國 立 清 華 大 學

碩 士 論 文

題目: 語 音 評 分

Speech Evaluation

系別 資訊工程學系 組別

學號姓名 894329 李俊毅 (Chun-Yi Lee)

指導教授 張智星博士 (Jyh-Shing Roger Jang)

中華民國九十一年六月

摘要

語音評分乃是結合了許多音訊處理以及語音辨識技術的一門學問，本論文從定義評分所需的特徵參數開始，實驗許多可行的特徵比對方式，期許建立一套合理的語音評分系統。

本論文包含兩個部分：第一部分為「利用標準語音資料的評分」，第二部分為「利用 HMM 及音高資料的評分」，分別從不同的觀點來對語音評分。

「利用標準語音資料的評分」，顧名思義我們可以想見這種評分方式將會有一個標準答案，亦即存在一標準語音，而測試的語音則要愈像此標準語音愈好，愈像者分數將會愈高；這部分所運用到的技術，包含特徵參數的擷取、圖樣比對方法的設計以及評分機制的建立等，其中特徵參數的部分我們是採用以下三個特徵，分別是音量強度曲線(Magnitude)、基頻軌跡(Pitch Contour)以及梅爾倒頻譜參數(Mel-Frequency Cepstral Coefficients)；音量強度曲線代表聲音音量強弱的變化趨勢；基頻軌跡代表聲音音高的起伏；梅爾倒頻譜參數則是代表聲紋，即語音的內容。在評分方面，我們使用「動態時間扭曲」(dynamic time warping)來比較這三個特徵的相似程度。

「利用 HMM 及音高資料的評分」主要是提供另一種語音評分的方式，以預先訓練好的聲學模型及聲調模型當成標準答案，經由語音辨認技術的使用，找出測試語音跟模型間的差異程度，並配合評分機制給與評分；此部分包含許多目前語音辨識常運用到的技術，例如在聲音辨識方面包含了隱藏式馬可夫模型(Hidden Markov Model)、樹狀網路(Tree Net)及維特比演算法(Viterbi Algorithm)等；在聲調辨識方面則包含了諸如 Orthogonal Expansion、Chebyshev Approximation、K-means 分群法及分類器的設計等等。在評分方面，我們利用測試語音在所有可能的 411 個音的排名並配合聲調給予評分。

Abstract

This thesis discusses several methods in speech evaluation, which is a study on computer evaluation of speech contents, fluency and intonation. It requires the techniques from audio signal processing and speech recognition. In order to develop an appropriate and consistent speech evaluation system, we define several useful speech features for our speech evaluation system and perform several experiments on feature matching methods. There are two parts in this thesis. The first one is "Evaluation using standard speech", and the other is "Evaluation using HMM and pitch contour".

"Evaluation using standard speech" is a method that evaluates the similarity between a test speech and the corresponding standard speech. We use various approaches for speech feature extraction, pattern matching, and similarity computation. In particular, we use magnitude contour, pitch contour, and mel-frequency cepstral coefficients as the features to generate a similarity score. Magnitude contours represent the variations in volume. Pitch contours represent the variations in pitches. Mel-frequency cepstral coefficients represent the contents of speech.

"Evaluation using HMM and pitch contour" is another speech evaluation paradigm that does not require the existence of a standard speech. Alternatively, we evaluate a test speech based on its similarity to a hidden Markov models (HMM) and tone models. Viterbi decoding is used to segment each character in a continuous sentence. Then the score of each character is computed through the ranking of 411 possible syllables and a tone recognition system.

致謝

在清華資工六年的歲月，是我一生中很重要的階段，從大學部到碩士班，我要特別感謝我的導師，也是我的指導教授－張智星老師，無論在專業領域的啟發或是做人處事的應對都讓我受益良多，尤其是在研究所這兩年的時光更讓我學習到很多理論與實作的技巧，並順利完成這篇論文。

另外，我要感謝在實驗室裡一同努力、一同成長的同學、學長以及學弟妹們，謝謝你們創造了這麼棒的研究環境，讓多媒體資訊檢索實驗室越來越好！

我也要感謝我的父母以及家人，你們是我最好的依靠，有你們的支持與鼓勵，讓我沒有後顧之憂，可以專心於學業的研究！

最後，我要感謝今年和我一起畢業的女友，謝謝你豐富了我的人生，謝謝你總是陪伴在我的身邊為我加油打氣。

目錄

圖表目錄

第1章 緒論

1.1 研究主題

本論文的研究主題是「語音評分」，包含「利用標準語音資料的評分」以及「利用 HMM 及音高資料的評分」，希望運用目前音訊處理和語音辨識的技術，分別從主觀和客觀兩個不同的角度來對語音評分。

利用標準語音資料的評分是一種比較主觀的評分方式，主要是使用圖樣比對(Pattern Matching)的方法，將欲測試的語音與標準語音資料作一比較，以期找出測試語音與標準語音的差異程度，並藉此對測試語音評分。

在本論文我們使用以下三個特徵參數：音量強度曲線(Magnitude)、基頻軌跡(Pitch Contour)以及梅爾倒頻譜參數(Mel-Frequency Cepstral Coefficients)；音量強度曲線代表聲音音量強弱的變化趨勢；基頻軌跡代表聲音音高的起伏；梅爾倒頻譜參數則是代表聲紋，即語音的內容。

利用 HMM 及音高資料的評分是一種比較客觀的評分方式，主要是從聲音以及聲調兩方面著手，希望找出測試語音與聲學模型及聲調模型的差異程度，並藉此對語音評分。

在本論文我們使用以下兩個特徵參數：基頻軌跡以及梅爾倒頻譜參數，分別做為聲調辨識以及聲音辨識的特徵參數;在實作上我們是先使用 Viterbi Decoding 將語音訊號切割成一個一個的字，即單音節，之後再對每一個音節比對聲音模型及聲調模型，並將辨識結果配合我們預先設計好的評分機制轉換成分數，即對此測試語音的評分。

1.2 語音評分系統簡介

本論文有兩個主體：「利用標準語音資料的評分」及「利用 HMM 及音高資料的評分」，我們分別實作兩個系統：第一個為英語評分系統，第二個是唐詩語音評分系統。

英語評分系統源自「利用標準語音資料的評分」，主要利用音量強度曲線、基頻軌跡及梅爾倒頻譜參數當作特徵參數，比較測試語音與標準語音的差異程度，並配合評分機制對兩者的差異程度評分。

唐詩語音評分系統源自「利用 HMM 及音高資料的評分」，主要是以聲學模型及聲調模型當作標準答案，將測試語音和這兩個模型比較，並依差異程度配合評分機制給與評分。

1.3 本論文研究方向和主要成果

本論文的研究方向為探討語音評分的方法，從定義評分所需的特徵開始，實驗許多可行的特徵比對方式，期許建立一套合理的語音評分系統。

本論文主要的成果為整合目前許多音訊處理及語音辨識相關的技術，運用在語音評分上，用以比較測試語音與標準語音的相似程度，並且建立合理的評分機制以及實作兩個不同機制的語音評分系統。

1.4 章節概要

本論文第二章將介紹「利用標準語音資料的評分」，並且將此單元所用到的一些技術於各小節中一一介紹，包含特徵參數的擷取、特徵參數正規化、圖樣比

對流程、相似度比對(DTW)、評分機制的建立及評分參數的調整等等。

第三章則是介紹另一種語音評分的機制－「利用 HMM 及音高資料的評分」，此章包含了許多語音辨識及聲調辨識的技巧，諸如語音辨識的隱藏式馬可夫模型(Hidden Markov Model)、語音辨識法則、樹狀網路以及語音訊號的切割等等；聲調辨識的部分則包含了基頻軌跡參數化的兩種方法：Orthogonal Expansion 及 Chebyshev Approximation 以及分群法則和分類法則等等。

第四章則是簡短的結論以及未來展望。

第2章 利用標準語音資料的評分

本章將介紹一種語音評分的方法:「利用標準語音資料的評分」,顧名思義我們可以想見這種評分方式將會有一個標準答案,亦即存在一標準語音,而測試的語音則要愈像此標準語音愈好,愈像者分數將會愈高。

在實作上我們主要是使用圖樣比對的方法,將欲評分的語音與標準語音作一比較,以期找出測試語音與標準語音之間的差異程度,並藉此對測試語音評分。

2.1 評分系統簡介

利用標準語音資料的評分系統流程如圖 2-1 所示,主要分為三大部分:第一部分為特徵參數的抽取(Feature Extraction),第二部分為圖樣比對(Pattern Matching)方法的設計,第三部分則為評分機制的建立;這三個部分將會在本章的各小節中逐一介紹;其中特徵參數的部分我們是採用以下三個特徵,分別是音量強度曲線(Magnitude)、基頻軌跡(Pitch Contour)以及梅爾倒頻譜參數(Mel-Frequency Cepstral Coefficients);經由圖樣比對之後我們提出一個評分機制,希望藉由此評分機制來對測試語音及標準語音之間的相似程度評分。

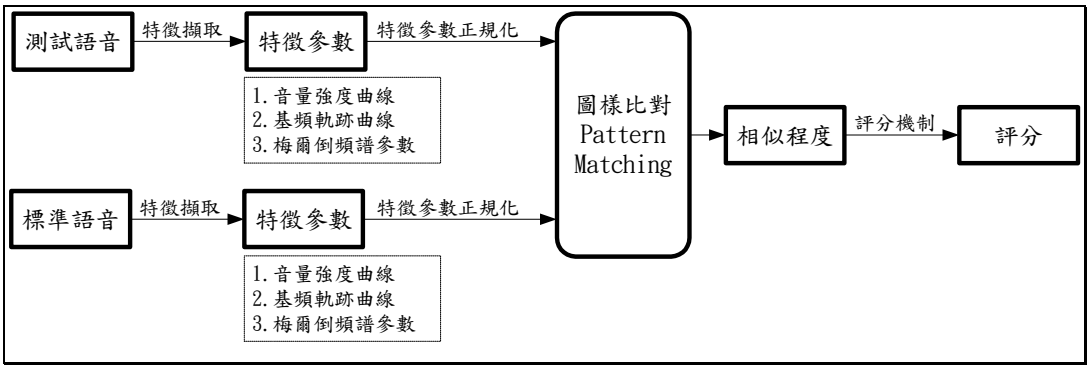


圖 2-1 利用標準語音資料之評分系統流程圖

2.2 特徵參數擷取

一般來說，聲音訊號是一種時變性(time varying)的訊號，其波形的變化相當快速，但是若將觀察此訊號的時間單位縮小，我們可以發現，其變化相當的緩慢，關於這種現象，我們稱其具有「短時間穩定」(short time stationary)之性質【1】，通常我們將此觀察的單位稱為一個「音框」(frame)，因此我們可以對聲音訊號做「短時距處理」(short time processing)，以切割音框(taking frame)的方法進行各種特徵參數的擷取；在利用標準語音資料評分的部分我們採用以下三個特徵參數，分別是音量強度曲線(Magnitude)、基頻軌跡(Pitch Contour)以及梅爾倒頻譜參數(Mel-Frequency Cepstral Coefficients)；各項特徵參數的擷取流程我們將在這節逐一介紹。

2.2.1 音量強度曲線

我們將取樣率 16kHz 的語音訊號經端點偵測(End-point Detection)【2】找出聲音頭尾端點之後取音框化，音框大小 512 點，約 32 毫秒，重疊(Overlap)為 170 點，約占一音框的三分之一，假設每一音框中的語音訊號以 $S_n(m)$ 表示，其中 $m = 0,1,...,M-1$，$n = 0,1,...,N-1$，N 為音框總數，亦即音量強度曲線的長度，M 為音框大小。

