-time phase spectrum，” ICASSP’86，TOKYO

[2] T. Abe， T. Kobayashi， and S. Imai， “Robust pitch estimation with harmonic enhancement in noisy environment based on instantaneous frequency，” *Proc. 4th ICSLP*， pp.1277-1280， Philadelphia， Oct. 1996.

[3] T. Tanaka， T. Kobayashi， D. Arifianto， T. Masuko， “Fundamental frequency estimation based on instantaneous frequency amplitude spectrum，” *Proc. ICASSP*， vol-I， pp.329-332， Orlando， Fl.， May 2002.

[4] D. Arifianto and T. Kobayashi， “IFAS-based voiced / unvoiced classification of speech signal，” *Proc. ICASSP*， vol.I， pp.812-815， Hong Kong， April 2003.

[5] D. Arifianto and T. Kobayashi，“Voiced/Unvoiced Determination of Speech Signal in Noisy Environment using Harmonicity Measure Based on Instantaneous Frequency，” Volume 1， March 18-23, 2005 Page(s)：877 - 880 Digital Object Identifier 10.1109/ICASSP.2005.1415254

[6] D.J. Liu and C.T. Lin， “Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure，” *IEEE Trans.， Speech and Audio Proc.*， vol. 9， no. 6， pp. 609-621， Sept. 2002.

[7] W.-y Lin and L.-s Lee， “Improve Tone Recognition for Fluent Mandarin speech Based On New Inter-Syllabic Features and Robust Pitch Extraction，”ASRU 2003

[8] Chen-Yu Chiang, Yih-Ru Wang and Sin-Horng Chen， “On the Inter-syllable Coarticulation Effect of Pitch Modeling for Mandarin Speech，” In

INTERSPEECH-2005, 3269-3272.

[9] D. Talkin, "A robust algorithm for pitch tracking (RAPT),"in *Speech coding and synthesis*, W. B. Kleijn and K. K.Paliwal, Eds.： Elsevier Science, 1995, pp. 495 –518.

[10] 陳鳳儀，蔡碧芳，陳克健，黃居仁，“中文句結構樹資料庫(Sinica Treebank)的構建”，中央研究院資訊所、中央研究院研究所。

[11] 曹登鈞，“利用統計方法之基週期偵測器與國語連續語音聲調辨認”，國立交通大學碩士論文，民國九十一年六月。

[12] L.S. Lee, C.Y. Tseng and M. Ouh-Young, “The Synthesis Rules in a Chinese Text-to-Speech System,” *IEEE Trans. Acoust., Speech, Signal Processing, Vol.37, No.9,* pp.1309-1320, Sep. 1998.