國 立 交 通 大 學

電信工程學系

碩士論文

使用 MLP 與韻律模型之聲調辨認

Tone Recognition Using MLP and Prosody Model

研 究 生：陳宏宇

指導教授：陳信宏　博士

中華民國九十六年七月

使用 MLP 與韻律模型之聲調辨認

Tone Recognition Using MLP and Prosody Model

研 究 生：陳宏宇　　Student：Hong-Yu Chen

指導教授：陳信宏　　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 Engineering and Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in Electrical Engineering

July 2007
Hsinchu, Taiwan, Republic of China

中華民國九十六年七月

使用 MLP 與韻律模型之聲調辨認

研究生：陳宏宇　　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，基本辨識系統上，對單一音節的辨認運用前後音節的特徵參數，並對於音高輪廓及能量區段化的方式，利用 MLP 辨認器進行聲調辨認，實驗於單一語者及非特定語者語料庫，辨認率分別為 87.74%及 83.27%；擴展特徵參數抽取方式至 tone pair 上，同樣利用 MLP 辨認器，加上利用 Viterbi search 對於 MLP 辨認器進行修正，辨認率分別為 88.15%及 85.81%；此外，利用音節的基頻軌跡、音節間的 pause duration 及 energy-deep level，訓練聲調模型、韻律模型、音節間的 break type 模型，利用 Viterbi search 做聲調辨認，對於單一語者語料庫，最高可得到辨識為 71.89%。

Tone Recognition Using MLP Tone Recognizer And Prosody Model

Student : Hong-Yu Chen Advisor : Dr. Sin-Horng Chen

Institute of Communication Engineering

National Chiao Tung University

Abstract

In this thesis, the features of the preceding and the succeeding syllable are used to help tone recognition on MLP (multi-layer perceptron) tone recognizer. The features include means and slopes of three uniformly divided-pitch contour, duration of the syllable and energy. Recognition rate are 87.74% and 83.27% for single speaker and multi-speaker database. If using the features of tone pair on MLP tone recognizer, the recognition rate are 88.15% and 85.81% respectively. Furthermore, using the features of pitch contour, pause duration and energy-dip level construct prosody model, tone model and break type model. Then we use Viterbi search algorithm to recognize. A recognition rate of 71.89% is achieved.

誌謝

時間過的很快，兩年的時間很快就過去了，首先感謝我的指導教授陳信宏老師與王逸如老師，感謝兩位老師兩年辛勤的指導與耐心的教誨，而可以學習到許多知識與研究的方法，在這裡表達我最誠摯的敬意。

這兩年也要感謝實驗的學長、同學及學弟，大家都給我許多的幫助，尤其感謝振宇學長及智合學長對我論文上的協助，也要謝謝輝哥、信德、希群及 Barking 學長們的幫助；感謝啟風、獻文、明彥、柏蒼、胤賢、友駿、小傅及小鄧，大家都幫了我不少忙，在研究所這段過程，才能解決許多的問題。另也要感謝一路上曾經幫助我、鼓勵我的朋友及同學，有大家的陪伴，生活才會充滿了色彩與歡笑。

最重要的要感謝我的父母及家人，尤其爸爸、媽媽能給我最大的支持及鼓勵，讓我在成長及求學路上，都能無憂慮地做自己想做的事，而未來我一定會更努力的。

目錄

表目錄

圖目錄

第一章　緒論

1.1 研究動機

在科技發展的新世代中，所有的事物不斷的在更新、改變與進步，電腦、手機、PDA 等，都是現代化的產品，在普及化的情況下，除了使用鍵盤或是手寫系統，人機介面便是重要的一步，更人性化的控制的方式，是使用者使用口語方式控制機器，這已經不再是不可能的事情，舉例來說，手機已有語音的辨識的功能，如何使語音辨識更進步，讓辨識技術更為可靠，仍是近年來研究的目標。語音辨識技術在聲調辨認上已經發展非常久的一段時間了，中文是一種聲調語言(tonal language)，音節可分解為基本音節及聲調，目前中文語音辨認系統，大都只是利用聲音的頻譜特徵參數進行基本音節辨認，對於聲調的辨認，則利用語言模型補償，因此，針對聲調辨認，可以更正確了解語意，達到完整的辨認結果。

中文聲調辨認研究上已有許多方法，例如 VQ (Vector Quantization)法、MLP (multi-layer perceptrons) 法 [1]、HMM(Hidden Markov Model)法、GMM(Gaussian Mixture Model)法等；除了辨認方法外，聲調辨認研究的另一重點在於辨認參數的選取及補償，一般使用音節內的特徵參數(MFCC、能量、基頻)及音節間特徵參數(inter-syllable features)，建立聲調變異的模型，進行辨認[2]。

1.2 研究方向

本論文之研究，主要使用 MLP 辨認器，並擴大辨認單元為 tone pair，考量參數的抽取，從基頻軌跡、能量及音節間的關係，抽取音節或雙音節前後的參數，

經由 MLP 辨認器進行聲調辨認。另考慮一新辨認方式，利用模型化的方法進行辨認，訓練時經由統計方式對語料庫中基頻軌跡的聲調變化，建立聲調模型、韻律模型及 break type 模型；辨識時，建立每一個音節的狀態候選者，從中挑選最佳的聲調序列，簡化搜尋方式辨認聲調。

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　使用 MLP 之聲調辨認器，分別以一個音節及兩個音節(tone pair)為辨認單元；另使用四維正交參數做為描述 pitch contour 的參數。

第三章　使用韻律模型來作中文聲調辨認。

第四章　實驗結果與分析：利用第二、三章的辨認方法，辨認中文連續語音的聲調，分析實驗的結果。

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　MLP 中文連續語音聲調辨認

本章節中使用 MLP 辨認器，分別對單一音節及兩個音節作聲調辨認；雙音節之 tone pair 辨認係利用 Viterbi search 來求取最佳聲調序列；此外，描述音節的音高輪廓的參數，也嘗試用四維正交參數描述。

2.1 國語聲調的特性

中文是一種聲調語言，中文音節的結構都是由 411 個基本音節與五種聲調構成，一個音節的聲調由音高輪廓(pitch contour)來描述，而音高(pitch)是指發音時聲帶震動的頻率，而震動的頻率就是聲音的基頻(fundamental frequency，F0)[3]，震動的頻率越高，音高也就越高。聲調就是指我們在發音的時候，隨著時間的變化，頻率會有不同的高低起伏變化而產生出不同聲調，因此，音節內的音高輪廓的形狀就是我們判斷聲調的重要依據，一般將聲調分為為一聲(high-level)、二聲(mid-rising)、三聲(mid-falling-rising)、四聲(high-falling)，此外還有五聲(輕聲，neutral tone)，五聲的音高輪廓軌跡通常是一不規則軌跡，五聲的音高輪廓是不固定的，大多和前後的音節影響有關，而五聲的特性通常是能量較低及音節長度較其他聲調來的短。

如果只針對單字音的發音來看，我們所發出的聲調，其音高輪廓之標準形式如圖 2.1 所示，各自具有其獨特的基頻軌跡分佈。圖中並沒有標示出第五聲的音高輪廓軌跡，這是因為五聲的音高輪廓是不固定的，大多和前後所發的音有關，通常是一不規則軌跡。

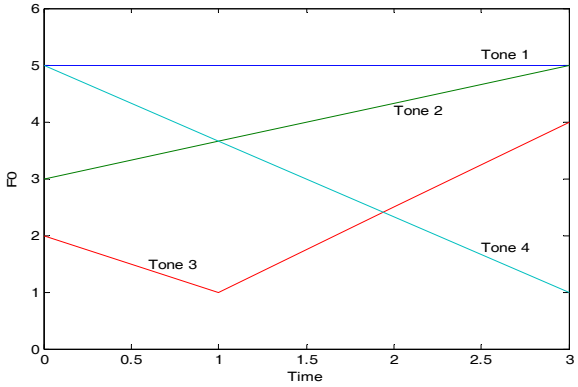

圖 2.1 音節聲調的基頻軌跡 vs. 時間關係圖

從單一的語者角度看，可以看出一聲的音高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。一般而言，每一個人的音高平均值、音高範圍都不相同，而女生所發出的頻率都比男生來的高，而連續語音中也會因語者說話的速度，造成音節長度的不同，因此圖 2.1 只是獨立音節發音的理想結果，如果在連續語音的情況下，音節就會受前後音節的影響，容易造成聲調辨識的混淆。

三聲是辨認中除了五聲外較易辨認錯誤的，連續語音的影響因素，大多是受到後面一個音節聲調的影響，遠比受到前一個音節的影響為多，而三聲是最常發生變調 (tone-sandhi) 的狀況，在三聲接三聲的情況下，前面的三聲大都會變調成二聲，如果有連續的三聲出現，常只有最後一個會是三聲，而前面的三聲都變調成二聲，此時音高輪廓就不是先降後升(falling-rising)。

此外，儘管知道連續語音音節之音高輪廓軌跡會受到前後音之聲調影響，但它也會受到其他因素的影響，如詞結構、語法、甚至語意等的影響，這是造成聲調辨識困難的主要原因。聲調辨認由基頻的抽取開始，本論文中並沒有討論基頻抽取的方式，而是使用 ESPS(Entropic Corp.) [4] 軟體求取基頻，論文重點在於如何做聲調辨認。

2.2 MLP 聲調辨認器

本章節中使用的基本辨認器為 MLP 聲調辨認器[1]，架構如圖 2.2，共分為三層；第一層為輸入層(input layer)，輸入的參數是從語音中所抽取的聲調辨認參數，第二層為隱藏層(hidden layer)，第三層為輸出層(output layer)；辨認器的訓練演算法是 back propagation algorithm。MLP 係為一正回饋 (feed forward) 網路，每一個第 m 層的神經元 (neuron) 的輸出 $O_k^{(m)}$，都是由第 m-1 層輸出的加權總和的非線性函數，數學表示為

$$O_k^{(m)} = f\left(\sum_{i=0}^{N_\Delta} w_{ki}^{(m)} \cdot O_i^{(\Delta)} + \theta_k^{(i)}\right) \tag{2.1}$$

$$f(x) = \frac{1}{1+e^{-x}} \tag{2.2}$$

其中 $f(x)$ 為一 sigmoid 函數，N_Δ 表示第 Δ 層的神經元的數目，$w_{ki}^{(m)}$ 表示由第 Δ 層的第 m 層的第 k 個神經元的加權值。

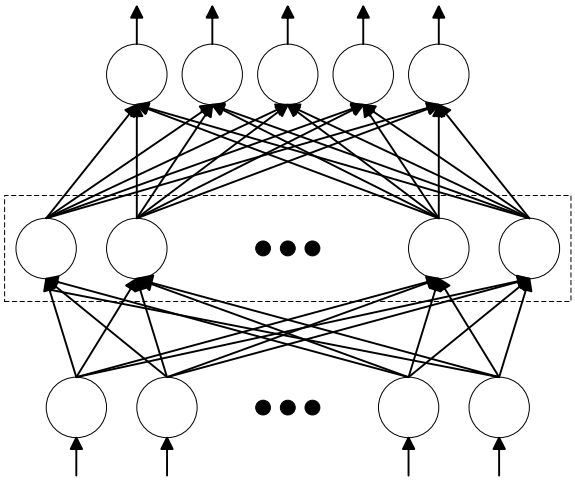

圖 2.2 MLP 聲調辨認架構圖

2.3 建立聲調模型的前處理

首先做前處理，利用隱藏式馬可夫模型對語料庫中每一 utterance 做 forced alignment，求出每一個 syllable 的起始及結束位置[5]，此外求出每個音框 (10ms) 的能量，使用 wavesurfer 求出基頻值。

針對每一個音檔所求出來非零的基頻做正規化 (normalization)，實驗的語料庫中，雖然有單一語者的語料庫，但是可能因為錄音時間的不同，造成當時音高的高低還是會有不同，而在非特定語者的語料庫，更因男女的基頻值有明顯的差異，因此做基頻正規化是必須的。首先將所有非零的基頻取對數值，計算出平均值(μ)，依照 $f_i^{'} = f_i - \mu$ 算出新的基頻值。

依據所求音節的起始結束位置，希望能在一個音節中，找到一段連續非零的基頻軌跡，因為對於之後的聲調辨認，將在連續的基頻上求取參數。會有基頻值為零的原因，是因為 syllable 有可能是由 unvoiced 聲母 (initial) 加上 voiced 韻母 (final) 所組成，unvoiced 會造成區間內的基頻值會有零發生的情況(如圖 2.3 的藍色部分)，所以在每一個音節前半部是可能會有 unvoiced 出現，造成在所求出的基頻值是零，因此依照起始位置，將開始位置往後延後至非零的位置；而 syllable 結束的位置因為原始切割可能有誤差，可能會切到後一個 syllable 的 unvoiced 部分或是沒有聲音的部份，造成出現為零的基頻值，所以結束位置也將提前至非零的位置，依照上述修正後的結果，在新的起始結束位置的區間內(如圖 2.3 紅色的部份)所有的基頻值將不為零，使得起始結束位置為此音節的 voiced 的起始結束位置，也就是此音節的基頻軌跡起始位置。此外，音節中的基頻軌跡，可能會因為求取基頻值發生錯誤，造成一個音節中出現兩段的軌跡，如果有一段的基頻軌跡是大於 40ms，而另一段的軌跡小於 40ms，則判斷為前者，如果上述條件不成