音量強度曲線定義為：

$$aveMag(n) = \frac{1}{M}\sum_{m=0}^{M-1}\left|S_n(m)\right| \quad ,n = 0,1,...,N-1$$

語音訊號經端點偵測後再求取音量強度曲線的示意圖如下頁圖 2-2 所示：

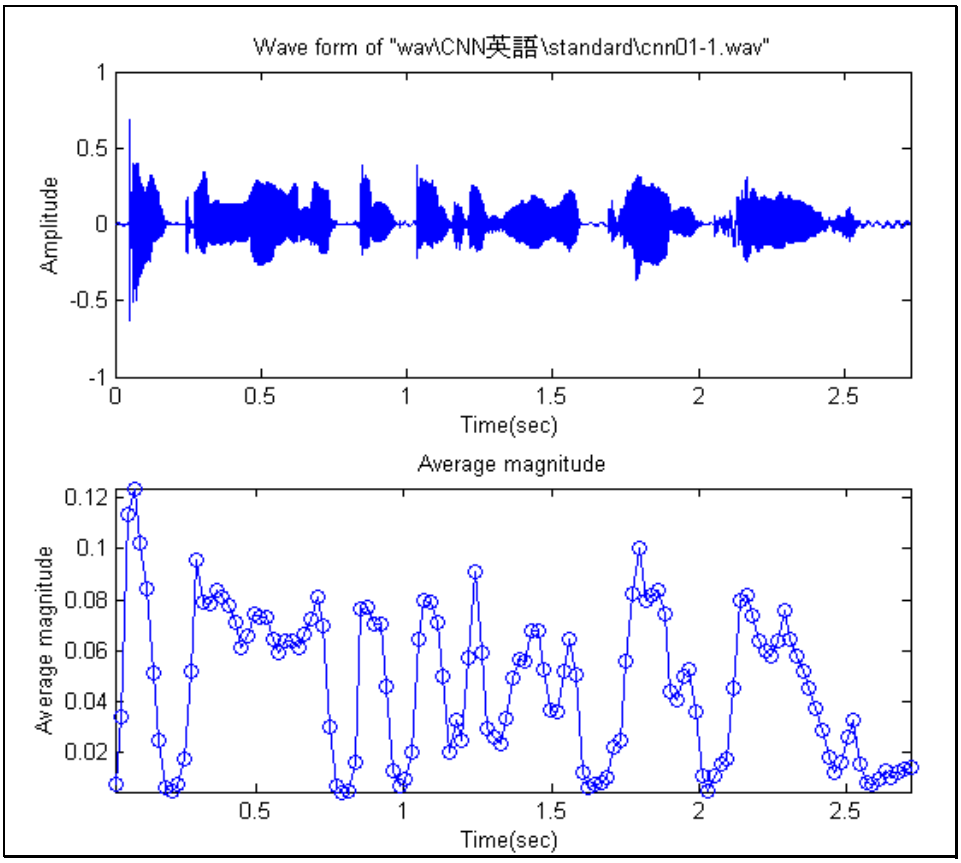


圖 2-2　音量強度曲線示意圖

2.2.2 基頻軌跡

求取基頻軌跡(pitch tracking)的方法有很多種，在此我們採用 Average Magnitude Difference Function【1】來擷取基頻，主要的流程如圖 2-3 所示：

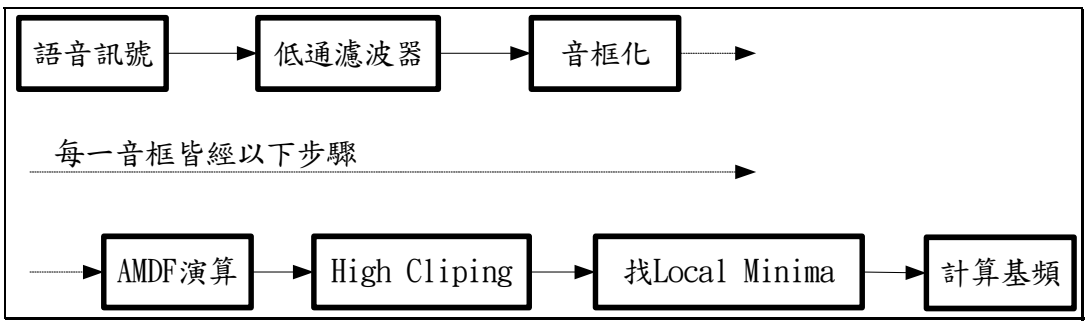


圖 2-3　基頻擷取流程圖

關於圖 2-3 我們將每一步驟條列如下：

1. **低通濾波器(low pass filter)**

我們將麥克風錄進來的聲音先做前處理，通過低通濾波器把雜訊和爆音過濾掉。

2. **取音框(taking frames)**

以 512 點為一框(frame)，音框和音框之間重疊 170 點，如此可以避免音框間的變化太過劇烈。

3. **AMDF 演算(Average Magnitude Difference Function)**

接下來對每個音框做 AMDF，找出相似波型重覆出現的週期，其公式如下：

$$AMDF_n(\tau) = \frac{1}{M}\sum_{m=0}^{M-\tau-1}\left|S_n(m) - S_n(m+\tau)\right|$$

M 為音框大小， τ 為平移量，由於 M 在此公式中為定值，我們可以省略計算。取一音框為例，其 AMDF 圖示如下：

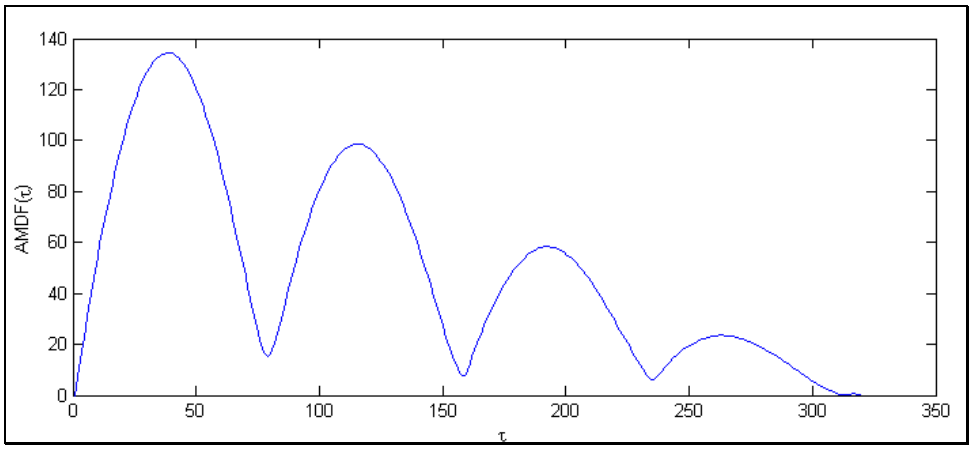


圖 2-4　AMDF 示意圖一

4. **High clipping**

經 AMDF 之後其 local minima 間的距離即是此聲音的週期，不過在算 local minima 之前，我們先用 High clipping 過濾這些訊號，示意圖如下：

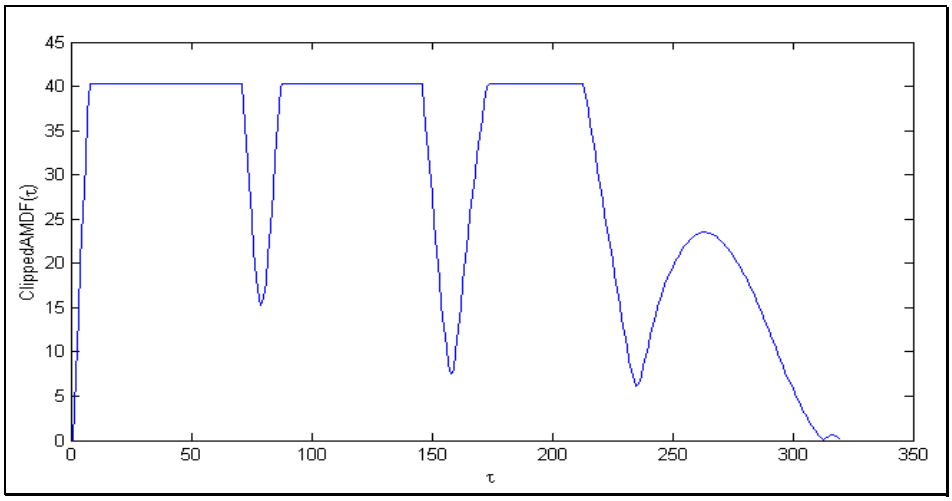


圖 2-5　AMDF 示意圖二

5. **找 local minima 及算出頻率**

經過 AMDF 及 high clipping 之後，我們可以利用相鄰 local minima 在時間軸上的距離找出聲音的週期，取其倒數即為基頻，圖示如下：

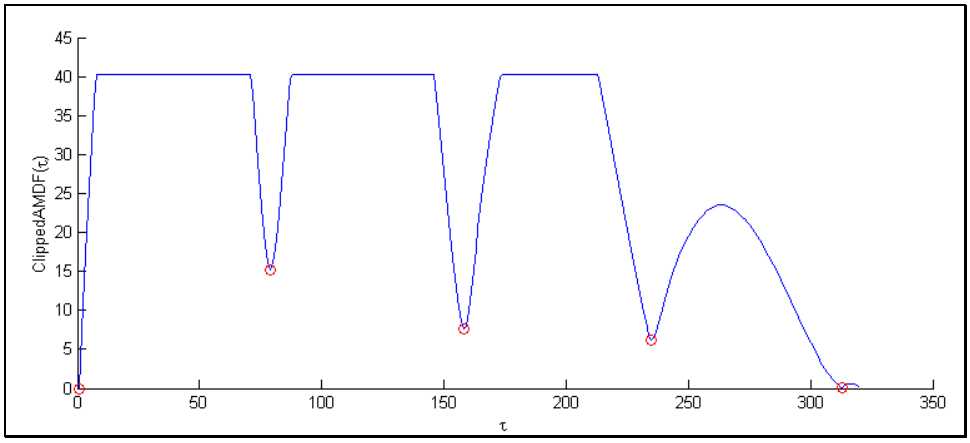


圖 2-6　AMDF 示意圖三

將每個音框重覆做步驟 3 到步驟 5 之後即可得到整個語音訊號的基頻軌跡，圖示如下：

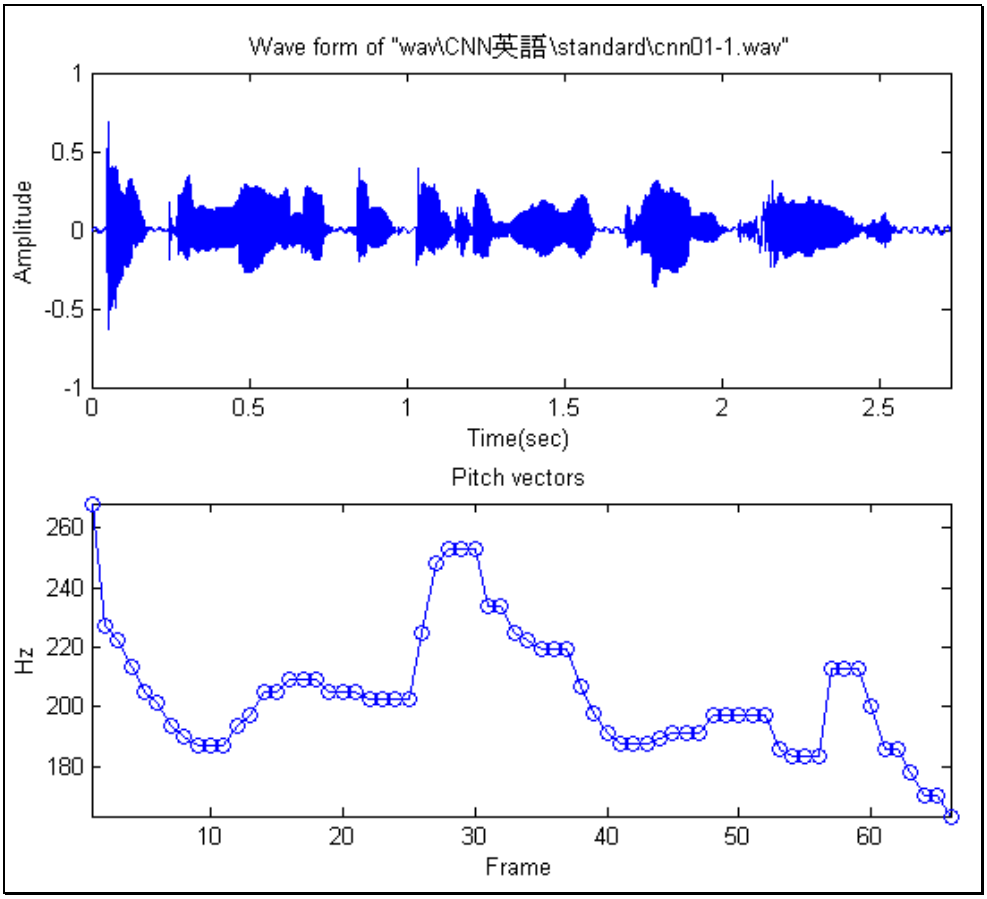


圖 2-7　基頻軌跡示意圖

2.2.3 梅爾倒頻譜參數

梅爾倒頻譜參數(Mel-Frequency Cepstral Coefficients)，縮寫為 MFCC【4】【18】，擷取方法簡述如下：

首先我們將 16KHz 語音訊號經預強調(Pre-emphasis)放大，主要是為了補償語音訊號受到發音系統所壓抑的高頻部分，係數是 0.975；之後取音框化，音框長度為 512 點，重疊 170 點，每個音框乘上漢明窗(Hamming window)，以補償

以音框為處理單位，在邊緣所造成的訊號不連續的現象；接著每個音框做快速傅利葉轉換(Fast Fourier Transform, FFT)，求出每個音框的頻譜；再帶入一組 20 個三角帶通濾波器(triangular bandpass filter)求出每一個頻帶的輸出對數頻譜 $m_j, j = 1,2,..,20$；最後再經餘弦轉換(cosine transform)即可求得 L 維的梅爾倒頻譜參數：

$$c_k = \sum_{j=1}^{P} m_j \cos(\frac{\pi k}{P}(j-0.5)), \quad k = 1,2,...L$$

其中 $p = 20$ 為三角帶通濾波器的數目，$L = 12$，即本論文使用 12 維的梅爾倒頻譜參數。

基本流程如下圖所述：

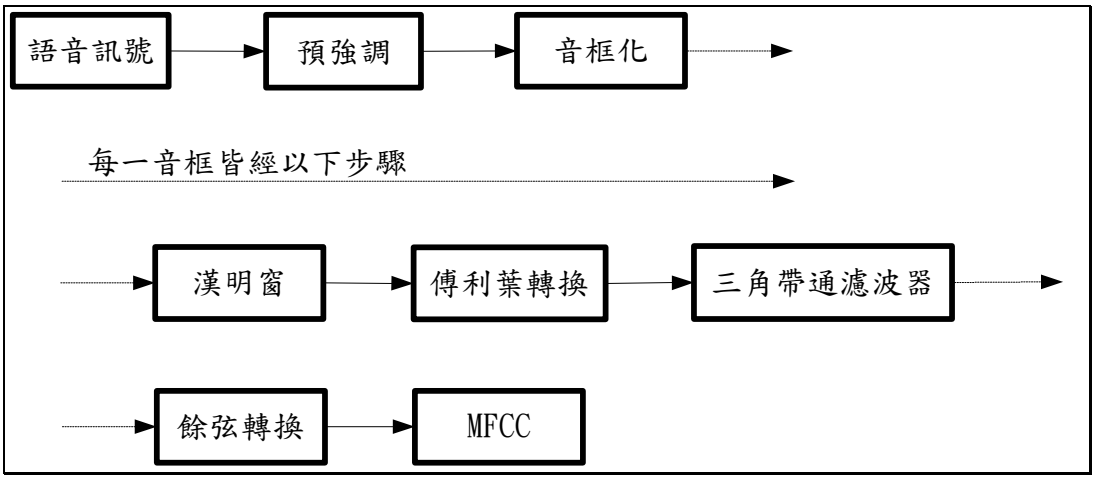


圖 2-8　12 維梅爾倒頻譜參數擷取流程

2.3 特徵參數正規化

在利用標準語音評分所使用的三個特徵參數中，梅爾倒頻譜參數在擷取的過程中已經使用 Vocal Track Length Normalization【19】的方法解決聲腔長度因人而異的變異性，然而音量強度曲線及基頻軌跡這兩個特徵仍存在一些個人的差異，例如麥克風的差異性及個人聲調長低不同等等，我們提出以下三種方法：Interpolation、Linear Scaling 及 Linear Shifting 以期將這兩個特徵參數正規化，另外我們也使用 Cepstral Means Subtraction 的方法，用來作為通道效應的補償，細節將在以下小節逐一介紹。

2.3.1 解決特徵參數長短不一的問題：Interpolation

由於測試語料與標準語料的長度不一定完全一樣，所以我們使用一維的內差法，可以在音量強度曲線及基頻軌跡的解析度盡量不失真前提下將這兩個特徵參數的長度伸長或縮短，此方法可以有效解決特徵參數長短不一的問題。

2.3.2 解決麥克風差異性：Linear Scaling

經由麥克風錄進來的音量大小會隨著麥克風的不同而使著音量強度曲線 $aveMag(n)$ 有些差異，我們假設其間存在一倍數的關係，若標準語料的音量強度曲線定義為 $aveMag_1(n)$，測試語料之音量強度曲線定義為 $aveMag_2(n)$，我們希望找出一參數 θ，使得兩曲線之誤差 $\bar{e}$ 越小越好，假設

$$\vec{A} = aveMag_2(n) = \begin{bmatrix} aveMag_2(0) \\ aveMag_2(1) \\ \cdots \\ aveMag_2(N-1) \end{bmatrix},$$

$$\vec{y} = aveMag_1(n) = \begin{bmatrix} aveMag_1(0) \\ aveMag_1(1) \\ \cdots \\ aveMag_1(N-1) \end{bmatrix}$$

兩曲線存在 $\vec{A}\theta + e = \vec{y}$ 的關係，由 Least-squares estimator【5】我們可以得知以下的結果：

$$\theta = (\vec{A}^T \vec{A})^{-1} \vec{A}^T \vec{y}$$

微調後的測試語料音量強度曲線假設為 $aveMag_2'(n)$，其公式如下：

$$aveMag_2'(n) = \vec{A}\theta = aveMag_2(n) * \theta, \quad n = 0,1,...,N-1$$

2.3.3 解決個人音高差異性：Linear Shifting

由於每個人的聲調高低不一致，通常女生的聲調較高，男生的聲調略低，而且我們語音評分著重在聲調的抑揚頓挫亦即其基頻軌跡高低起伏之變化趨勢，因此我們有必要對聲調高低作一平移(Shifting)【6】的動作以期解決此個人的差異性，我們所用的方法如下：

假設標準語料之基頻軌跡以 $f_1(x)$、測試語料之基頻軌跡以 $f_2(x)$ 表示，其中 $x = 0,1,...,N-1$，N 為基頻軌跡的長度，我們以 $f_1(x)$ 為基準調整 $f_2(x)$，調整後的基頻軌跡假設為 $\hat{f}_2(x)$，定義如下：

$$\hat{f}_2(x) = f_2(x) - \frac{1}{N}\sum_{k=0}^{N-1} f_2(k) + \frac{1}{N}\sum_{k=0}^{N-1} f_1(k)$$

此平移的動作即是將兩段基頻軌跡的平均值調成一致。

2.3.4 解決未知的通道效應：Cepstral Mean Subtraction

倒頻譜平均值消去法(Cepstral Mean Subtraction)【7】主要的精神就是將倒頻譜參數每一維視為隨機變數，將其期望值改成零。

公式如下：

$$\tilde{x}_c(t) = y_c(t) - \bar{b}$$

$$\bar{b} = \frac{1}{T}\sum_{t=1}^{T} y_c(t)$$

其中為$\bar{b}$估測出來的通道值，$y_c(t)$為觀察到的特徵參數，共有T個音框，$\tilde{x}_c(t)$為補償後的特徵參數。

2.3.5 語者正規化：Vocal Tract Length Normalization

我們使用 VTLN (Vocal Tract Length Normalization)的方法【19】解決語者正規化的問題，主要原理是利用第三共振峰(F3)頻率較不易變動的特性，藉以調整測試語音的線性頻率尺度，正規化係數α定義為：

參考語音的平均 F3 值除以測試語音的平均 F3 值。

有了正規化係數α之後就可以用下式來調整測試語音的頻率尺度：

$$F_N = \alpha \cdot F$$

在實作上正規化係數α是運用在圖 2-8 的傅利葉轉換及三角帶通濾波器之間，主要的精神即是將測試語音的原始頻率F利用正規化係數α對應到調整後的頻率F_N。

2.4 圖樣比對方法設計

本節我們將介紹如何找出測試語音及標準語音兩者之間的差異程度，主要的精神是利用圖樣比對的方法【8】，針對我們使用的三個特徵分別設計不同的比對方式，對於音量強度曲線我們使用 Interpolation 及 Linear Scaling 來調整參數，調整後我們再求取此特徵的最小 Dynamic Time Warping 平均誤差；對於基頻軌跡我們使用 Interpolation 及 Linear Shifting 來調整參數，調整後我們一樣找其最小的 DTW 平均誤差；至於梅爾倒頻譜參數的比對我們首先是將特徵先經由 Cepstral Mean Subtraction 的方法解決未知的通道效應，再採取動態時間扭曲(Dynamic Time Warping)的方法，找出測試語音與標準語音最相似的音框對應及平均 DTW 距離。Dynamic Time Warping 的實作我們將於下一單元介紹。

2.4.1 音量強度曲線比對方法

音量強度曲線的比對方法設計如圖 2-9 所示，假設標準語音的音量強度曲線為 v1，測試語音之音量強度曲線為 v2，我們以 v1 為基準去調整 v2 後再比對即可以得到兩者的距離 $dist_1$，代表其差異程度。

我們以 Interpolation 來解決特徵長度不一致的問題，並以 Linear Scaling 解決麥克風差異性；在算距離時我們是採用 DTW 的方法，求取兩段特徵最相近的平均距離。