立，則由兩段能量平均值來判斷，能量較大者判定為此音節的基頻軌跡。

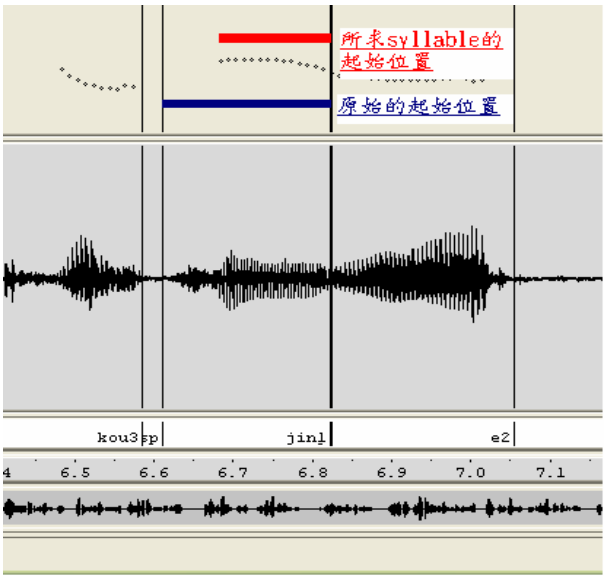

圖 2.3 基頻軌跡的起始位置示意圖

2.4 聲調特徵參數的抽取

依照 2.2 節的聲調模型的前處理，要使用 MLP 的聲調辨認器，再來就是聲調參數的抽取，依照參考資料[1]，對音節做參數的抽取。

2.4.1 基頻軌跡的區段化

將基頻軌跡的起始結束位置內的基頻軌跡平均分成三段，在每一段中分別抽取三個特徵參數，分別為此段的基頻平均值、能量的平均值及此段基頻斜率值，基頻的平均值為前置處理正規化後的基頻的平均值，而能量的平均值為 2.3 節所求能量的平均值，斜率的求法是依照每段區間內音框的基頻值求出的最小平方直線 (Least Square Line) 的斜率值。依照上述，每一個音節可以求出九個特徵參數，此外利用音節的起始結束位置，求出音節的長度 (duration)，所以每一個音

節一共有十個特徵參數。

2.4.2 使用前後音節的特徵參數

另因中文連續語音每一個音節都會受到前後音節的影響，所以加入前後的特徵參數來建立音節的聲調模型，依照 2.4.1 節的方法，先求出前後音節三段的九個參數，加入前一個音節最後一段及後一個音節第一段的三個參數(基頻平均值、能量平均值及基頻斜率值)，一共六個參數。此外，依照音節的基頻軌跡的起始結束位置，求出音節和前一個音節間 pause 的長度，同樣加入與後一個音節間 pause 的長度，所以一共再加入了兩個參數。

綜合所述，針對一個聲調模型，一共抽取了十八個特徵參數。但是因為一個音檔中的第一個和最後一個音節不能求出前一個和後一個音節所影響的參數，所以受前後影響參數均為零。因此加入兩個 binary indicator，一個來標示是否存在相鄰的前一個音節，如果存在相鄰的前一個音節就將 indicator 標示為 0，否則標示成 1；同理，另一個 indicator 以相同的方式來標示是否存在後一個音節。每一個音節一共抽出了二十個特徵參數，特徵參數抽取的圖示如下圖 2.4。

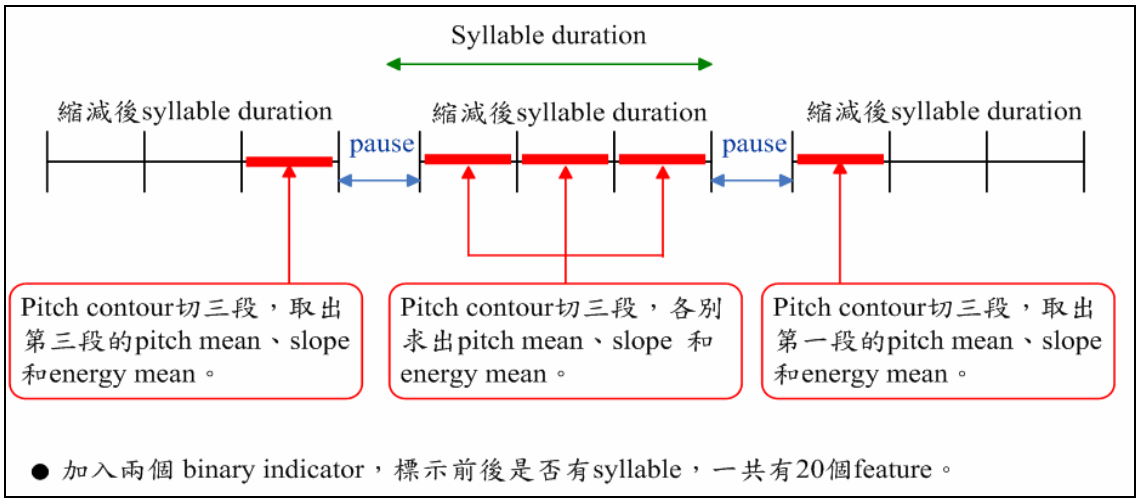

圖 2.4 前後文相關特徵參數抽取示意圖

2.5 使用 MLP 來做 Tone Pair 辨認

在 2.4 節中我們建立五個聲調的辨認，在此小節我們擴展 MLP 辨認器做 tone pair 的辨認，一共會有二十五個 tone pairs，辨認的時候則是一次取出兩個音節來辨認，利用 MLP 辨認器先計算出分數，再用 Viterbi search 做最後的辨認。

2.5.1 特徵參數的抽取

依照 2.4.1 節所述，針對一個音節分成三段，分別求出基頻平均值、能量的平均值及基頻斜率值，以及音節的長度，一共有十個特徵參數，因為是進行 tone pair 辨認，一次會抽取兩個音節的參數，故總共二十個特徵參數，另外加入兩個音節音高輪廓間的 pause 長度，以及此 tone pair 和前一個及後一個音節音高輪廓間的 pause 長度，因此會有二十三的特徵參數(如圖 2.5)。除了二十三個參數外，加入了兩個 binary indicator，來標示每個 tone pair 的前後是否有音節的存在，所以一共是二十五個特徵參數。

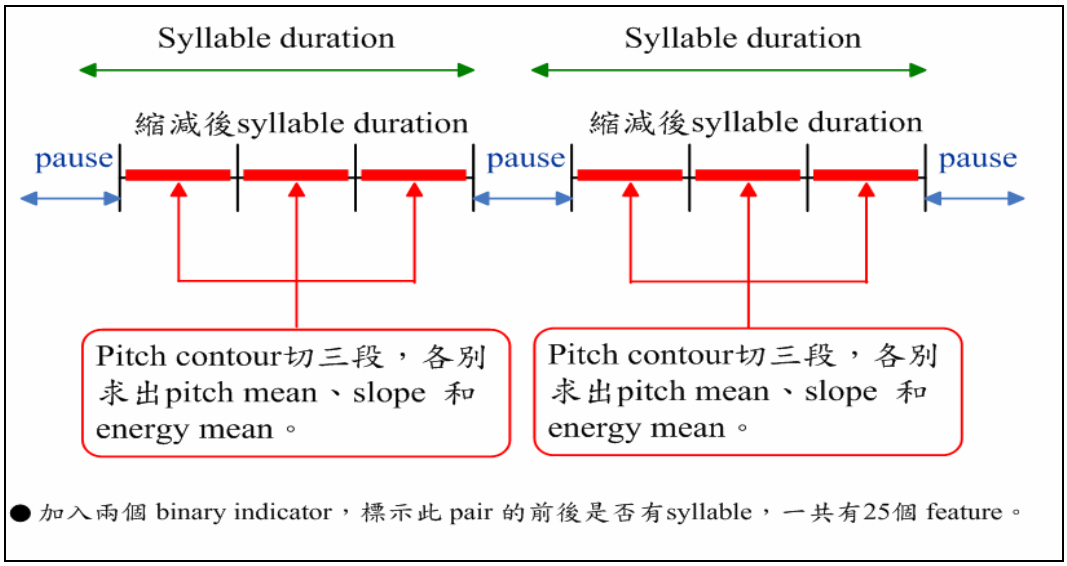

圖 2.5 Tone pair model 特徵參數抽取示意圖

Tone pair model 特徵參數的抽取是從一個句子中去抽取，因此會利用到標點符號的位置來當做一個句子的開始和結束，假設一個句子有 N 個音節，第一個音節和第二個音節抽出一個 tone pair，而第二個音節和第三個音節抽取出一個 tone pair…依此類推，最後是第(N-1)個音節和第 N 個音節建立一個 tone pair，一個 N 個音節的句子，會取出(N-1)個 tone pair，用圖示如下圖 2.6。

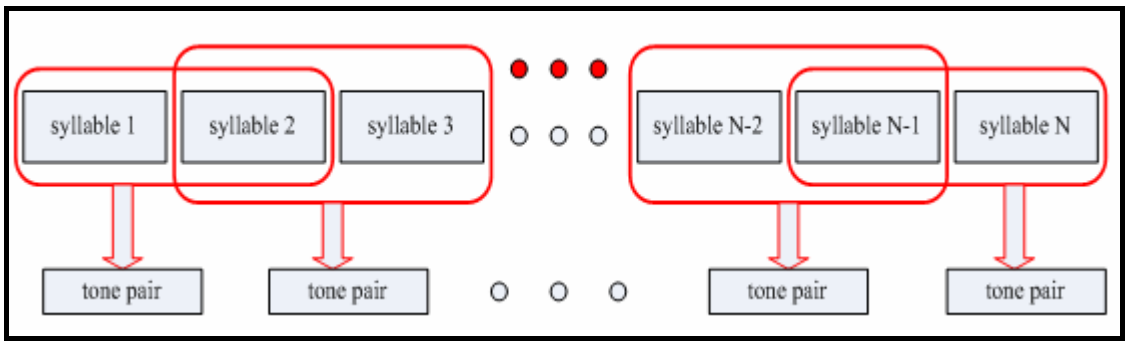

圖 2.6 Tone pair 抽取示意圖

2.5.2 Tone Pair 辨認方式

MLP 之訓練係依照圖 2.2 進行，而輸出層會有二十五個輸出，一次拿一個句子辨認，辨認器如圖 2.7，假設有 N 個音節，則會求出(N-1)個 tone pair，每一個 tone pair 都會經由 MLP 辨認器得到二十五個分數，利用 Viterbi search，希望能得到一個最佳的路徑，得到辨認結果。在尋找最佳路徑的時候，假設第 M 個 tone pair 為 p_q，則第(M+1)個只允許是 q_t，t=1~5，也就是說最佳路徑有重疊的地方只允許相同聲調。如下圖 2.8 所示。