在實作上，我們會將測試語音的特徵在時間軸上左右平移幾個音框，以彌補特徵可能沒有對齊的因素，再重覆比對流程並取距離最小者為兩特徵之間的平均距離 $dist_1$。

比對流程如下頁圖 2-9 所示：

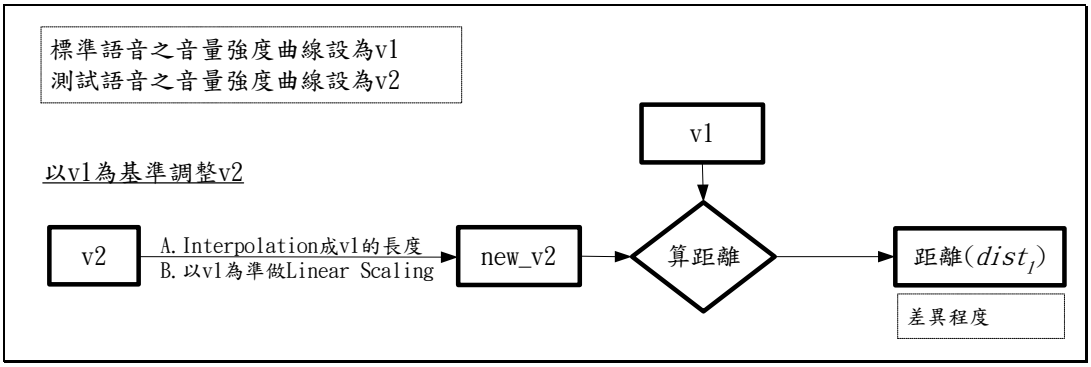


圖 2-9　音量強度曲線比對流程圖

2.4.2 基頻軌跡比對方法

基頻軌跡的比對方法設計大致如同上一小節音量強度曲線的方法，所不同的是特徵在 Interpolation 成相同長度後是以 Linear Shifting 來解決個人音高不一致的差異；在算距離時我們是採用 DTW 的方法，求取兩段特徵最相近的平均距離。

在實作上，我們會將測試語音的特徵在時間軸上左右平移幾個音框，以彌補特徵可能沒有對齊的因素，再重覆此比對流程並取距離最小者為兩特徵之間的平均距離 $dist_2$。

詳細比對流程如下圖所示：

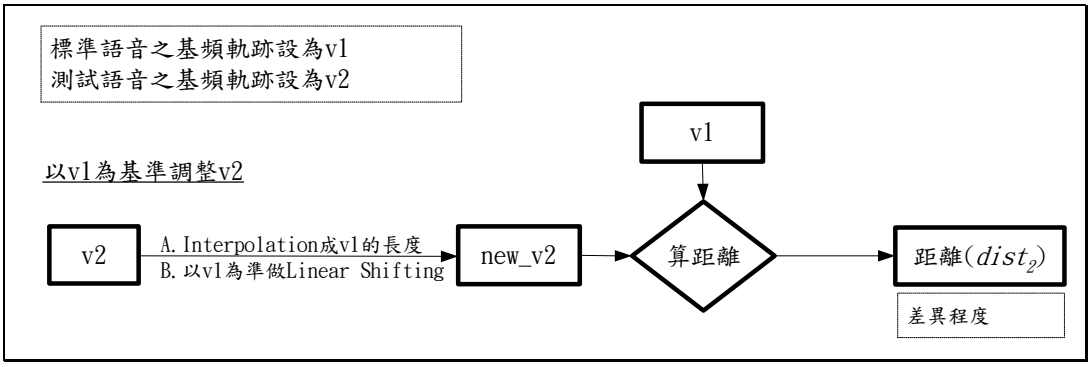


圖 2-10　基頻軌跡比對流程圖

2.4.3 梅爾倒頻譜參數比對方法

梅爾倒頻譜參數的比對我們首先是將特徵先經由 Cepstral Mean Subtraction 的方法解決未知的通道效應，再採取動態時間扭曲(Dynamic Time Warping)的方法，找出測試語音與標準語音最相似的音框對應及平均距離 $dist_3$，主要的流程如圖 2-11 所示；在實作上，我們會將測試語音的特徵在時間軸上左右平移幾個音框，以彌補特徵可能沒有對齊的因素，再重覆此比對流程並取距離最小者為兩特徵之間的平均距離 $dist_3$。

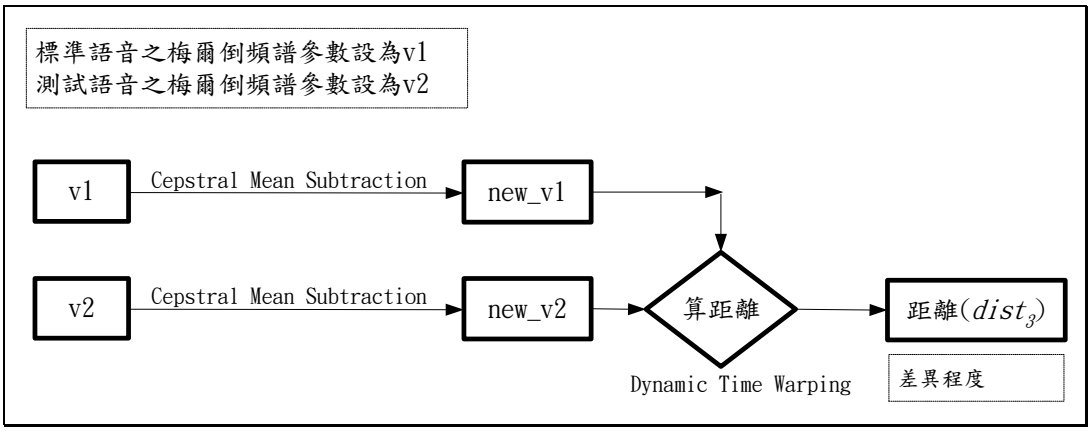


圖 2-11　梅爾倒頻譜參數比對流程圖

2.5 相似度比對：Dynamic Time Warping

兩段特徵參數的距離估測我們採用「動態時間扭曲」(dynamic time warping)的方法，簡稱 DTW【6】【17】，此方法在語音訊號處理中是一種很常用來做相似度比對的方法，其主要的精神在於提供一個具有更大彈性的相似度比對法，使測試資料能透過伸展或壓縮，找到與參考資料間最小誤差的非線性對應。

舉一例子，假設我們的測試資料為 **t**，長度為 **I**，參考資料為 **r**，長度為 **J**，下圖是常見的動態時間扭曲比對示意圖：

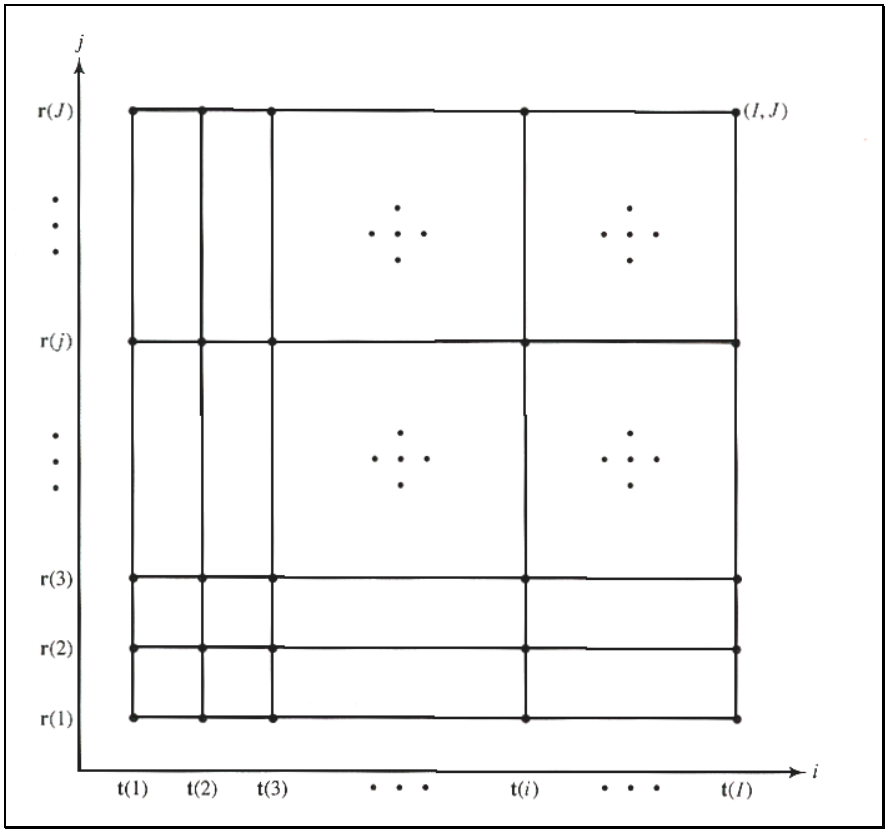


圖 2-12　動態時間扭曲比對示意圖【1】

DTW 的主要目的便是在 **t**、**r** 構成的平面上找出一條最佳的對應路徑 $path(i_k, j_k)$，即是使得測試資料與參考資料間的距離 D 為最小，並且使得 $t(i_k)$ 對應到 $r(j_k)$，其中，$k = 1, 2, \ldots, \mathrm{K}$， i_k 與 j_k 都必須遞增，以數學式子表示如下：

$$D = \sum_{k=1}^{K} d(i_k, j_k)$$

$$d(i_k, j_k) = dist(t(i_k), r(j_k))$$

其中 $d(i_k, j_k)$ 可以為任意一種距離測量方式，最常見的就是歐幾里得距離，在此我們是計算測試語音及標準語音之兩組梅爾倒頻譜參數以音框為單位的歐幾里得距離。

在實際運算上，我們可以透過動態規劃的方式找出最佳的路徑。首先我們先定義出可能的起始點與終點，A、D 分別為參考資料可能的起點範圍與終點範圍；B、C 分別為測試資料可能的起點範圍與終點範圍, 如圖 2-8 所示。因此，我們可以定義出在起始點與終點的 D ：

$$D(i_k, 0) = 0,\ i_k \in B$$
$$D(i_k, 0) = Inf,\ i_k \notin B$$
$$D(0, j_k) = 0,\ j_k \in A$$
$$D(0, j_k) = Inf,\ j_k \notin A$$

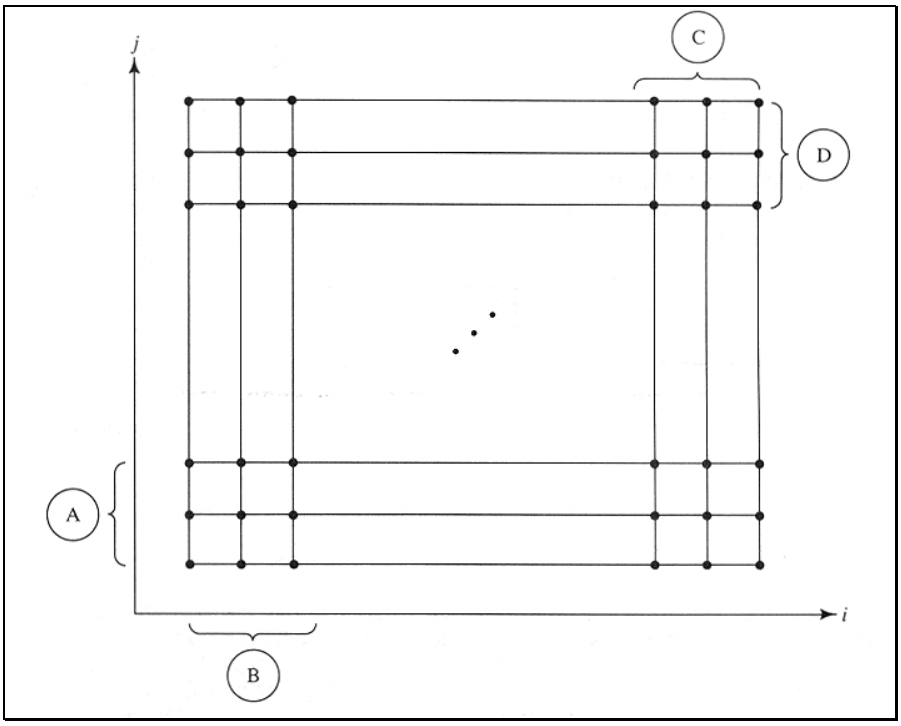


圖 2-13　DTW 彈性起始點與終點示意圖【1】

至於終點範圍的限制，則可以在動態規劃最後回溯最佳路徑時予以限制。

基於局部最佳值能導致整體最佳值的概念，定義出每一點 $D(i_k, j_k)$ 可能的路徑來源 $parent(i_k, j_k)$ 如下：

$$D(i_k, j_k) = parent(i_k, j_k) + d(i_k, j_k)$$
$$parent(i_k, j_k) = \min(D(p, q))\ ,\ p \le i_k, q \le j_k$$
$$d(i_k, j_k) = dist(t(i_k), r(j_k))$$

其中，p 與 q 的限制可以根據各種不同的比對資料及問題類型予以變化。下圖為幾種在語音辨識上較常見的 p、q 限制：

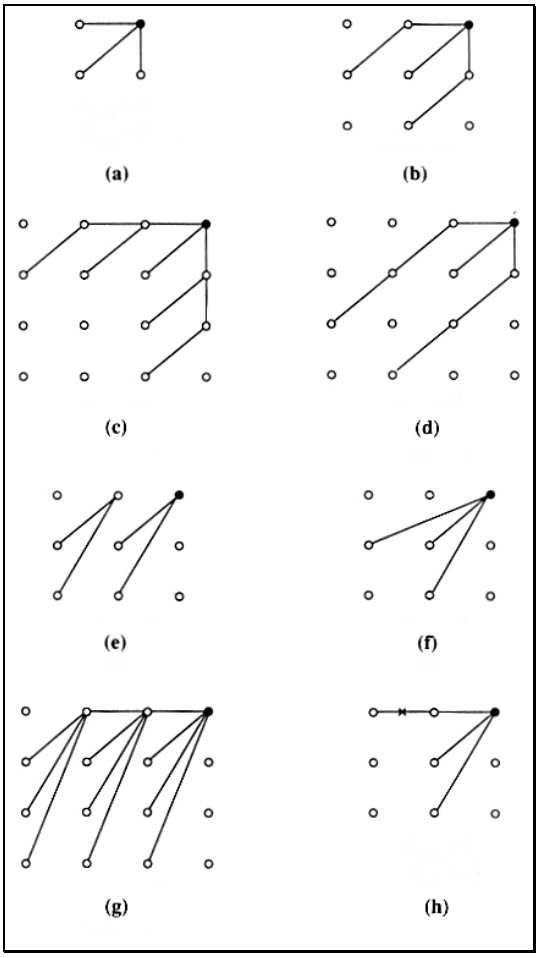


圖 2-14　常見的 DTW 限制條件【1】

2.6 評分機制

本小節將介紹利用標準語音資料評分的評分機制，首先我們先設定測試語音與標準語音兩相同特徵比對之後的結果，其距離與分數間的關係，設定公式如下：

$$score = \frac{100}{1 + a(dist)^b}$$

由這個公式我們就可以將距離轉換成分數，舉一例子，假設我們可以經由實驗得到測試語音與標準語音的基頻軌跡相似度在分數為 90 分時，距離大約為 5；分數在 60 分時，距離大約為 6，如此我們就可以求得 a 和 b，有了 a 和 b 之後就可以得到距離和分數的關係圖：

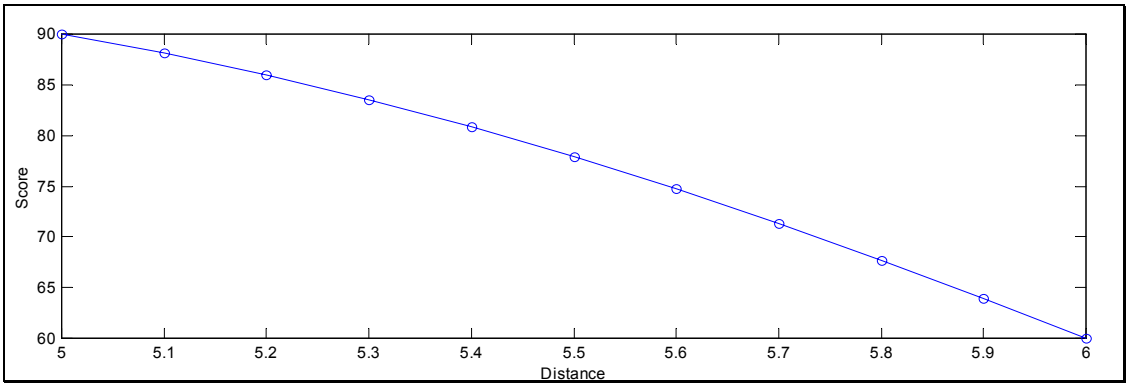


圖 2-15　距離轉換成分數示意圖一

使用本論文的公式即使距離比我們設定的大或小時皆可以合理的轉換分數到 100 跟 0 的區間裡，如下圖所示：

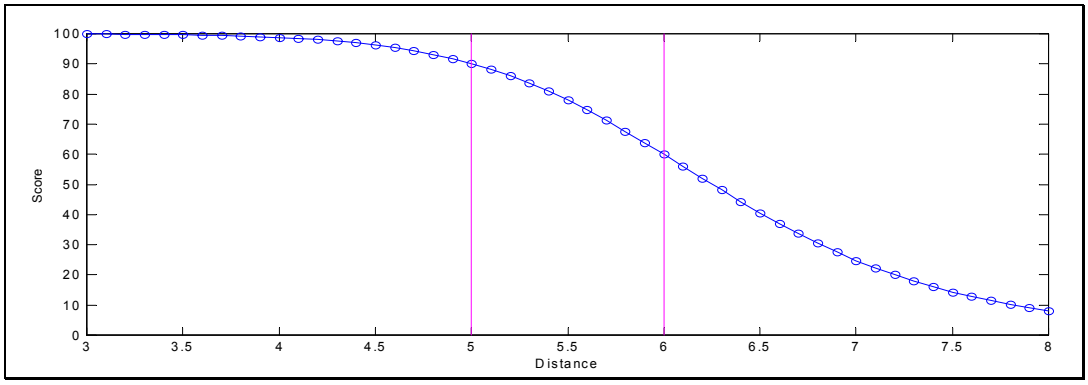


圖 2-16　距離轉換成分數示意圖二

2.7 評分參數調整

當一測試語音進來時，我們分別和標準語音比較音量強度曲線、基頻軌跡及梅爾倒頻譜參數三個特徵，分別算出距離$[dist_1, dist_2, dist_3]$後，距離轉分數的公式可以定義如下：

$$score = w_1 \cdot \frac{100}{1 + a_1 (dist_1)^{b_1}} + w_2 \cdot \frac{100}{1 + a_2 (dist_2)^{b_2}} + w_3 \cdot \frac{100}{1 + a_3 (dist_3)^{b_3}}$$

$$a_1, a_2, a_3, b_1, b_2, b_3 > 0, \quad w_1 + w_2 + w_3 = 1$$

$a_1, a_2, a_3, b_1, b_2, b_3$為距離轉成分數的參數，$w_1, w_2, w_3$為三個特徵的權重。為了求得$a_1, a_2, a_3, b_1, b_2, b_3, w_1, w_2, w_3$我們設計了以下實驗：首先我們先收集 10 句 CNN 互動英語的句字，當成標準語音，再請實驗室同學依此 10 句錄音，當成測試語音，總共收集了 320 句測試語音，每一句跟標準答案比對，會算出以下三個距離：

$$[dist_1, dist_2, dist_3]$$

假設每句的分數皆為 90 分；另將每句測試語音再跟其它標準語音(內容不同者)比對，亦可以得到三個距離，假設其分數為 30 分，收集了這些距離及分數之後，我們使用 MATLAB 的 fminsearch 函式，利用 Simplex Downhill Search 的方法，即可找出不錯的$a_1, a_2, a_3, b_1, b_2, b_3, w_1, w_2, w_3$組合，有了這組參數之後，只要我們算出測試語音及標準語音三個特徵的距離$[dist_1, dist_2, dist_3]$，就可以轉換其對應的分數。

第3章 利用 HMM 及音高資料的評分

3.1 評分系統簡介

本章以長庚大學語音實驗室【9】所收集的語料為基礎，從聲音及聲調兩方面著手，期望由比較客觀的角度來對語音訊號評分。

利用 HMM 及音高資料的評分包含許多目前語音辨識常運用到的技術，例如在聲音辨識方面包含了隱藏式馬可夫模型(Hidden Markov Model)、樹狀網路(Tree Net)及維特比演算法(Viterbi Algorithm)等；在聲調辨識方面則包含了諸如 Orthogonal Expansion、Chebyshev Approximation、K-means 分群法及分類器的設計等等技術，這些方法將在之後的小節裡有詳細的介紹。

簡單的語音評分流程如下圖所示，一句測試語音進來之後，我們先斷詞將一句話分成一個一個的字，再分別對每個字做聲音及聲調的辨識，之後再依辨識結果的排名配合評分機制給與評分。

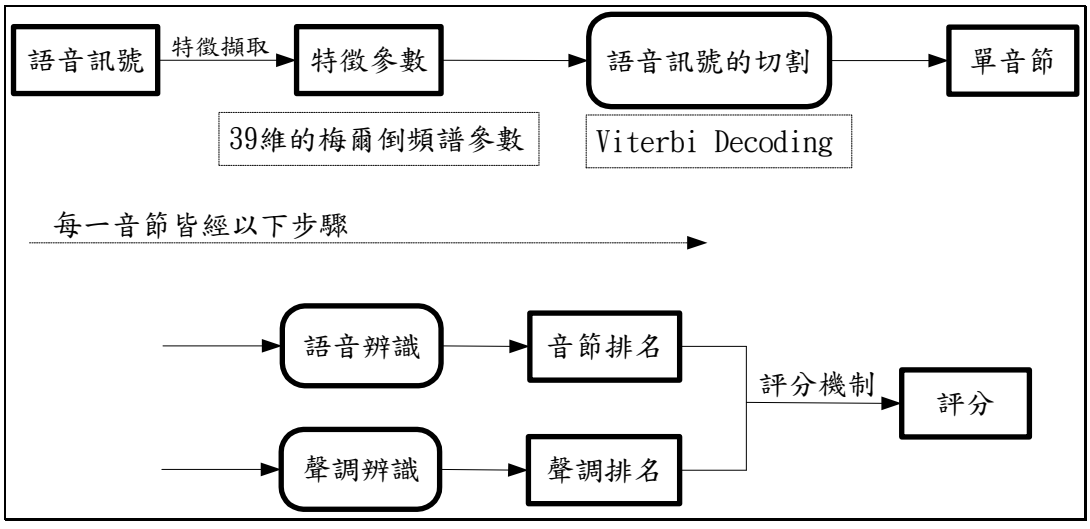


圖 3-1　利用 HMM 及音高資料的評分流程圖

3.2 語音辨識

3.2.1 語音辨識流程

語音辨識整個流程包含特徵向量的擷取、隱藏式馬可夫模型訓練以及語音辨識和音節轉文字模組等，以下為其基本流程圖：

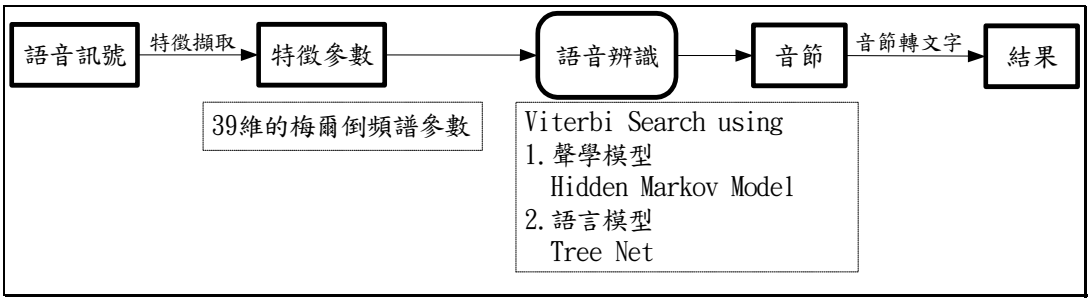


圖 3-2　語音辨識流程圖

整個語音辨識的流程如上圖所示，從一語音訊號經過了特徵擷取，取出語音中的特徵，在此我們採用 39 維的梅爾倒頻譜參數當作特徵參數，然後利用這些特徵參數透過聲學模型及語言模型，利用維特比演算法(Viterbi algorithm)找出最相似的音節，之後再經過音節轉文字即可輸出結果；其中特徵參數擷取、聲學模型：隱藏式馬可夫模型、語言模型：樹狀網路、維特比演算法等都是語音辨識中很重要的核心技術，我們將會在以下小節逐一介紹。