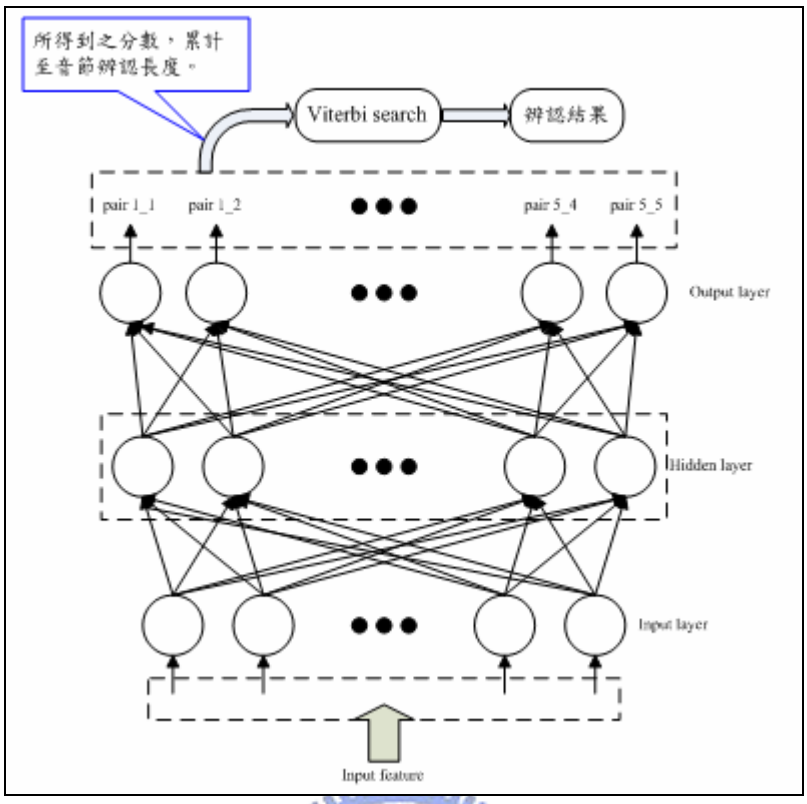

圖 2.7 利用 tone pair 於 MLP 的辨認方式示意圖

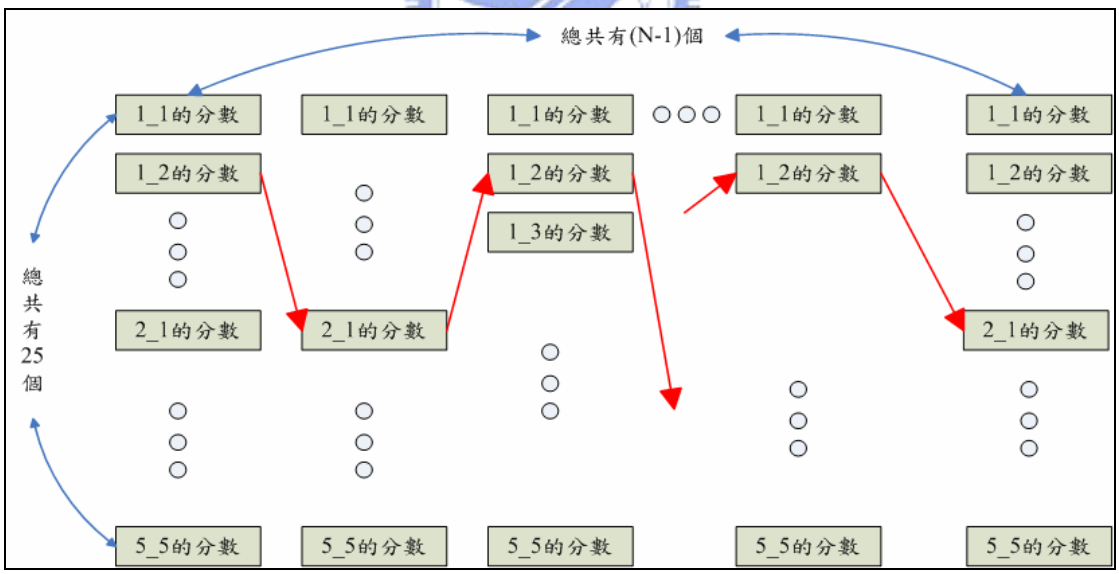

圖 2.8 利用 tone pair 的 Viterbi search 辨認示意圖

2.6 MLP 聲調辨認器使用四維正交參數

本小節中利用四維正交參數(orthogonal expansion coefficients)來描述每一個音節的音高輪廓，能量則是用區段能量的極大值，希望可以得到與 2.4 及 2.5 節有相當的辨認結果。

2.6.1 四維正交參數

對於一個音節的音高輪廓，使用四維正交參數來描述音高輪廓的形狀，表示方式如下(其中的 $f(i/N)$ 為對非零的基頻取對數值，之後對一個音檔做正規化後的值)：

$$a_j = \frac{1}{N+1}\sum_{i=0}^{N} f\left(\frac{i}{N}\right)\cdot\phi_j\left(\frac{i}{N}\right),\ \text{for}\ \ j=0,1,2,3 \tag{2.3}$$

其中，$\phi_0\left(\frac{i}{N}\right)=1$

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

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

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

重建基頻軌跡方式為

$$\hat{f}\left(\frac{i}{N}\right)=\sum_{j=0}^{3} a_j\phi_j\left(\frac{i}{N}\right) \tag{2.4}$$

2.6.2 四維正交參數的 MLP 辨認器

由上述對於每一個音節先依照 2.3 節的方式求出音高輪廓的起始位置，再來依照前一小節的公式求出一組四維正交參數，此外，加上了音高輪廓的長度(N+1)，因此描述此音節音高輪廓的參數一共使用了五個特徵參數。

只對單一的聲調建立五個聲調模型，除了使用五維的參數描述一個音節的音高輪廓外，將音高輪廓平均分成三段，分別求出三段的能量極大值，做為描述此音節的能量，及音節的長度(duration)，一共九維參數。聲調會受到前後音節的影響，故同樣加入前後描述音節的四維參數、前後音節的長度 (N+1)及前後音高輪廓能量的極大值，一共十維參數。另加入此音節音高輪廓和前後音節的音高輪廓間 pause 的長度。最後加入兩個 binary indicator 來描述此音節前後是否有音節存在，一共有二十五個參數，詳細圖示如下圖 2.9，辨認時利用 2.2 節所述的辨認器。

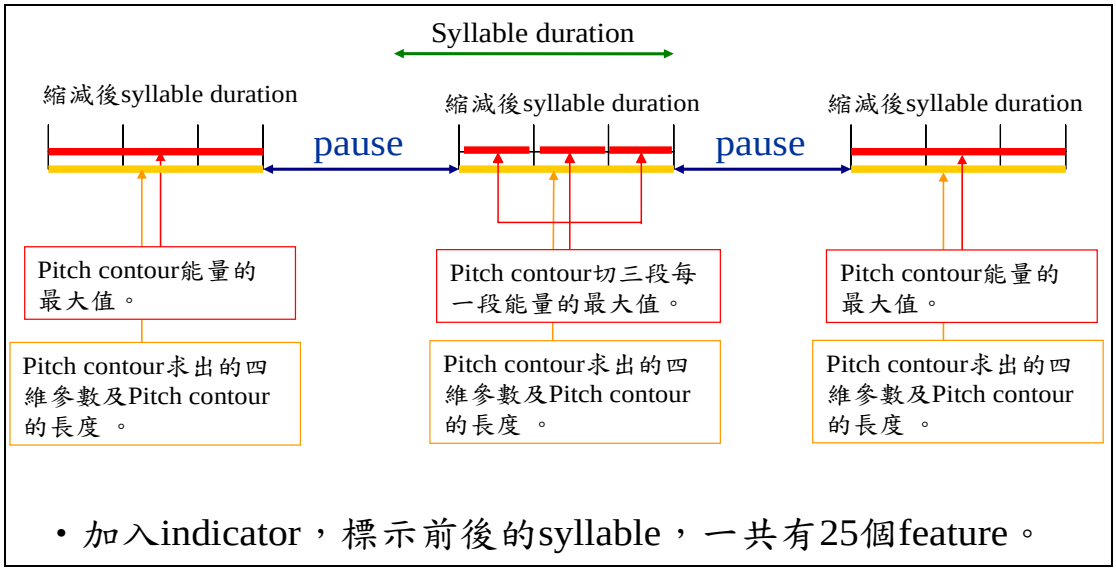

圖 2.9 MLP 辨認器使用四維正交參數的示意圖

2.6.3 Tone pair model 使用四維正交參數

建立 tone pair 辨認器時同樣的將描述音高輪廓更改為四維正交參數。每一個 tone pair 的兩個音節均是用九維參數描述(四維參數、音高輪廓的長度、三段能量分別的極大值及音節長度)，另外加入兩個音節音高輪廓間的 pause 長度，以其此 tone pair 和前一個及後一個音節音高輪廓間的 pause 長度，最後加入兩個 binary indicator 來標示 tone pair 前後是否有音節的存在，綜合上述一共有二十三個參數來描述 tone pair，圖示如下圖 2.10，辨認方式同為，2.25 節的辨認器。

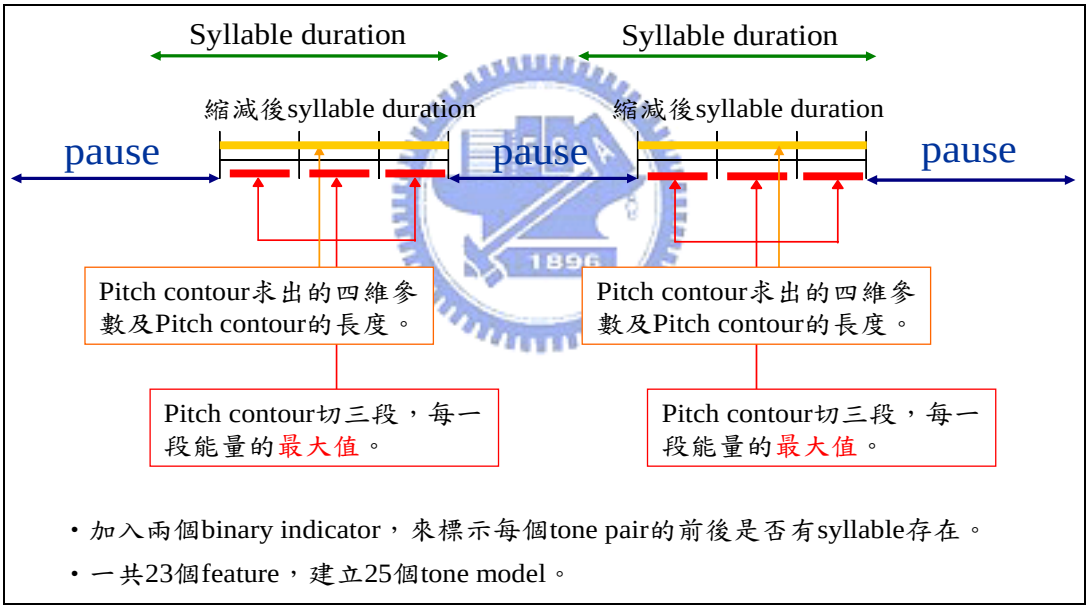

圖 2.10 Tone pair model 特徵參數抽取使用四維正交參數示意圖

第三章　運用韻律模型之聲調辨認

影響基頻軌跡的因素有很多，包括了聲調、音節的韻律狀態及語者，當然說話的速度、音節間的停頓，或是其他種種的因素，都會有影響，本章中，最重要就是運用建立新的韻律模型及音節間的 break type 模型來辨認聲調。

3.1 聲調模型及韻律模型建立

依照參考資料[6][7][8]，假設基頻軌跡的影響因素有四種，一是此音節的聲調，二是此音節的韻律狀態，三是語者的因素，四是前後音節的影響因素。前後音節的影響因素，包含了前後聲調的影響，以及此音節與前後音節間 pause 的 break type，假設上述影響因素是具有加成性的，圖 3.1 說明了音節所受的影響因素，而數學式表示如下，

$$\mathbf{sp}_{k,n} = \mathbf{sp}^{r}_{k,n} + \boldsymbol{\beta}_{t_{k,n}} + \boldsymbol{\beta}_{p_{k,n}} + \boldsymbol{\beta}^{f}_{B_{k,n-1},tp_{k,n-1}} + \boldsymbol{\beta}^{b}_{B_{k,n},tp_{k,n}} + \boldsymbol{\mu} \tag{3.1}$$

其中

$\mathbf{sp}_{k,n}$ 為第 k 音檔中的第 n 個音節的音高輪廓的四維正交參數(其中 $k = 1\sim K$，$n = 1\sim N_k$，K =語料庫中音檔的數目，N_k =第 k 個音檔中的音節數目)

$\mathbf{sp}^{r}_{k,n}$ 為 normalized(i.e., residual)後的音高輪廓

$\boldsymbol{\beta}_{t_{k,n}}$ 為聲調(tone)的影響因素，$t_{k,n} \in \{1,2,3,4,5\}$

$\boldsymbol{\beta}_{p_{k,n}}$ 為韻律狀態(prosodic state)的影響因素(本論文中將韻律狀態分成十六個狀態，因此 $p_{k,n} \in \{1\sim 16\}$)

$\boldsymbol{\mu}$ 為語者的平均值

$\boldsymbol{\beta}^{f}_{B_{k,n-1},tp_{k,n-1}}, \boldsymbol{\beta}^{b}_{B_{k,n},tp_{k,n}}$ 為此音節分別受前後音節連音的影響因素($tp_{k,n}$ 表示 tone pair($t_{k,n}, t_{k,n+1}$)，而 $B_{k,n}$ 表示兩音節($t_{k,n}, t_{k,n+1}$)間的 break type，此處的 break type 是用來定義兩音節間的 pause 的狀態，定義方式如參考資料[7]，一共有五種狀態，分別為 *B0*、*B1*、*B2*、*B3*、*B4*，而 *B2* 又分為兩類，為 *B2-1* 及 *B2-2*，圖 3.1 說明了音節所受的影響因素。)