3.2.2 特徵參數擷取

本小節將介紹梅爾倒頻譜參數(Mel-Frequency Cepstral Coefficients)的求法，縮寫為 MFCC，以下為擷取的基本流程：

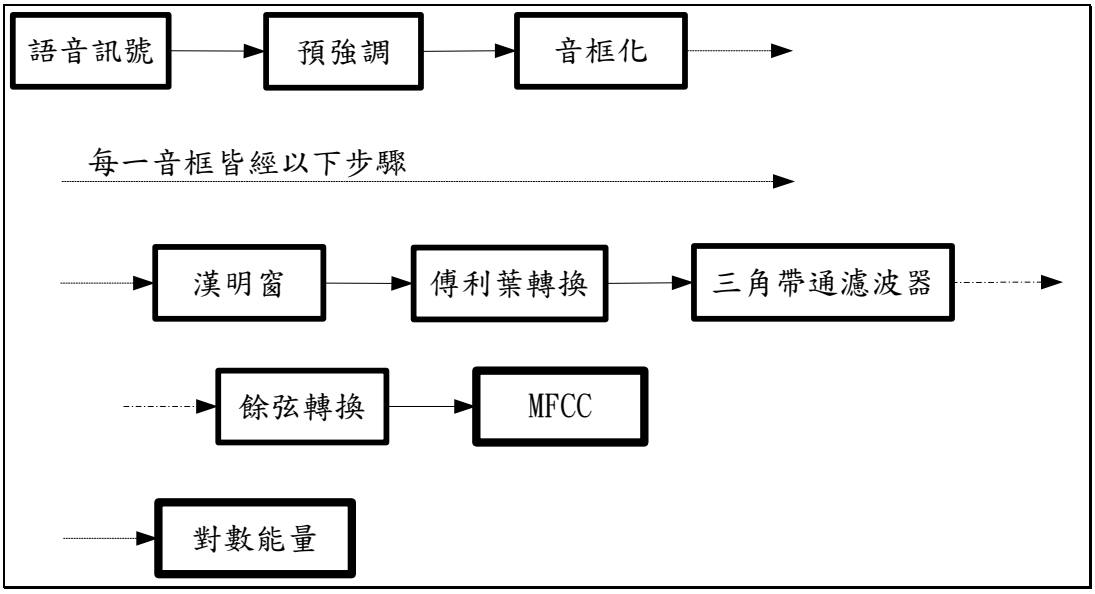


圖 3-3　梅爾倒頻譜參數擷取流程

首先我們將 16KHz 語音訊號經預強調(Pre-emphasis)放大，主要是為了補償語音訊號受到發音系統所壓抑的高頻部分，係數是 0.975；之後取音框化，音框長度為 512 點，重疊 170 點，每個音框乘上漢明窗(Hamming window)，以補償以音框為處理單位，在邊緣所造成的訊號不連續的現象；接著每個音框做快速傅利葉轉換(Fast Fourier Transform, FFT)，求出每個音框的頻譜；再帶入一組 20 個三角帶通濾波器(triangular bandpass filter)求出每一個頻帶的輸出對數頻譜 $m_j, j = 1,2,..,20$，經由研究【10】發現，人類對於低頻聲音的感知能力較強，約呈線性關係；而對於高頻聲音的感知能力則較弱，約呈對數關係，所以在設計三角帶通濾波器時就以低頻多取、高頻少取為其精神所在；最後再經餘弦轉換(cosine transform)即可求得 L 維的梅爾倒頻譜參數：

$$c_k = \sum_{j=1}^{P} m_j \cos(\frac{\pi k}{P}(j-0.5)), \quad k = 1,2,...L$$

其中 $p = 20$ 為三角帶通濾波器的數目，$L = 12$，即本論文使用 12 維的梅爾倒頻譜參數，

本論文使用 12 維的梅爾倒頻譜參數與 1 維的對數能量，組成基本的 13 維特

徵參數，再以這 13 維做為基礎，取其一階差量倒頻譜參數與二階差量倒頻譜參數，全部合起來總共 39 維的梅爾倒頻譜特徵參數，示意圖如圖 3-4 所示。

差量的意義為倒頻譜參數相對於時間的斜率，也就是代表倒頻譜參數在時間上的動態變化程度。其公式如下:

$$\Delta C_m(t) = \frac{\sum_{\tau=-M}^{M} \tau \cdot C_m(t+\tau)}{\sum_{\tau=-M}^{M} \tau^2} = \frac{\sum_{\tau=1}^{M} \tau\left(C_m(t+\tau) - C_m(t-\tau)\right)}{2 \cdot \sum_{\tau=1}^{M} \tau^2}, m = 1,2,...,L$$

這裡的 M 取 2，代表視窗寬度為 5 個音框，*t* 代表哪一個音框。

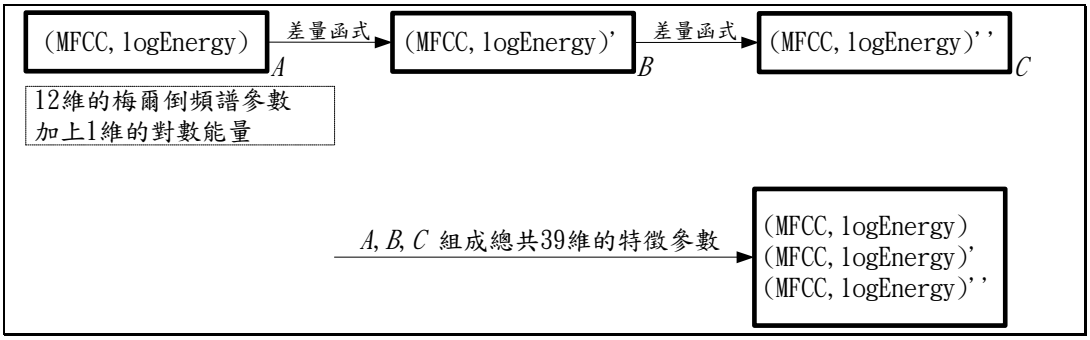


圖 3-4　39 維梅爾倒頻譜特徵參數示意圖

3.2.3 聲音單元介紹

國語語音中每一個音節即代表一個字，而音節又是由音素或是聲韻母所組成，目前我們使用的聲學模型是不考慮聲調的，因此國語語音的音節約有 415 個，在本章我們使用的聲音單元【9】是音節內右相關的聲學單位，舉一個例子，例如「家」這個字(即一個音節)，其長庚拼音為「jia」，其音節內右相關的聲音單元為「sil+j」,「j+i」,「i+a」,「a+sil」，此聲音單元將視為語音的最小單位，並為每一單元訓練其聲學模型，意即每一單元都有一個模型(Model)，本論文的聲學模型是採用隱藏式馬可夫模型。

3.2.4 隱藏式馬可夫模型

本章語音辨識所用到的聲學模型是以隱藏式馬可夫模型(Hidden Markov Model, HMM)為基礎所訓練出來的，經由前人的研究【9】 【10】【11】，我們得知隱藏式馬可夫模型基本上是一種雙重隨機過程，而之所以稱為隱藏式是因為其中有一組隨機過程是隱藏的，看不見的，在語音中就如同人類在發聲的過程中其發聲器官狀態變化是看不見的，好比喉嚨、舌頭與口腔的變化是不可能從可觀測的語音訊號序列看出來的。而另一組隨機過程稱為觀測序列(observation sequence)，它是由狀態觀測機率(state observation probability)來描述在每個狀態下觀測到各種語音特徵參數的機率分佈。

HMM 的特性正好適用於描述語音的特性，我們可以把每個狀態看成是聲道(vocal tract)正處於某個發聲組態(articulatory configuration)，而狀態觀測機率則描述了在某個發聲狀態下聽到各種聲音的可能性。

HMM 的狀態觀測機率函式$b_j(o_t)$是採用高斯混合密度函數或稱高斯混合模型(Gaussian Mixture Model, GMM)。

在本論文中每一個聲音單元皆有一個 HMM，一個模型有 3 或 5 個狀態(State)，示意圖如下：

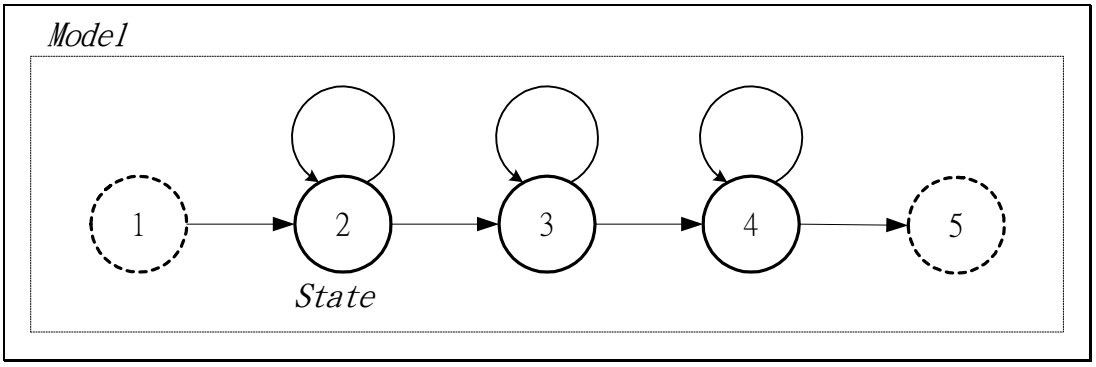


圖 3-5　模型(Model)與狀態(State)示意圖

狀態觀測機率函式$b_j(o_t)$定義【12】為：

$$b_j(o_t)=\prod_{s=1}^{\#S}\left[\sum_{m=1}^{\#M_s} w_{jsm}G_{jsm}\right]^{r_s}$$

其中$\#S$代表 Stream 的數目；r_s為 Stream 的權重(weight)，在本論文中為 1；$\#M_s$代表 Steam 為s時，mixture 的數目；w_{jsm}及G_{jsm}則分別代表在狀態j下，Steam 為s時，mixture 為m時高斯函數的權重及高斯機率密度函數，G_{jsm}的定義如下：

$$G_{jsm}=g(x;\mu,\Sigma)=\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

其中d為維度，μ及Σ分別代表此高斯機率密度函數的平均值(mean)及共變異矩陣(Covariance Matrix)，這些參數決定了此機率密度函數的特性，諸如函數形狀的中心點、寬窄及走向等。