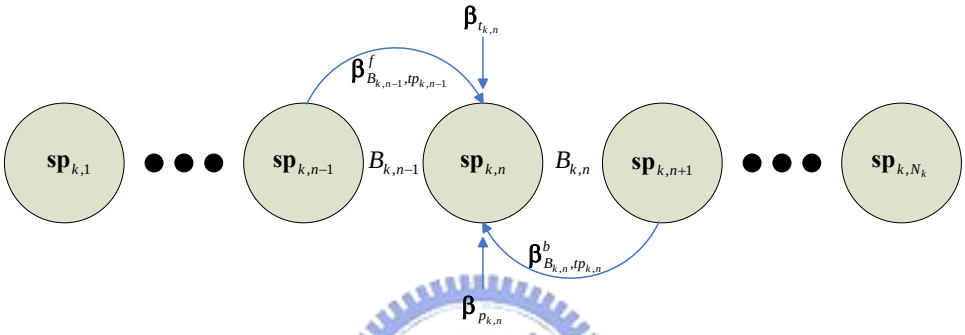

圖 3.1 音節的音高輪廓手前後影響的示意圖

依照參考資料[8]，為一對韻律狀態自動標記之方法，本論文將利用訓練後之模型進行辨認。依照定義的 likelihood function (Q，如 3.2 式)，利用 ML criterion (Maximum likelihood criterion)不斷更新模型的參數，直到 likelihood function 收斂為止，3.2 節將利用訓練出的模型，辨認聲調，訓練方式不在此詳述。

$$Q=\left\{\left\{\prod_{k=1}^{K}\prod_{n=1}^{N_k}\left[N\left(\mathbf{sp}_{k,n};\boldsymbol{\beta}_{t_{k,n}}+\boldsymbol{\beta}_{p_{k,n}}+\boldsymbol{\beta}^{f}_{B_{k,n-1},tp_{k,n-1}}+\boldsymbol{\beta}^{b}_{B_{k,n},tp_{k,n}}+\boldsymbol{\mu},\mathbf{R}\right)\times P\left(pd_{k,n},pe_{k,n}\mid B_{k,n},\mathbf{I}_{k,n}\right)\right]\right\}\right.$$

$$\left.\times\left\{\prod_{k=1}^{K}\left\{\left[P\left(p_{k,1}\right)\prod_{n=2}^{N_k}P\left(p_{k,n}\mid p_{k,n-1},B_{k,n-1}\right)\right]\left[\prod_{n=1}^{N_k}P\left(B_{k,n}\mid\mathbf{I}_{k,n}\right)\right]\right\}\right\}\right\} \qquad (3.2)$$

其中 $pd_{k,n}$ 為 pause duration、 $pe_{k,n}$ 為 energy-deep level；而 $\mathbf{I}_{k,n}$ 則是 linguistic feature。

由 3.2 式可以訓練出 $\boldsymbol{\beta}_{t_{k,n}}$、$\boldsymbol{\beta}_{p_{k,n}}$、$\boldsymbol{\beta}^{f}_{B_{k,n-1},tp_{k,n-1}}$、$\boldsymbol{\beta}^{b}_{B_{k,n},tp_{k,n}}$ 及 $\mathbf{R}$，而 $P\left(pd_{k,n}, pe_{k,n} \mid B_{k,n}\right)$ 則是分別對 pause duration 訓練 Gamma distribution model，而 energy-deep level 訓練 normal distribution model，機率模型如圖 3.2。

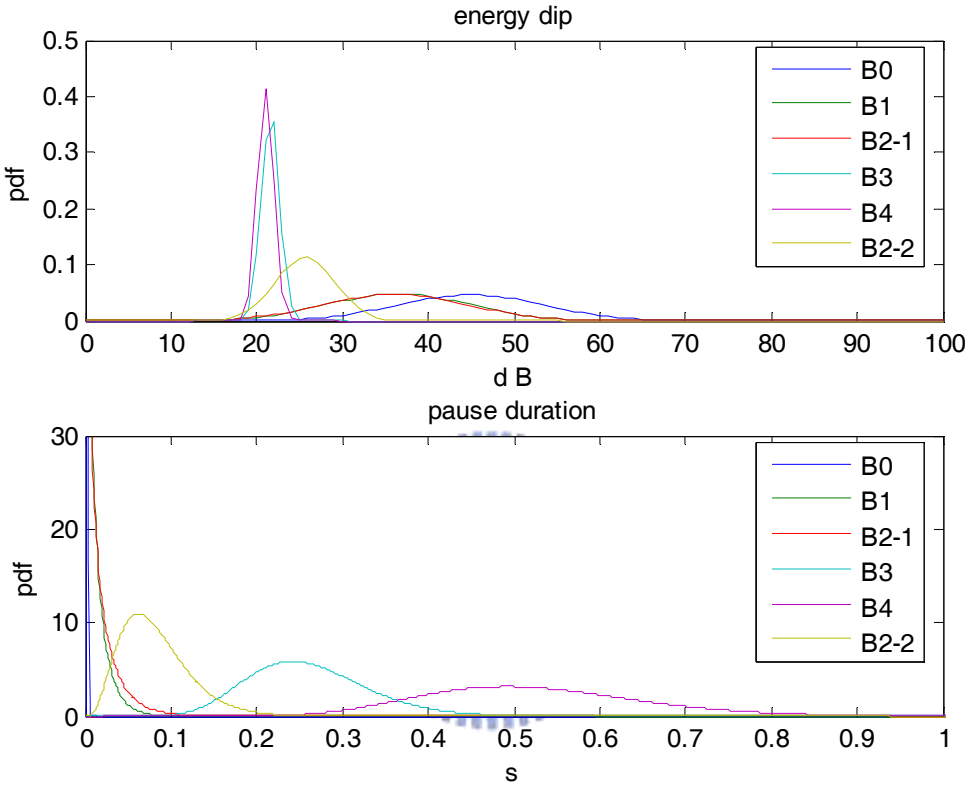

圖 3.2 pause duration 與 energy-deep level 機率模型的示意圖

而 $P\left(p_{k,n} \mid p_{k,n-1}, B_{k,n-1}\right)$ 則是統計最後訓練兩音節韻律狀態在五種 break type 下的轉移機率。$P\left(B_{k,n} \mid B_{k,n-1}\right)$ 則是統計 break type 間的 bigram 的機率。$P\left(p_{k,1}\right)$ 及 $P\left(B_{k,1}\right)$ 則是統計初始音節發生的機率。

上述的方法是已知音節的聲調的情況下訓練各種模型，下小節將利用上述模型，去猜測聲調發生的狀況。

3.2 利用模型的聲調辨認

本小節中利用 3.1 節所求出來的模型來做聲調辨認，將說明如何利用前述的模型辨認聲調的原因，在這裡是一次針對一個音檔裡的所有音節求出最佳的聲調序列，使得 $P(\mathbf{T}_k \mid \mathbf{X}_k)$ 有最大值。

$$\begin{aligned}\mathbf{T}_k^* &= \arg\max_{\mathbf{T}} P(\mathbf{T}_k \mid \mathbf{X}_k) \\ &\approx \arg\max_{\mathbf{T}} P(\mathbf{T}_k, \mathbf{X}_k) \\ &= \arg\max_{\mathbf{T}} [P(\mathbf{X}_k \mid \mathbf{T}_k) \times P(\mathbf{T}_k)] \end{aligned} \tag{3.3}$$

其中

$\mathbf{X}_k$ ：能從第 k 個音檔求出用來做為辨認的參數序列

$\mathbf{T}_k = \{ t_{k,n} \mid n = 1 \sim N_k \}$，$k = 1 \sim K$ (一次只針對一個音檔的聲調序列)

$\mathbf{T} = \{ t_{k,n} \mid n = 1 \sim N_k; k = 1 \sim K \}$

假設已知一個辨認句子中每一個音節基頻軌跡的四維正交參數($\mathbf{SP}_k$)，音節間的 pause 的長度($\mathbf{PD}_k$)及 energy-deep level($\mathbf{PE}_k$)，利用上述三個參數來做音節聲調辨認所求取的參數，$P(\mathbf{X}_k \mid \mathbf{T}_k)$ 的化簡結果如下式(3.4)，。假設 $\mathbf{X}_k = (\mathbf{SP}_k, \mathbf{PD}_k, \mathbf{PE}_k)$，則

$$\begin{aligned} P(\mathbf{X}_k \mid \mathbf{T}_k) &= \sum_{\mathbf{P}_k} \sum_{\mathbf{B}_k} P(\mathbf{X}_k, \mathbf{P}_k, \mathbf{B}_k \mid \mathbf{T}_k) \\ &= \sum_{\mathbf{P}_k} \sum_{\mathbf{B}_k} [P(\mathbf{X}_k \mid \mathbf{P}_k, \mathbf{B}_k, \mathbf{T}_k) \times P(\mathbf{P}_k, \mathbf{B}_k \mid \mathbf{T}_k)] \\ &\approx \max_{\mathbf{P}_k, \mathbf{B}_k} [P(\mathbf{X}_k \mid \mathbf{P}_k, \mathbf{B}_k, \mathbf{T}_k) \times P(\mathbf{P}_k, \mathbf{B}_k \mid \mathbf{T}_k)] \end{aligned}$$

$$= \max_{\mathbf{P}_k, \mathbf{B}_k} \left[P\left(\mathbf{SP}_k, \mathbf{PD}_k, \mathbf{PE}_k \mid \mathbf{P}_k, \mathbf{B}_k, \mathbf{T}_k \right) \times P\left(\mathbf{P}_k, \mathbf{B}_k \mid \mathbf{T}_k \right) \right] \quad (3.4)$$

其中

$\mathbf{B}_k = \left\{ B_{k,n} \mid n = 1 \sim N_k \right\}$，$k = 1 \sim K$ ； $B_{k,n} \in \{B0, B1, B2\text{-}1, B2\text{-}2, B3, B4\}$

$\mathbf{P}_k = \left\{ p_{k,n} \mid n = 1 \sim N_k \right\}$，$k = 1 \sim K$； $p_{k,n} \in \{1 \sim P\}$為 prosodic state

$\mathbf{SP}_k = \left\{ \mathbf{sp}_{k,n} \mid n = 1 \sim N_k \right\}$，$k = 1 \sim K$ 為四維正交參數

$\mathbf{PD}_k = \left\{ pd_{k,n} \mid n = 1 \sim N_k \right\}$，$k = 1 \sim K$ ；

$\mathbf{PE}_k = \left\{ pe_{k,n} \mid n = 1 \sim N_k \right\}$，$k = 1 \sim K$ ；

將(3.4)式的結果代入(3.3)式，經過化簡得到式(3.5)

$$\begin{aligned} \mathbf{T}_k^* &= \arg\max_{\mathbf{T}_k} \left\{ \max_{\mathbf{P}_k, \mathbf{B}_k} \left[P\left(\mathbf{SP}_k, \mathbf{PD}_k, \mathbf{PE}_k \mid \mathbf{P}_k, \mathbf{B}_k, \mathbf{T}_k\right) P\left(\mathbf{P}_k, \mathbf{B}_k \mid \mathbf{T}_k\right) \right] P\left(\mathbf{T}_k\right) \right\} \\ &\approx \arg\max_{\mathbf{T}_k} \left[\max_{\mathbf{P}_k, \mathbf{B}_k} P\left(\mathbf{SP}_k, \mathbf{PD}_k, \mathbf{PE}_k \mid \mathbf{P}_k, \mathbf{B}_k, \mathbf{T}_k\right) P\left(\mathbf{P}_k \mid \mathbf{B}_k\right) P(\mathbf{B}_k) P(\mathbf{T}_k) \right] \end{aligned} \quad (3.5)$$

更進一步將(3.5)式化簡並取對數值可以得到下面的(3.6)。

$$\begin{aligned} &\log\left[P\left(\mathbf{SP}_k, \mathbf{PD}_k, \mathbf{PE}_k \mid \mathbf{P}_k, \mathbf{B}_k, \mathbf{T}_k\right) P\left(\mathbf{P}_k \mid \mathbf{B}_k\right) P(\mathbf{B}_k) P(\mathbf{T}_k) \right] \\ &\approx \log\left\{ \prod_{n=1}^{N_k} \left[N\left(\mathbf{sp}_{k,n}; \boldsymbol{\beta}_{t_{k,n}} + \boldsymbol{\beta}_{p_{k,n}} + \boldsymbol{\beta}^f_{B_{k,n-1}, tp_{k,n-1}} + \boldsymbol{\beta}^b_{B_{k,n}, tp_{k,n}} + \boldsymbol{\mu}, \mathbf{R} \right) g\left(pd_{k,n}; \alpha_{B_{k,n}}, \beta_{B_{k,n}} \right) \right. \right. \\ &\left. \left. N\left(pe_{k,n}; \mu_{B_{k,n}}, \sigma_{B_{k,n}} \right) \right] \right\} + \log\left[P\left(p_{k,1}\right) \prod_{n=2}^{N_k} P\left(p_{k,n} \mid p_{k,n-1}, B_{k,n-1} \right) \right] \\ &+ \log\left[P\left(B_{k,1}\right) \prod_{n=2}^{N_k} P\left(B_{k,n} \mid B_{k,n-1} \right) \right] + \log\left[P\left(t_{k,1}\right) \prod_{n=2}^{N_k} P\left(t_{k,n} \mid t_{k,n-1} \right) \right] \end{aligned} \quad (3.6)$$