在本論文中我們使用 3 個 Stream，即$\#S=3$，mixture 數目則有兩組，一組為(6, 2, 2)，另一組為(10, 10, 10)，每組 3 個數值依序代表每一個 stream 包含 mixture 的個數，以第一組(6, 2, 2)為例，圖示如下：

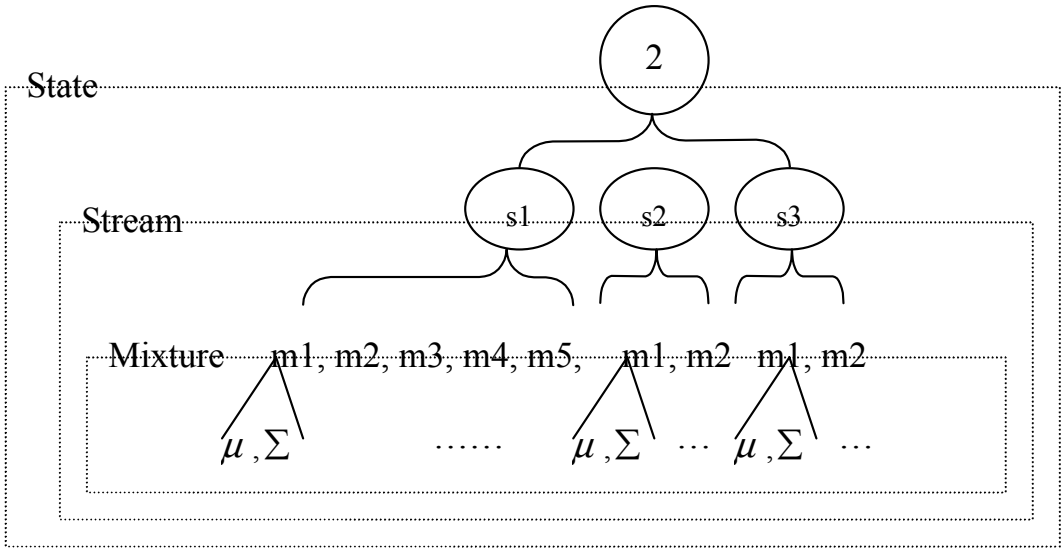


圖 3-6　State, Stream, Mixture 示意圖

3.2.5 語音辨識法則

在本論文中的聲學模型是採用長庚大學語音實驗室所訓練出來的，我們是做語音訊號與已建立之聲學模型的比對動作，當然，在辨識前，語音訊號也是經由特徵參數擷取的動作，且參數的定義也如之前章節所述；而要辨識這一段語音訊號，其實就是決定觀測序列究竟由哪些模型的序列來描述是最恰當的，我們使用維特比演算法(Viterbi algorithm)來找出與觀測序列匹配的最佳狀態序列。

首先我們先介紹在實作上如何求狀態觀測機率 $b_j(o_t)$，由上一小節我們得知

$$b_j(o_t)=\prod_{s=1}^{3}\left[\sum_{m=1}^{\#M_s} w_{jsm}G_{jsm}\right]$$

取對數(log)後得到

$$\sum_{s=1}^{3}\log\left(\sum_{m=1}^{\#M_s} w_{jsm}G_{jsm}\right)$$

我們以 stream1 為例，介紹其對數機率求取過程，stream2 及 stream3 亦同理，在 stream1 的對數機率為

$$\log\left(w_1G_1+w_2G_2+...\right)$$

可簡化成【12】

$$\begin{aligned}&\log\left(w_1G_1\right)+\log\left(w_2G_2\right)+...\\&=\log\left(w_1\right)+\log\left(G_1\right)+\log\left(w_2\right)+\log\left(G_2\right)+...\end{aligned}$$

而高斯函數如上一小節所述

$$G=\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

因此

$$\log(G)=-\frac{1}{2}\log\left[(2\pi)^d|\Sigma|\right]-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$$

我們定義

$$GConst = \log\left((2\pi)^d |\Sigma|\right)$$

GConst 可以依下式求得

$$GConst = d\log(2\pi) + \log(|\Sigma|)$$

$$= 13 \cdot \log(2\pi) + \sum_{i=1}^{13} |\text{var}[i]|$$

另一部分$(x-\mu)^T \Sigma^{-1}(x-\mu)$亦可依以下方法求得

$$(x-\mu)^T \Sigma^{-1}(x-\mu)$$

$$= \underset{1*13}{\begin{bmatrix} A & B & .. & .. \end{bmatrix}} \cdot \underset{13*13}{\begin{bmatrix} C & 0 & .. & 0 \\ 0 & D & 0 & .. \\ .. & 0 & .. & 0 \\ 0 & .. & 0 & .. \end{bmatrix}} \cdot \underset{13*1}{\begin{bmatrix} A \\ B \\ .. \\ .. \end{bmatrix}}$$

$$= \underset{1*13}{\begin{bmatrix} AC & BD & .. & .. \end{bmatrix}} \cdot \underset{13*1}{\begin{bmatrix} A \\ B \\ .. \\ .. \end{bmatrix}}$$

$$= \underset{1*1(scalar)}{A^2C + B^2D + ...}$$

了解如何求取狀態觀測機率$b_j(o_t)$之後，接著介紹維特比演算法：

首先假設觀察序列$\vec{O} = \{o_1, o_2, ..., o_T\}$的最佳狀態序列為$\vec{q} = \{q_1, q_2, ..., q_T\}$，並以$\delta_t(i)$代表從頭開始，直到時間點$t$時的觀測值$o_t$為狀態$i$的最大機率，以下式表示：

$$\delta_t(i) = \max_{q_1, q_2, ..., q_{t-1}} P\left(q_1, q_2, ..., q_{t-1}, q_t = i, \vec{O} \mid \lambda\right)$$

由【4】我們可得知

$$\delta_{t+1}(i) = [\max_i \delta_t(i) \cdot a_{ij}] \cdot b_j(o_t)$$

其中λ為 Hidden Markov Models，a_{ij}為狀態 i 跳到狀態 j 的轉移機率(Transition probability)，$b_j(o_t)$為狀態 j 時出現o_t的觀測機率。

維特比演算法【4】的步驟如下：

1. 初始化：

$$\delta_1(i) = \pi_i b_i(o_1)$$
$$\psi_1(i) = 0, \quad 1 \le i \le N$$

$\psi_t(j)$用以回溯(Backtracking)

2. 遞迴步驟：

$$\delta_t(j) = [\max_{1 \le i \le N} \delta_{t-1}(i) \cdot a_{ij}] \cdot b_j(o_t)$$
$$\psi_t(j) = \arg\max_{1 \le i \le N} [\delta_{t-1}(i) \cdot a_{ij}], \quad 2 \le t \le N, \quad 1 \le j \le N$$

3. 結束(Termination)：

$$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$
$$q_T^* = \arg\max_{1 \le i \le N} [\delta_T(i)]$$

4. 回溯步驟(state sequence backtracking)

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, ..., 1$$

如此即可找出最佳狀態序列。

3.2.6 樹狀網路

經由研究【9】我們可以得知：樹狀網路(tree net)可以有效的提升辨識率及降低音節的複雜度，因此本論文採用樹狀網路來做為語言模型，樹狀網路的基本精神以圖 3-7 為例作一說明，假設欲辨識的詞只有「台北縣」、「台中市」、「新竹

縣」、「新竹市」四個詞，我們可以看出「台北縣」及「台中市」兩個詞，都是以「台」為樹根，分別長出「北」跟「中」兩字，因為只接這兩字而已，所以搜尋的空間相對的減少，除了提升辨識率之外，搜尋的速度也會提升許多。

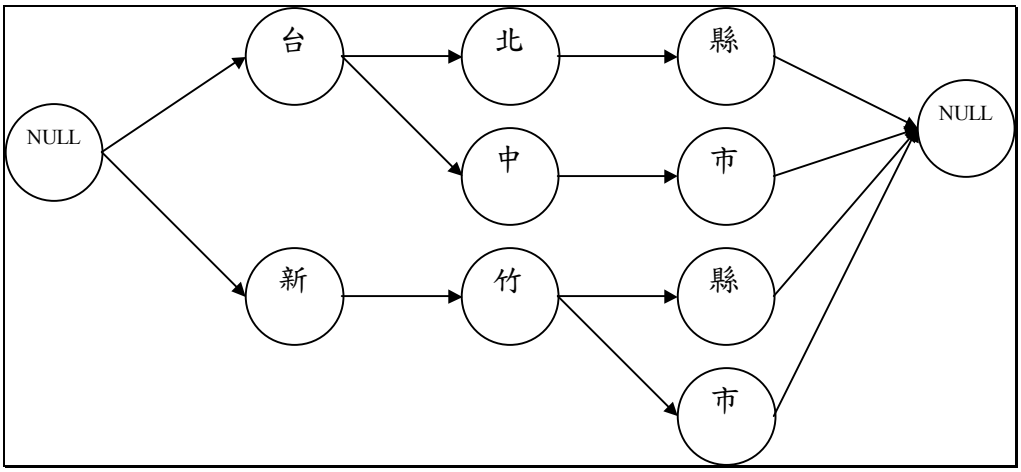


圖 3-7　樹狀網路示意圖

3.2.7 語音訊號的切割

我們將語音訊號切割主要的目的是希望將連續的語音切割成獨立的音節，這樣才容易對一句話中的每一個字評分。在本論文中，我們使用 Viterbi decoding 的方法【13】，亦即在已知語音訊號內容的前提下，解出語音訊號的狀態序列，圖 3-8 說明了切音的流程。

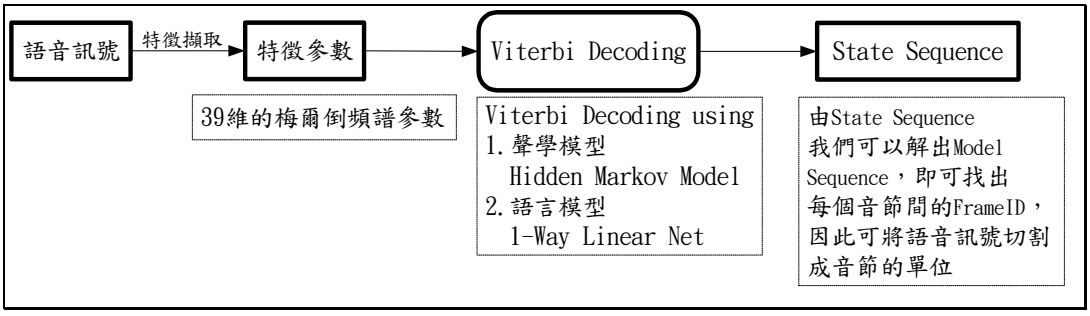


圖 3-8　切音流程圖

3.3聲調辨識

3.3.1 國語聲調簡介

國語語音有兩個明顯的特徵：一是有聲調性、二是單音節，亦即每一個中文字皆對應到一個音節，每一個音節有一個聲調；在聲調方面，有一聲調、二聲調、三聲調、四聲調及輕聲調共五種聲調(本論文不考慮輕聲調)。由前人的研究【14】我們可以得知這幾種聲調間的差異性，主要在於基頻軌跡(pitch contour)變化的趨勢，各聲調的基頻軌跡大致如下圖所示：