由上述的式(3.5)，利用 Viterbi search 找到最佳的 $\mathbf{T}_k$、$\mathbf{P}_k$ 及 $\mathbf{B}_k$ 序列，使得(3.5)

有最大的值。從一個句子中的音節(不包括句首和句尾的音節)來看，此音節聲調發生的可能性有 5 種，韻律狀態有 16 種可能，受到前一個音節的影響因素 $\beta^{f}_{B_{k,n-1},tp_{k,n-1}}$ 有 30 種可能(包含受到前一個可能的 5 種聲調，以及兩音節間的 6 種 break type 影響，因此有 30 種可能)；同理，受後一個音節的音響因素 $\beta^{b}_{B_{k,n},tp_{k,n}}$ 也有 30 種可能，綜合上述共有 5×16×30×30=72000 種的可能。但是每一個句子的第一個音節沒有受前一個音節的影響，最後一個音節同樣只有 2400 種的可能，因此它們有 2400 種的可能。

每一個音節的可能所代表的除了是此音節可能的狀態，其實也包含了前後音節聲調的狀態，及與前後音節間 break type 狀態，因此限制了 Viterbi search 時路徑搜尋的可能，如此限制可以將搜尋的路徑降低，但還是範圍過大，其實改變搜尋的方式，可以將搜尋量降低，下段的搜尋方式，將每一個音節的可能降低至 2400 種，較之前的 72000 種可以降低，而搜尋路徑時，只有搜尋 480 種可能的來源。

假設一個音節和後音節的關係用 state 來表示如下，

$$q_{k,n} = \left(t_{k,n}, p_{k,n}, B_{k,n}, t_{k,n+1} \right) \tag{3.7}$$

第一、二維代表了此音節的聲調及韻律狀態，第三、四維代表此音節和後一個音節間 pause 的 break type 及下一個音節的聲調，上述的表示可以將每一個音節的狀態下降至 2400 種可能；因此，前一個音節的 state 可表示成

$$q_{k,n-1} = \left(t_{k,n-1}, p_{k,n-1}, B_{k,n-1}, t_{k,n} \right) \tag{3.8}$$

在找尋最佳路徑的時候，目前音節的可能 state 就會受到前一個音節 state 的限制，而不是所有可能 state 都是可以選擇的路徑，因此，當在尋找 $q_{k,n}$ 前一個可能的音節 $q_{k,n-1}$ 時，$q_{k,n-1}$ 中的第四維，就會受到限制，必定和 $q_{k,n}$ 的第一維相同，因此只可能有 16×5×6=480 種 state，圖示如下圖 3.3。

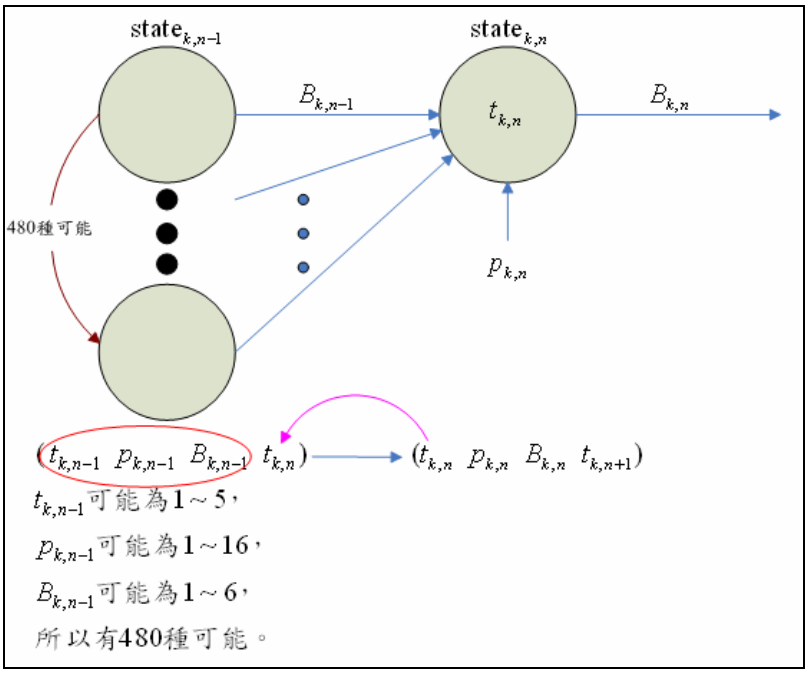

圖 3.3 音節的 state 所允許路徑的示意圖

上述的說明只是概念上的說明，實際上的搜尋方式，在下一節中將會有詳細的數學表示方式。

加入的韻律狀態轉移機率($P(p_{k,n} \mid p_{k,n-1}, B_{k,n-1})$)，來強化前後韻律狀態在 break type 情況下的轉移機率。加入 pause duration 及 energy-deep level 在 break type 下發生的機率($g\left(pd_{k,n};\alpha_{B_{k,n}},\beta_{B_{k,n}}\right)$ and $N\left(pe_{k,n};\mu_{B_{k,n}},\sigma_{B_{k,n}}\right)$)及 break type 間的轉移機率 $P\left(B_{k,n} \mid B_{k,n-1}\right)$。

除此之外，搜尋最佳路徑時，加入音節 tone 之間的轉移機率，強調前後 tone 的相關性，每個模型是等價的關係，但是實際上，每一個聲調發生的機率不相同，尤其是五聲發生的機率大約只有 5%發生的機率，如果沒有加入聲調的轉移機率，可能會造成五聲發生的機率大為提高，另可強化聲調發生的相關性，假設相鄰音節聲調為 $t_{k,n-1}$ 與 $t_{k,n}$ ，則加入 $p\left(t_{k,n} \mid t_{k,n-1} \right)$。

3.3 聲調辨認的Viterbi Search Algorithm

找尋最佳的路徑的時候，就是在每一個 state 找出 3.2 節圖 3.3 所限制 480 種可能分數的最大值，更明確的計算方式，下列數學式表示。當 $B \in \{B0, B1, B2\text{-}1, B2\text{-}2, B3, B4\}$、$t \in \{1,2,3,4,5\}$、$p \in \{1 \sim 16\}$，每一個音節的 state 有 2400 種可能，則

$$q_{k,n} = \left(t_{k,n}, p_{k,n}, B_{k,n}, t_{k,n+1} \right) \quad \text{for } n = 1 \sim N_k - 1 \tag{3.9}$$

但當 $n = N_k$ 時，假設 B_{k,N_k} 固定為 $B4$ ，且因為 t_{k,N_k+1} 不存在，每一個音節的 state 只有 80 種可能，演算法如下所表示。

Viterbi Search Algorithm

(1) **Initialization** (for $n = 1$)

$$q_{k,1} = \left(t_{k,1}, p_{k,1}, B_{k,1}, t_{k,2} \right)$$

$$L\left(q_{k,1}\right) = \log\left[N\left(\mathbf{sp}_{k,1}; \boldsymbol{\beta}_{t_{k,1}} + \boldsymbol{\beta}_{p_{k,1}} + \boldsymbol{\beta}^{b}_{B_{k,1}, tp_{k,1}} + \boldsymbol{\mu}, \mathbf{R} \right) \right] + \log\left[g\left(pd_{k,1}; \alpha_{B_{k,1}}, \beta_{B_{k,1}} \right) \right]$$

$$+ \log\left[N\left(pe_{k,1}; \mu_{B_{k,1}}, \sigma_{B_{k,1}} \right) \right] + \log\left[P\left(B_{k,1} \right) \right] + \log\left[P\left(p_{k,1} \right) \right] + \log\left[P\left(t_{k,1} \right) \right]$$

(3.10)

因 $t_{k,0}$ 不存在，所以沒有 forward coarticulation affecting pattern$\left(\boldsymbol{\beta}^{b}_{B_{k,0},tp_{k,0}}\right)$。

此處的 $L(q_{k,1})$，可以計算出 2400 個分數。

(2) **Recursion** (for $n = 2 \sim N_k - 1$)

$$L\left(q_{k,n}\right) = \max_{q_{k,n-1}\in Q(q_{k,n})}\left\{L\left(q_{k,n-1}\right) + \log\left[N\left(\mathbf{sp}_{k,n};\boldsymbol{\beta}_{t_{k,n}} + \boldsymbol{\beta}_{p_{k,n}} + \boldsymbol{\beta}^{f}_{B_{k,n-1},tp_{k,n-1}} + \boldsymbol{\beta}^{b}_{B_{k,n},tp_{k,n}} + \boldsymbol{\mu}, \mathbf{R}\right)\right]\right.$$
$$\left. + \log\left[P\left(B_{k,n} \mid B_{k,n-1}\right)\right] + \log\left[P\left(p_{k,n} \mid p_{k,n-1}, B_{k,n-1}\right)\right] + \log\left[p\left(t_{k,n} \mid t_{k,n-1}\right)\right]\right\}$$
$$+ \log\left[g\left(pd_{k,n};\alpha_{B_{k,n}},\beta_{B_{k,n}}\right)\right] + \log\left[N\left(pe_{k,n};\mu_{B_{k,n}},\sigma_{B_{k,n}}\right)\right] \qquad (3.11)$$

where $Q\left(q_{k,n}\right) = \left\{\, q_{k,n-1} \mid \left(\, t_{k,n-1}, p_{k,n-1}, B_{k,n-1}, t'_{k,n}\right), t'_{k,n} = t_{k,n} \right\}$，搜尋前一個音節可能的個數為 5×16×6=480 個；此處的 $L(q_{k,n})$，可以計算出 2400 個分數。另外為了找出最佳路徑的 state indices，定義

$$\delta\left(q_{k,n}\right) = \arg\max_{q_{k,n-1}\in Q(q_{k,n})}\left\{L\left(q_{k,n-1}\right) + \log\left[N\left(\mathbf{sp}_{k,n};\boldsymbol{\beta}_{t_{k,n}} + \boldsymbol{\beta}_{p_{k,n}} + \boldsymbol{\beta}^{f}_{B_{k,n-1},tp_{k,n-1}} + \boldsymbol{\beta}^{b}_{B_{k,n},tp_{k,n}} + \boldsymbol{\mu}, \mathbf{R}\right)\right]\right.$$
$$\left. + \log\left[P\left(B_{k,n} \mid B_{k,n-1}\right)\right] + \log\left[P\left(p_{k,n} \mid p_{k,n-1}, B_{k,n-1}\right)\right] + \log\left[p\left(t_{k,n} \mid t_{k,n-1}\right)\right]\right\}$$

(3.12)

來記錄 state $q_{k,n}$ 的前一個分數最高 state $q_{k,n-1}$ 的狀態。

(3) **Termination** (for $n = N_k$)

$$L\left(q_{k,N_k}\right) = \max_{q_{k,N_k-1}\in Q(q_{k,N_k})}\left\{L\left(q_{k,N_k-1}\right) + \log\left[N\left(\mathbf{sp}_{k,N_k};\boldsymbol{\beta}_{t_{k,N_k}} + \boldsymbol{\beta}_{p_{k,N_k}} + \boldsymbol{\beta}^{f}_{B_{k,N_k-1},tp_{k,N_k-1}} + \boldsymbol{\mu}, \mathbf{R}\right)\right]\right.$$
$$\left. + \log\left[P\left(B_{k,N_k} \mid B_{k,N_k-1}\right)\right] + \log\left[P\left(p_{k,N_k} \mid p_{k,N_k-1}, B_{k,N_k-1}\right)\right] + \log\left[p\left(t_{k,N_k} \mid t_{k,N_k-1}\right)\right]\right\}$$

(3.13)

where $Q(q_{k,N_k})=\{q_{k,N_k-1}|(t_{k,N_k-1},p_{k,N_k-1},B_{k,N_k-1},t'_{k,N_k}),t'_{k,N_k}=t_{k,N_k}\}$，因 t_{k,N_k+1} 不存在，所以沒有 backward coarticulation affecting pattern $\left(\boldsymbol{\beta}^{f}_{B_{k,N_k},tp_{k,N_k}}\right)$，另此處 B_{k,N_k} 固定為 $B4$，因此 $L(q_{k,N_k})$，可以計算出 80 個分數。

$$\delta(q_{k,N_k})=\arg\max_{q_{k,N_k-1}\in Q(q_{k,N_k})}\left\{L(q_{k,N_k-1})+\log\left[N(\mathbf{sp}_{k,N_k};\boldsymbol{\beta}_{t_{k,N_k}}+\boldsymbol{\beta}_{p_{k,N_k}}+\boldsymbol{\beta}^{f}_{B_{k,N_k-1},tp_{k,N_k-1}}+\boldsymbol{\mu},\mathbf{R})\right]\right.$$
$$\left.+\log\left[P(B_{k,N_k}|B_{k,N_k-1})\right]+\log\left[P(p_{k,N_k}|p_{k,N_k-1},B_{k,N_k-1})\right]+\log\left[p(t_{k,N_k}|t_{k,N_k-1})\right]\right\} \quad (3.14)$$

(4)**Path Backtracking**

$$\mathrm{R}_{k,N_k}=(t^{*}_{k,N_k},p^{*}_{k,N_k},B^{*}_{k,N_k})=\arg\max_{q}L(q_{k,N_k}) \quad (3.15)$$

辨識的結果 $\mathbf{RT}_{k,N_k}=t^{*}_{k,N_k}$；依照下列式子計算 for $n=N_k$ 至 $n=2$，就可以計算出答案。

$$\mathrm{R}_{k,n-1}=\delta(\mathrm{R}_{k,n})，\quad \mathbf{RT}_{k,n-1}=t^{*}_{k,n-1} \quad (3.16)$$

3.4 修正聲調模型的辨認

3.3 節所述的辨認模型 $N(\mathbf{sp}_{k,n};\boldsymbol{\beta}_{t_{k,n}}+\boldsymbol{\beta}_{p_{k,n}}+\boldsymbol{\beta}^{f}_{B_{k,n-1},tp_{k,n-1}}+\boldsymbol{\beta}^{b}_{B_{k,n},tp_{k,n}}+\boldsymbol{\mu},\mathbf{R})$ 中，目的是為了找出聲調、韻律、前後音節影響因素組合加成後，最相近組合的分數，之後利用 Viterbi search 做辨認搜尋，也就是希望將各種因素扣除後，能與零相近，共用了共變異數矩陣(covariance matrix，R)，本小節中，利用訓練後的結果，將基頻軌跡的四維正交參數扣除聲調模型以外的因素，再將每一種聲調的四維參數各自建立 mixture Gaussian model，希望可以對聲調有更好的模型，下段將更明確的說明。

訓練的步驟結束後，將會對訓練語料標示每一個音節的韻律狀態及音節間的 break type，還有音節間在五種 break type 下對基頻軌跡的影響因素 $(\boldsymbol{\beta}^{f}_{B_{k,n-1},tp_{k,n-1}},\boldsymbol{\beta}^{b}_{B_{k,n},tp_{k,n}})$，分別依照音節所標示的結果扣除韻律的影響因素、受前後音節影響因素及語者平均，

$$\mathbf{sp}'_{k,n}=\mathbf{sp}_{k,n}-\boldsymbol{\beta}_{p_{k,n}}-\boldsymbol{\beta}^{f}_{B_{k,n-1},tp_{k,n-1}}-\boldsymbol{\beta}^{b}_{B_{k,n},tp_{k,n}}-\boldsymbol{\mu} \tag{3.17}$$

分別依照聲調對 $\mathbf{sp}'_{k,n}$ 訓練 mixture Gaussian model。

3.3 節所述的 Viterbi Search 的辨識方式，Gaussian model 本身為一維的高斯函數（$N\left(\mathbf{sp}_{k,n};\boldsymbol{\beta}_{t_{k,n}}+\boldsymbol{\beta}_{p_{k,n}}+\boldsymbol{\beta}^{f}_{B_{k,n-1},tp_{k,n-1}}+\boldsymbol{\beta}^{b}_{B_{k,n},tp_{k,n}}+\boldsymbol{\mu},\mathbf{R}\right)$），將修正成：

$$\sum_{q=1}^{M}c_q N_q\left(\mathbf{sp}_{k,n};\boldsymbol{\beta}_{t_{k,n},q}+\boldsymbol{\beta}_{p_{k,n}}+\boldsymbol{\beta}^{f}_{B_{k,n-1},tp_{k,n-1}}+\boldsymbol{\beta}^{b}_{B_{k,n},tp_{k,n}}+\boldsymbol{\mu},\mathbf{R}_{t_{k,n},q}\right) \tag{3.18}$$

$$\sum_{q=1}^{M}c_q=1$$

其中 M 為 mixture number，c_q 為每個 mixture 的權重，$\boldsymbol{\beta}_{t_{k,n},q}$ 為聲調 $t_{k,n}$ 的第 q 個 mixture 的影響因素，$\mathbf{R}_{t_{k,n},q}$ 為聲調 $t_{k,n}$ 的第 q 個 mixture 的共變異數矩陣。

重新更新模型的參數之後，而最後的 Viterbi search 演算的方式與 3.3 節所述的方式相同。

第四章　實驗結果與分析

本章將說明實驗所使用的語料庫，及第二、三章所述的實驗結果，分析聲調辨識的結果。

4.1 使用語料

本小節將介紹實驗所使用的兩套語料庫，以及對語料庫所做的修正。

4.1.1 Treebank 語料庫

Treebank 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)[9]，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。一共為 379 個音檔的乾淨語料，共 52,192 個音節，其中使用 341 個音檔作為訓練語料，共有 47,093 音節，外部測試選用了剩餘 38 個音檔，共 5,099 音節。

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。此外，對於三聲接三聲的變調情況，已將變調情況做修正的工作。

實驗是對中文聲調去做辨認，所以針對訓練及測試語料做五聲調的統計，統計資料如下表 4.1。

表 4.1 Treebank 語料庫訓練語料與測試語料的聲調統計

訓練語料		
	字數	%
tone1	8801	18.56
tone2	12647	26.67
tone3	7417	15.64
tone4	16081	33.92
tone5	2471	5.21
總數	47417	100

測試語料		
	字數	%
tone1	942	19.73
tone2	1175	24.61
tone3	798	16.71
tone4	1611	33.74
tone5	249	5.21
總數	4775	100

4.1.2 TCC300 語料庫

實驗也使用 TCC300 的語料庫，為一多語者的語料庫，本論文實驗的訓練語料和測試語料一共使用二七四位的語者，錄製的語料均有長短句，測試語料中分別從中各取出五位男性及女性語者的語料，所以一共有三十位語者的語料來當做測試語料，統計資料如下表 4.2。

表 4.2 TCC300 語料庫訓練語料與測試語料的性別及人數統計

	訓練語料	測試語料
語者數	244 (男性 122 人、女性 122 人)	30 (男性 15 人、女性 15 人)
音節數	269011	31323

語料中常因為破音字的問題而有聲調標錯的問題，依據實驗室的破音字常用字表(附錄一)中的一百二十六字以及「一」和「不」去撿查是否有聲調標示錯誤的問題。「不」是依照規則去更正，「不」後面所接的聲調如果是四聲，則「不」是念二聲，其餘情況「不」是念四聲。其餘的檢查方式，為將斷詞後的結果，將出現破音字位置的前後詞分別取出，之後用人工效正，對於「一」和「不」一共修正 1835 個音節錯誤，而其餘破音字則修正 1192 個音節。

實驗是對中文聲調去做辨認，所以針對訓練及測試語料做五聲調的統計，統計資料如下表 4.3。

表 4.3 TCC300 語料庫訓練語料與測試語料的聲調統計

訓練語料		
	字數	%
tone1	54267	20.17
tone2	69237	25.74
tone3	40986	15.24
tone4	92852	34.51
tone5	11669	4.34
總數	269011	100

測試語料		
	字數	%
tone1	6059	19.34
tone2	8040	25.67
tone3	4818	15.38
tone4	11057	35.30
tone5	1349	4.31
總數	31323	100

4.2 MLP 辨認器之實驗結果

變調在中文連續語音聲調辨認是常發生的事情，針對連續三聲的變調情況，實驗中 Treebank 語料庫已有修正的答案，而 TCC300 並沒有做修正，因此，只利用規則的方式進行修正，規則是判斷連續三聲出現的情況下，音節間的 pause 是否相近(此實驗是依照 30m 為基準)，如果判斷相近則是更改為三聲。

依照第二章所述的辨認方法，本節將對 4.1 節所介紹的語料庫進行實驗及分析，分別利用 MLP 辨認器對單一音節辨認，同樣對 tone pair 進行辨認。

4.2.1 對單一音節辨認之實驗結果

對於一個音節抽取 2.4 節所述的 20 個特徵參數，利用 MLP 聲調辨認器，使用軟體為 LNKnet，實驗對象為 Treebank 語料庫，訓練時設定的 Hidden Layer 為 150 個，而 step size 設定為 0.01，而訓練的次數為 500 次。對測試語料的整體辨

識率為 87.74%，訓練語料的辨識率為 91.24%。而從單一聲調來看，四聲的辨識率最高為 92.29%，五聲的辨識率最差只有 54.43%，內部測試的實驗詳細結果如下表 4.4。

表 4.4 Treebank 使用 MLP 辨認結果

輸入聲調	聲調辨認 Confusion Matrix				
	一聲	二聲	三聲	四聲	五聲
一聲	89.3%	5.4%	0.7%	4.0%	0.6%
二聲	2.5%	91.2%	2.5%	1.7%	2.1%
三聲	0.9%	9.1%	80.7%	7.3%	2.0%
四聲	2.8%	1.2%	2.6%	92.3%	1.1%
五聲	3.4%	14.8%	16.0%	11.4%	54.4%
				正確率	87.74%

針對 TCC300 語料庫進行相同的實驗，TCC300 為一多語者語料庫，測試語料中的語者並無出現在訓練語料中，實驗設定與前實驗相同。對測試語料的整體辨識率為 83.27%，訓練語料的辨識率為 84.45%。而從單一聲調來看，四聲的辨識率最高為 90.2%，五聲的辨識率最差只有 50.9%，內部測試的實驗詳細結果如下表 4.5，實驗測試語料中每一位語者均有多個音檔進行測試，是針對一個音檔中非零的基頻進行正規化，如果是針對語者進行正規化，則外部測試的辨識率為 83.15%，與本實驗結果 83.27%是差不多的。

表 4.5 TCC300 使用 MLP 辨認結果

輸入聲調	聲調辨認 Confusion Matrix				
	一聲	二聲	三聲	四聲	五聲
一聲	85.3%	7.3%	0.9%	6.3%	0.2%
二聲	5.2%	82.9%	8.4%	2.4%	1.1%
三聲	0.9%	16.3%	74.3%	5.6%	2.9%
四聲	4.2%	2.3%	2.6%	90.2%	0.7%
五聲	2.0%	17.2%	17.7%	12.2%	50.9%
				正確率	83.27%

同樣的對於一個音節抽取 2.6.2 節所述的四維正交參數的 MLP 辨認器，將抽取的參數更換成 2.6.2 節所述的 25 個特徵參數，分別對兩語料庫進行辨認。對測試語料的整體辨識率分別為 87.98% 及 84.11%，訓練語料的辨識率分別為 93.00% 及 85.02%。內部測試的實驗詳細結果如下表 4.6 及表 4.7。

表 4. 6 Treebank 使用四維正交參數的 MLP 辨認結果

輸入聲調	聲調辨認 Confusion Matrix				
	一聲	二聲	三聲	四聲	五聲
一聲	89.1%	5.0%	0.8%	4.6%	0.5%
二聲	3.5%	90.9%	2.8%	1.5%	1.3%
三聲	1.1%	8.6%	80.7%	7.5%	2.1%
四聲	3.4%	0.9%	2.1%	92.4%	1.2%
五聲	4.2%	13.1%	13.9%	8.0%	60.8%
				正確率	87.98%

表 4.7 TCC300 使用四維正交參數的 MLP 辨認結果

輸入聲調	聲調辨認 Confusion Matrix				
	一聲	二聲	三聲	四聲	五聲
一聲	85.9%	5.3%	0.7%	8.0%	0.1%
二聲	5.3%	82.9%	8.2%	2.6%	1.0%
三聲	1.0%	14.4%	77.5%	4.6%	2.5%
四聲	5.0%	1.8%	2.0%	90.6	0.6%
五聲	1.7%	14.8%	18.1%	11.3%	54.1%
				正確率	84.11%

綜合上述四個實驗結果，Treebank 為單一語者語料，辨識率較 TCC300 多語者語料辨識率為高，另一原因 Treebank 語料庫為專業錄音員，發音較為清晰正確；另一方面，使用四維正交參數描述音高輪廓的辨識率與將音高輪廓區段化的結果比較，能得到相近的辨識率。

4.2.2 對 tone pair 辨認之實驗結果

依照 2.5 節建立 tone pair model 的 MLP 聲調辨認器，對於一個 tone pair 抽取 25 個特徵參數，同樣對於兩語料辨認，MLP 訓練時設定的 hidden layer 為 250 個，辨認結果如下表 4.8 及表 4.9，Treebank 測試語料辨識率為 85.15%，訓練語料為 92.12%，TCC300 測試語料辨識率為 85.81%，訓練語料為 87.00%

表 4.8 Treebank 使用 tone pair model 的 MLP 辨認結果

輸入聲調	聲調辨認 Confusion Matrix				
	一聲	二聲	三聲	四聲	五聲
一聲	88.0%	5.6%	0.7%	5.1%	0.6%
二聲	2.6%	91.3%	2.9%	1.6%	1.6%
三聲	1.4%	8.6%	81.7%	6.2%	2.1%
四聲	2.8%	1.2%	2.4%	92.4%	1.2%
五聲	3.4%	13.1%	11.0%	9.3%	63.2%
				正確率	88.15%

表 4.9 TCC300 使用 tone pair model 的 MLP 辨認結果

輸入聲調	聲調辨認 Confusion Matrix				
	一聲	二聲	三聲	四聲	五聲
一聲	87.8%	4.6%	0.6%	6.9%	0.1%
二聲	3.6%	86.4%	6.9%	2.1%	1.0%
三聲	0.7%	15.6%	77.8%	3.4%	2.5%
四聲	4.3%	1.6%	1.9%	91.6%	0.6%
五聲	1.0%	18.5%	17.8%	8.5%	54.2%
				正確率	85.81%

將抽取 tone pair 的參數更換成 2.6.3 節所述的 23 個特徵參數，利用四維正交參數描述音高輪廓，Treebank 與 TCC300 語料庫實驗結果如下表 4.10 及表 4.11，

整體的辨識率，Treebank 測試語料辨識率為 89.10%，訓練語料辨識率為 92.56%，TCC300 測試語料辨識率為 85.04%，訓練語料辨識率為 85.91%。

表 4.10 Treebank 使用四維正交參數的 tone pair model 之 MLP 辨認結果

輸入聲調	聲調辨認 Confusion Matrix				
	一聲	二聲	三聲	四聲	五聲
一聲	89.4%	4.1%	1.1%	4.7%	0.7%
二聲	3.0%	91.1%	2.7%	1.5%	1.7%
三聲	0.7%	7.6%	85.3%	5.1%	1.3%
四聲	3.3%	1.1%	2.3%	92.3%	1.0%
五聲	2.5%	11.8%	12.7%	6.3%	66.7%
				正確率	89.10%

表 4.11 TCC300 使用四維正交參數的 tone pair model 之 MLP 辨認結果

輸入聲調	聲調辨認 Confusion Matrix				
	一聲	二聲	三聲	四聲	五聲
一聲	86.3%	4.2%	0.7%	8.7%	0.1%
二聲	4.1%	84.8%	7.5%	2.6%	1.0%
三聲	0.7%	14.5%	78.8%	3.6%	2.4%
四聲	4.8%	1.6%	2.1%	90.9%	0.6%
五聲	0.7%	16.1%	19.2%	8.8%	55.2%
				正確率	85.04%

本小節中，使用兩種不同參數的結果顯示，在 Treebank 語料庫使用四維正交參數的 tone pair 辨識結果較高的，而 TCC300 語料庫則是相反，沒有太大的差異。而與 4.2.1 節的結果比較來看，使用 tone pair 的辨識率均較使用單一音節模型的結果為佳。

4.3 利用韻律模型之實驗結果

本小節中將討論 3.3 節及 3.4 節的辨認結果。3.3 節測試結果如下表 4.12，整體的辨識率為 68.90%，其中五聲的辨識率不佳，只有 12.5%。我們將 3.4 節中對模型修正，對於每一個聲調模型升高至 4 及 8 mixture Gaussian model，結果如下表 4.13，整體辨識率為 71.10% 及 71.89%，分別提高 2.2% 及 2.99%。

表 4.12 Treebank 利用韻律模型之辨認結果

輸入聲調	聲調辨認 Confusion Matrix				
	一聲	二聲	三聲	四聲	五聲
一聲	75.5%	11.4%	3.2%	8.3%	1.6%
二聲	8.9%	77.2%	8.8%	2.6%	2.5%
三聲	3.4%	22.7%	56.7%	13.7%	3.5%
四聲	10.8%	4.4%	7.6%	75.5%	1.7%
五聲	17.3%	33.1%	15.7%	21.4%	12.5%
				正確率	68.90%

表 4.13 Treebank 利用韻律模型之辨認結果(4 mixture)

輸入聲調	聲調辨認 Confusion Matrix				
	一聲	二聲	三聲	四聲	五聲
一聲	84.8%	5.6%	2.0%	5.6%	2.0%
二聲	12.6%	71.5%	9.3%	3.2%	3.4%
三聲	5.3%	19.3%	56.9%	14.0%	4.5%
四聲	8.8%	2.6%	7.0%	79.2%	2.4%
五聲	26.6%	21.8%	14.9%	17.7%	19.0%
				正確率	71.10%

表 4.14 Treebank 利用韻律模型之辨認結果(8 mixture)

輸入聲調	聲調辨認 Confusion Matrix				
	一聲	二聲	三聲	四聲	五聲
一聲	84.6%	6.4%	2.0%	5.1%	1.9%
二聲	11.2%	74.4%	9.1%	2.6%	2.7%
三聲	4.6%	19.6%	58.0%	13.0%	4.8%
四聲	7.9%	3.6%	6.9%	78.9%	2.7%
五聲	24.5%	23.0%	13.3%	19.0%	20.2%
				正確率	71.89%

訓練及辨識中，變調的情況並沒有在文字上修正，分析表 4.12 測試結果中，三聲是有 135 個音節是有變調情況，比對表 4.13 的結果，因為變調的原因(答案仍為三聲)，造成有 80 個音節辨認成二聲，辨認成三聲的有 29 個音節，而辨識成一聲的有 20 個音節。如果把三聲中有變調的情況更正，則三聲的辨識率上升為 62.6%，二聲則下降為 72.6%，整體辨識結果 69.97%。而同樣的，表 4.13、表 4.14 的實驗結果辨識率變為 72.17%、72.96%。

假設放寬條件已知 break type 的情況，依照 3.3 節方法，辨認識率為 74.83% (表 4.15)；而已知韻律狀態的情況下，辨認率為 96.21% (表 4.16)；如果在兩條件均已知的情況下，辨識結果可達 97.89% (表 4.17)。辨認的結果顯示辨識率均到達上限 (Up bound)，其中韻律狀態提供最重要的資訊，最能有效減低聲調混淆的情況，而辨識最困難的五聲，也因韻律狀態的資訊，辨識率提高到九成以上。

表 4.15 Treebank 已知 break type 之辨認結果

輸入聲調	聲調辨認 Confusion Matrix				
	一聲	二聲	三聲	四聲	五聲
一聲	82.4%	7.4%	1.6%	7.7%	0.9%
二聲	7.0%	81.8%	6.8%	2.3%	2.1%
三聲	2.9%	18.7%	64.8%	10.2%	3.4%
四聲	11.2%	2.1%	4.5%	80.1%	2.0%
五聲	14.9%	31.0%	14.9%	18.6%	20.6%
				正確率	74.83%

表 4.16 Treebank 已知韻律狀態之辨認結果

輸入聲調	聲調辨認 Confusion Matrix				
	一聲	二聲	三聲	四聲	五聲
一聲	97.8	0%	0%	2.1%	0.1%
二聲	0%	96.9%	2.3%	0.5%	0.3%
三聲	0.1%	4.9%	94.1%	0.2%	0.6%
四聲	2.9%	0.1%	0%	96.8%	0.3%
五聲	0.4%	4.8%	1.2%	2.0%	91.5%
				正確率	96.21%

表 4.17 Treebank 已知韻律狀態與 break type 之辨認結果

輸入聲調	聲調辨認 Confusion Matrix				
	一聲	二聲	三聲	四聲	五聲
一聲	99.6%	0%	0%	0.4%	0%
二聲	0%	97.6%	1.7%	0.4%	0.3%
三聲	0%	4.0%	95.6%	0.1%	0.2%
四聲	1.2%	0%	0%	98.7%	0.1%
五聲	0.4%	2.0%	1.2%	0.4%	96.0%
				正確率	97.89%

第五章　結論與展望

5.1 結論

本論文中，使用 MLP 聲調辨認器，利用前後音節之特徵參數，與利用 tone pair 之辨認方式，對於單一語者與非特定語者語料庫均有 85%以上的辨識率，而在利用四維正交參數於描述音高輪廓的結果，辨識率有同樣甚至較佳的結果，分析結果得知，五聲辨識率不佳，均只有五成至六成的辨識率，五聲外，三聲仍然是辨識率較差的部份，分別約有 8%、15%辨認成二聲，一聲與四聲的混淆情況也有超過 5%的情形。

對於利用韻律模型與 Viterbi search 的辨認方法，辨識率最高至 71.89%，其中三聲與五聲的辨識率較差，五聲辨識率除外，三聲與二聲的混淆情況最為嚴重，三聲約有 20%辨識成二聲，15%辨識成三聲。經由已知韻律狀態與 break type 的條件下，五種聲調的辨識結果均有效提昇，這是因為韻律狀態與 break type 態提供了聲調辨識重要的資訊。

5.2 未來之展望

本論文結果得知，韻律狀態與 break type 預測與的準確性能影響辨識的結果，實驗中，雖然加入韻律狀態及 break type 的 bigram 轉移機率，但實際上轉移機率範圍應該是更大範圍是有相關性的，不單只是與相鄰音節狀態有關，如能提供更有效預測方法，必能有效提升聲調的辨識率；本論文中計算量的降低，更快速的辨認方式，也是未來研究的方向。

此外，本方法中，未用到與前後音節相關的參數來做辨認，而是希望能直接

將影響因素扣除，如能直接加入前後音節參數於模型中，也許可以提高辨識的可能，另一原因，也因為使用音節參數較少之原因，例如音節長度、音節能量，這些資訊也是提供音調的資訊，如何有效利用更多參數幫助辨認，也是未來研究的重點。。

參考文獻

[1] Yih-Ru Wang , Sin-Horng Chen ,“Tone Recognition of Continuous Mandarin Speech Based on Neural Networks,”*IEEE Trans. On Speech and Audio Processing* ,*Vol*.3,*No*2,pp.146-150.March 1995.

[2] 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

[3] 王小川，“語音訊號處理”，全華科技圖書，中華民國九十三年三月。

[4] Wavesurfer Homepage：http：//www.speech.hth.se/wavesurfer/

[5] S. Young , G. Evermann , M. Gales , T. Hain , D. Kershaw , G. Moore , J. Odell , D. Ollason , D. Povey , V. Valtchev , P. Woodland , “The HTK Book”, Cambridge University , 2005

[6] 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.

[7] Chen-Yu Chiang, Xiao-Dong Wang, Yuan-Fu Liao, Yih-Ru Wang, Sin-Horng Chen, Keikichi Hirose, “Latent prosody model of continuous Mandarin speech,” *ICASSP* 2007

[8] Chen-Yu Chiang, Hsiu-Min Yu, Yih-Ru Wang and Sin-Horng Chen, “An Automatic Prosody Labeling Method for Mandarin Speech,” *Interspeech* 2007

[9] 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。

附錄一：破音字表

編號	國字	讀音	說明	編號	國字	讀音	說明
1	中	ㄓㄨㄥ	中央、中國	17	匹	ㄆㄧ	單槍匹馬、馬匹
		ㄓㄨㄥˋ	中的、中毒			ㄆㄧˇ	布匹、匹夫之勇
2	乾	ㄑㄧㄢˊ	乾卦、乾坤	18	占	ㄓㄢ	占卜、占夢
		ㄍㄢ	餅乾、乾杯			ㄓㄢˋ	占有、占據
3	了	ㄌㄧㄠˇ	了解、了不起	19	參	ㄘㄢ	參議員、參觀
		ㄌㄜ˙	「做完了！」			ㄕㄣ	人參、動如參商
4	供	ㄍㄨㄥ	口供、供給	20	吐	ㄊㄨˇ	吐痰、吐露
		ㄍㄨㄥˋ	供品、供奉			ㄊㄨˋ	吐血、嘔吐
5	便	ㄅㄧㄢˋ	方便、便利	21	否	ㄈㄡˇ	否定、不置可否
		ㄆㄧㄢˊ	便宜、便辟			ㄆㄧˇ	否卦、否極泰來
6	倒	ㄉㄠˇ	倒閉、跌倒	22	和	ㄏㄜˊ	總和、大和民族
		ㄉㄠˋ	倒影、倒轉			ㄏㄜˋ	唱和、附和
7	假	ㄐㄧㄚˇ	假借、假裝			ㄏㄢˋ	我和你
		ㄐㄧㄚˋ	假期、放假	23	哄	ㄏㄨㄥ	哄傳、哄堂大笑
8	傍	ㄅㄤˋ	依山傍水、依傍			ㄏㄨㄥˇ	連哄帶騙
		ㄅㄤ	傍晚、傍午	24	咽	ㄧㄢ	咽喉
9	傳	ㄔㄨㄢˊ	傳單、傳神			ㄧㄝˋ	哽咽
		ㄓㄨㄢˋ	左傳、傳記	25	哪	ㄋㄚˇ	哪知、哪個
10	冠	ㄍㄨㄢ	皇冠、雞冠			ㄋㄚ˙	「你又走哪？」
		ㄍㄨㄢˋ	冠禮、冠軍	26	喪	ㄙㄤ	居喪、喪亡
11	切	ㄑㄧㄝˋ	密切、一切			ㄙㄤˋ	喪失、喪心病狂
		ㄑㄧㄝ	切麵、切斷	27	喝	ㄏㄜ	喝水、喝酒
12	分	ㄈㄣ	分數、分析			ㄏㄜˋ	喝責、喝采
		ㄈㄣˋ	部分、本分	28	嘔	ㄡˇ	嘔吐、嘔心瀝血
13	創	ㄔㄨㄤˋ	開創、創舉			ㄡ	嘔歌、嘔啞
		ㄔㄨㄤ	重創、創傷			ㄡˋ	嘔氣、存心嘔我
14	勒	ㄌㄜˋ	韁勒、勒碑	29	嚇	ㄏㄜˋ	恐嚇、嚇嚇
		ㄌㄟ	勒緊、勒死			ㄒㄧㄚˋ	嚇唬、嚇了一跳
15	勞	ㄌㄠˊ	功勞、勞工	30	嚼	ㄐㄧㄠˊ	嚼舌、細嚼慢嚥
		ㄌㄠˋ	慰勞、勞軍			ㄐㄩㄝˊ	咀嚼
16	勝	ㄕㄥˋ	戰勝、尋幽覽勝	31	圈	ㄑㄩㄢ	圓圈、圈套
		ㄕㄥ	勝任、不勝枚舉			ㄐㄩㄢˋ	豬圈、羊圈

編號	國字	讀音	說明	編號	國字	讀音	說明
32	地	ㄉㄧˋ	大地	47	從	ㄘㄨㄥˊ	跟從、力不從心
		ㄉㄜ˙	慢慢地			ㄗㄨㄥˋ	侍從、從兄弟
33	塞	ㄙㄜˋ	阻塞、推諉塞責	48	悶	ㄇㄣˋ	煩悶、悶得慌
		ㄙㄞˋ	邊塞、要塞			ㄇㄣ	悶熱、悶飯
		ㄙㄞ	塞車、活塞	49	惡	ㄜˋ	罪惡、惡化
34	奇	ㄑㄧˊ	奇怪、奇門遁甲			ㄨˋ	羞惡、交惡
		ㄐㄧ	奇數、奇拜	50	應	ㄧㄥ	應該、應非難事
35	好	ㄏㄠˇ	好人好事			ㄧㄥˋ	應驗、應對
		ㄏㄠˋ	投其所好	51	扇	ㄕㄢˋ	門扇、扇子
36	宿	ㄙㄨˋ	宿舍、宿命			ㄕㄢ	扇風、扇惑
		ㄒㄧㄡˇ	一宿、整宿	52	挑	ㄊㄧㄠ	挑選、挑夫
37	將	ㄐㄧㄤ	將軍、打將下去			ㄊㄧㄠˇ	挑撥、挑燈
		ㄐㄧㄤˋ	勇將、上將	53	掃	ㄙㄠˇ	掃地、掃興
38	少	ㄕㄠˇ	缺少、少頃			ㄙㄠˋ	掃帚、掃把
		ㄕㄠˋ	少傅、少尉	54	擔	ㄉㄢ	負擔、擔擱
39	屬	ㄕㄨˇ	親屬、金屬			ㄉㄢˋ	重擔、扁擔
		ㄓㄨˇ	屬託、屬文	55	教	ㄐㄧㄠˋ	佛教、諄諄教誨、
40	差	ㄔㄚ	差數、差不多			ㄐㄧㄠ	教學生、教書匠
		ㄔㄞ	郵差、出差	56	散	ㄙㄢˋ	分散、散步
41	幾	ㄐㄧˇ	幾何學			ㄙㄢˇ	丸散、一盤散沙
		ㄐㄧ	庶幾、幾乎	57	數	ㄕㄨˋ	數目、數學
42	度	ㄉㄨˋ	制度、度日如年			ㄕㄨˇ	數來寶、數落
		ㄉㄨㄛˋ	忖度、度長絜大	58	暈	ㄩㄣ	頭暈眼花、暈倒
43	強	ㄑㄧㄤˊ	強壯、強權政治			ㄩㄣˋ	月暈、燈暈
		ㄑㄧㄤˇ	勉強	59	暴	ㄅㄠˋ	暴虐、暴躁
		ㄐㄧㄤˋ	倔強			ㄆㄨˋ	暴露、一暴十寒
44	彈	ㄉㄢˋ	彈弓、炸彈	60	曲	ㄑㄩ	歪曲、委曲求全
		ㄊㄢˊ	彈性、彈劾			ㄑㄩˇ	歌曲、曲高和寡
45	待	ㄉㄞˋ	對待、坐以待斃	61	曾	ㄗㄥ	曾孫、姓
		ㄉㄞ	待不住、待會兒			ㄘㄥˊ	曾經
46	得	ㄉㄜˊ	得到、得意	62	會	ㄏㄨㄟˋ	農會、都會
		ㄉㄟˇ	總得			ㄏㄨㄟˇ	限於「一會兒」、「多會兒」等詞音
		ㄉㄜ˙	飛得高、跳得遠				

編號	國字	讀音	說明	編號	國字	讀音	說明
63	朝	ㄓㄠ	朝露、朝氣蓬勃	78	省	ㄕㄥˇ	省分、中書省
		ㄔㄠˊ	朝代、朝廷			ㄒㄧㄥˇ	反省、省親
64	校	ㄒㄧㄠˋ	學校、上校	79	看	ㄎㄢˋ	看見、看病
		ㄐㄧㄠˋ	校量、校稿			ㄎㄢ	看門、看守
65	樂	ㄩㄝˋ	音樂、姓（如戰國時燕國名將樂毅）	80	相	ㄒㄧㄤ	相像、相親相愛
						ㄒㄧㄤˋ	福相、吃相
		ㄌㄜˋ	快樂、樂此不疲	81	禁	ㄐㄧㄣˋ	宵禁、禁止
66	橫	ㄏㄥˊ	縱橫、			ㄐㄧㄣ	弱不禁風、禁受
		ㄏㄥˋ	橫政、橫死	82	禪	ㄔㄢˊ	禪坐、禪語
67	沒	ㄇㄛˋ	沉沒、沒收			ㄕㄢˋ	禪讓、封禪
		ㄇㄟˊ	沒有、沒用	83	種	ㄓㄨㄥˇ	種子、種類
68	泥	ㄋㄧˊ	泥土、爛醉如泥			ㄓㄨㄥˋ	種田、接種
		ㄋㄧˋ	拘泥、泥古	84	稱	ㄔㄥ	稱號、稱讚
69	漲	ㄓㄤˋ	熱漲冷縮			ㄔㄥˋ	稱職、對稱
		ㄓㄤˇ	漲潮、水漲船高	85	空	ㄎㄨㄥ	天空、空歡喜
70	漂	ㄆㄧㄠ	漂浮、漂泊			ㄎㄨㄥˋ	空閒、空白
		ㄆㄧㄠˇ	漂母、漂白	86	答	ㄉㄚˊ	答數、回答
		ㄆㄧㄠˋ	漂亮、漂賬			ㄉㄚ	答應、羞答答
71	為	ㄨㄟˊ	行為、天下為公	87	累	ㄌㄟˇ	累犯、累積
		ㄨㄟˋ	為什麼			ㄌㄟˋ	家累、連累
72	率	ㄕㄨㄞˋ	表率、率由舊章	88	給	ㄐㄧˇ	年給、給事中
		ㄌㄩˋ	速率、或然率			ㄍㄟˇ	「給我拿來！」
73	甚	ㄕㄣˋ	甚好、欺人太甚	89	縫	ㄈㄥˊ	裁縫、縫紉
		ㄕㄣˊ	甚麼、作甚			ㄈㄥˋ	門縫、衣縫
74	畜	ㄔㄨˋ	畜生、家畜	90	署	ㄕㄨˇ	官署、環保署
		ㄒㄩˋ	畜產、畜養			ㄕㄨˋ	部署、署理
75	當	ㄉㄤ	當權、安步當車	91	翹	ㄑㄧㄠˊ	翹楚、翹舌
		ㄉㄤˋ	當鋪、勾當			ㄑㄧㄠˋ	翹翹板、翹辮子
76	的	ㄉㄧˋ	目的、標的	92	聽	ㄊㄧㄥ	聽講、垂簾聽政
		ㄉㄧˊ	的確			ㄊㄧㄥˋ	聽其自然
		ㄉㄜ˙	美麗的、慢慢的	93	背	ㄅㄟˋ	背後、離鄉背井
77	盛	ㄕㄥˋ	盛氣凌人、興盛			ㄅㄟ	背書包
		ㄔㄥˊ	盛飯、滾盛	94	脈	ㄇㄞˋ	動脈、脈搏
						ㄇㄛˋ	脈脈含情

編號	國字	讀音	說明	編號	國字	讀音	說明
95	興	ㄒㄧㄥ	興建、興旺	110	轉	ㄓㄨㄢˇ	轉學、颱風轉向
		ㄒㄧㄥˋ	興趣、興匆匆			ㄓㄨㄢˋ	暈頭轉向、公轉
96	荷	ㄏㄜˊ	荷花、薄荷	111	還	ㄏㄨㄢˊ	還原、償還
		ㄏㄜˋ	荷鋤、負荷			ㄏㄞˊ	時間還早
97	著	ㄓㄨˋ	名著、著作	112	那	ㄋㄚˋ	那個、那麼著
		ㄓㄨㄛˊ	棋高一著、著手			ㄋㄚˇ	那有這種事？
		ㄓㄠˊ	著火、睡著了！	113	都	ㄉㄨ	首都、都市
		ㄓㄠ	著涼、著急			ㄉㄡ	大都如此、都是
		ㄓㄜ˙	坐著	114	釘	ㄉㄧㄥ	鐵釘、補釘
98	藏	ㄘㄤˊ	藏匿			ㄉㄧㄥˋ	釘書機、釘門牌
		ㄗㄤˋ	西藏、三藏	115	重	ㄓㄨㄥˋ	體重、慎重
99	藉	ㄐㄧㄝˋ	憑藉、藉口			ㄔㄨㄥˊ	重複、重來
		ㄐㄧˊ	藉藉、聲名狼藉	116	量	ㄌㄧㄤˋ	容量、量力而為
100	處	ㄔㄨˋ	住處、益處			ㄌㄧㄤˊ	量體重、思量
		ㄔㄨˇ	處理、相處	117	鑽	ㄗㄨㄢ	鑽研、鑽木取火
101	號	ㄏㄠˋ	記號、坐號			ㄗㄨㄢˋ	鑽子、鑽戒
		ㄏㄠˊ	號哭、呼號	118	長	ㄔㄤˊ	專長、長短
102	行	ㄒㄧㄥˊ	人行道、行書			ㄓㄤˇ	尊長、首長
		ㄏㄤˊ	行家、太行山	119	間	ㄐㄧㄢ	隔間、房間
103	衝	ㄔㄨㄥ	要衝、衝突			ㄐㄧㄢˋ	間隙、間諜
		ㄔㄨㄥˋ	衝南走、太衝	120	阿	ㄜ	山阿、阿房宮
104	要	ㄧㄠˋ	摘要、需要			ㄚ	阿拉伯、阿伯
		ㄧㄠ	要求、要功	121	降	ㄐㄧㄤˋ	降落傘、霜降
105	覺	ㄐㄩㄝˊ	知覺、發覺			ㄒㄧㄤˊ	降龍伏虎、投降
		ㄐㄧㄠˋ	睡覺	122	難	ㄋㄢˊ	難堪、進退兩難
106	說	ㄕㄨㄛ	邪說、小說			ㄋㄢˋ	災難、問難
		ㄕㄨㄟˋ	說客、游說	123	養	ㄧㄤˇ	養育、撫養小孩
107	調	ㄊㄧㄠˊ	調羹、調色			ㄧㄤˋ	奉養、供養父母
		ㄉㄧㄠˋ	租庸調法、聲調	124	鮮	ㄒㄧㄢ	新鮮、海鮮
108	車	ㄐㄩ	車馬炮、學富五車			ㄒㄧㄢˇ	鮮有、鮮少
		ㄔㄜ	車衣服、姓	125	更	ㄍㄥ	變更、三更半夜
109	載	ㄗㄞˋ	刊載、載重量、姓			ㄍㄥˋ	自力更生、更好
		ㄗㄞˇ	一年半載				