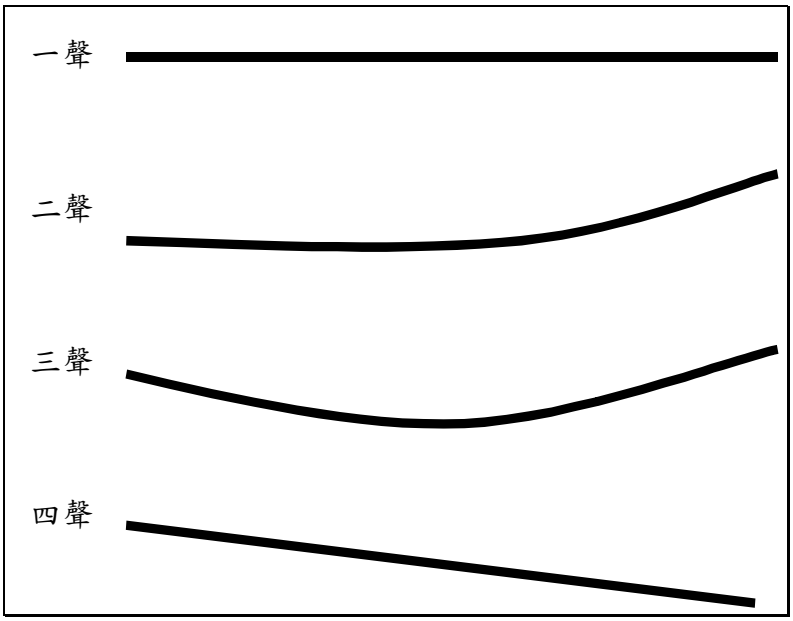


圖 3-9　各聲調基頻軌跡趨勢示意圖

由上圖我們可以觀察得知：一聲調的基頻軌跡接近一水平線，二聲調的基頻軌跡先持平後持續上揚，三聲調的基頻軌跡先平緩下降後再上揚，四聲調的基頻軌跡則是由高處持續下降。在台灣地區的語言習慣中，三聲調經常是發音不完全的，因此其尾端上揚的特性並不明顯，如此也增加了聲調辨認的困難度。

3.3.2 聲調辨識流程

這一小節我們要介紹聲調辨識系統的流程，包括特徵擷取、語料訓練以及聲調辨識等，如下圖所示：

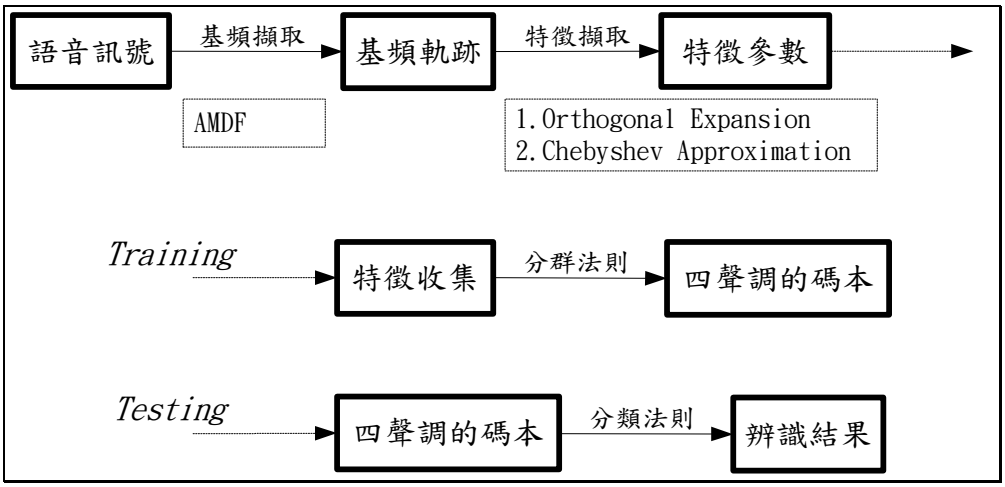


圖 3-10　聲調訓練及辨識流程圖

整個聲調辨識系統如上圖所示，我們可以得知，從輸入一語音訊號，經過基頻擷取(pitch tracking)之後，我們得到了一段基頻軌跡(pitch contour)，由於基頻擷取我們在第二章已經介紹過了，所以在此不在贅述，我們從特徵參數的擷取，使用 Orthogonal Expansion 及 Chebyshev polynomial fitting 來表示一段基頻軌跡開始介紹，之後我們介紹聲調辨識系統主要的兩個部分，一個是語料的訓練(Training)，一個是聲調的辨別(Testing)。在語料訓練方面，我們採取 K-means 分群法，將每一聲調的所有特徵係數分群找其代表點，即得到代表此一聲調的碼本(Code Book)，由於每一個聲調有兩種參數化表示法(Orthogonal Expansion 及 Chebyshev polynomial fitting)，因此每一聲調有兩組碼本，分別定義為 $CodeBook_O(j), j = 1,...4$ 及 $CodeBook_C(j), j = 1,...4$ ；在聲調辨別方面，我們將欲測試的語料求取其特徵參數後與之前訓練出來的碼本進行 1-Nearest Neighbor 的分類，即會得到 $MinDist_O(j), j = 1,...4$ 及 $MinDist_C(j), j = 1,...4$ ，經距離的正規化後找其距離最小的類別，即為聲調辨識的結果。

3.3.3 基頻軌跡參數化(一)：Orthogonal Expansion

每一段基頻軌跡我們可以用一個三次的多項式來表示，如

$$f(x) = a_0 + a_1 x^1 + a_2 x^2 + a_3 x^3, \quad x = 0,1,...,N$$

不過由於每段基頻軌跡的長度不一定，因此我們有必要對其長度做一正規化，如下式所示，即將其長度正規化至[0,1]的區間：

$$f(\frac{i}{N}) = a_0 + a_1(\frac{i}{N})^1 + a_2(\frac{i}{N})^2 + a_3(\frac{i}{N})^3, \quad i = 0,1,...,N$$

接著我們利用 Gram-Schmidt Orthogonalization Procedure 將基底

$$\{1, \frac{i}{N}, (\frac{i}{N})^2, (\frac{i}{N})^3\}$$

轉成

$$\{\phi_0(\frac{i}{N}), \phi_1(\frac{i}{N}), \phi_2(\frac{i}{N}), \phi_3(\frac{i}{N})\}$$

兩兩互相垂直且內積(Inner Product)等於一的基底【15】，如下列式子所示：

$$\phi_0(\frac{i}{N}) = 1$$

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

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

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

詳細之公式推導請參考附錄一。

有了這些基底之後，每一段的基頻軌跡即可利用這些基底來逼近，如下式所示：

$$\hat{f}(\frac{i}{N}) = \sum_{j=0}^{3} a_j \cdot \phi_j(\frac{i}{N}), \quad i = 0,1,...,N$$

$$where \qquad a_j =< f(\frac{i}{N}), \phi_j(\frac{i}{N}) >= \frac{1}{N+1}\sum_{i=0}^{N} f(\frac{i}{N}) \cdot \phi_j(\frac{i}{N})$$

3.3.4 基頻軌跡間的距離估測

兩段基頻軌跡間的距離我們定義為其 root-mean-square 的距離。假定兩段等長的基頻軌跡 $f_1(\frac{i}{N})$ 及 $f_2(\frac{i}{N})$，其距離可表示為：

$$D = \{\frac{1}{N+1}\sum_{i=0}^{N}[f_1(\frac{i}{N}) - f_2(\frac{i}{N})]^2\}^{1/2}$$

由上節可知我們可以將基頻軌跡 $f_1(\frac{i}{N})$ 及 $f_2(\frac{i}{N})$ 以相同的基底逼近，如下式：

$$\hat{f}_1(\frac{i}{N}) = \sum_{j=0}^{3} a_{1j} \cdot \phi_j(\frac{i}{N}), \quad i = 0,1,...,N$$

$$\hat{f}_2(\frac{i}{N}) = \sum_{j=0}^{3} a_{2j} \cdot \phi_j(\frac{i}{N}), \quad i = 0,1,...,N$$

因此兩段逼近後基頻軌跡的距離如下式所示：

$$D \approx D' = \{\frac{1}{N+1}\sum_{i=0}^{N}[\hat{f}_1(\frac{i}{N}) - \hat{f}_2(\frac{i}{N})]^2\}^{1/2}$$

$$= \{\frac{1}{N+1}\sum_{i=0}^{N}[\sum_{j=0}^{3} a_{1j} \cdot \phi_j(\frac{i}{N}) - \sum_{j=0}^{3} a_{2j} \cdot \phi_j(\frac{i}{N})]^2\}^{1/2}$$

$$= \{\frac{1}{N+1}\sum_{i=0}^{N}[\sum_{j=0}^{3}\sum_{k=0}^{3}(a_{1j} - a_{2j})(a_{1k} - a_{2k}) \cdot \phi_j(\frac{i}{N})\phi_k(\frac{i}{N})]\}^{1/2}$$

$$= \{\sum_{j=0}^{3}\sum_{k=0}^{3}(a_{1j} - a_{2j})(a_{1k} - a_{2k}) \cdot <\phi_j(\frac{i}{N}), \phi_k(\frac{i}{N}) >\}^{1/2}$$

$$= \{\sum_{j=0}^{3}(a_{1j} - a_{2j})^2\}^{1/2}$$

由上式我們可得知：兩段基頻軌跡間的距離可以逼近於$(a_{10}, a_{11}, a_{12}, a_{13})$與$(a_{20}, a_{21}, a_{22}, a_{23})$的 Euclidean distance。

3.3.5 基頻軌跡參數化(二)：Chebyshev Approximation

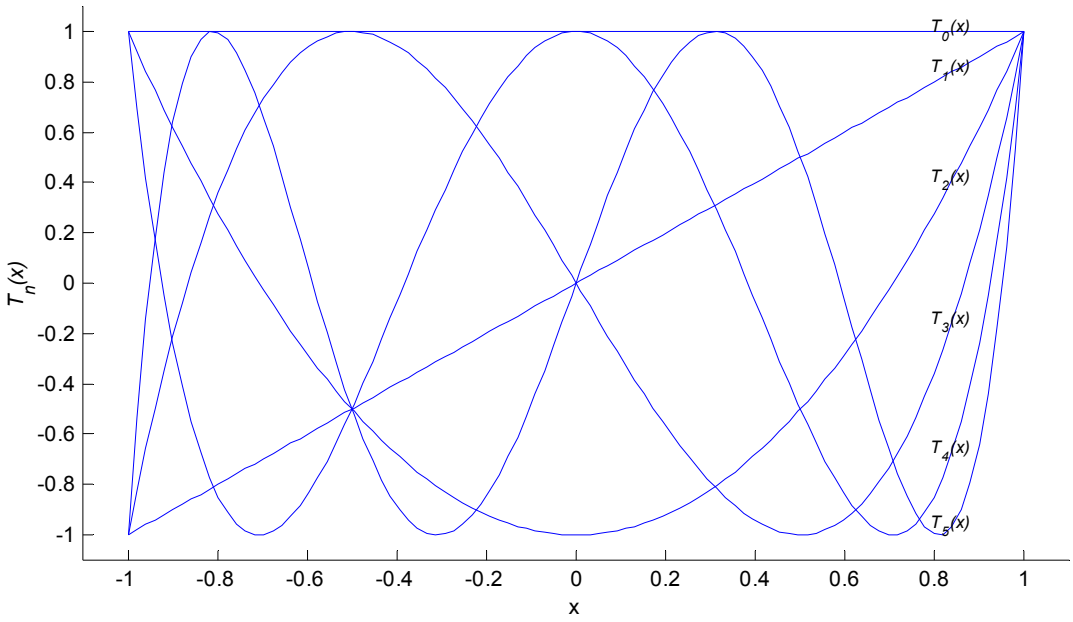
Chebyshev 多項式【16】可以表示成$T_n(x)$，n 為其 degree，其公式如下：

$$T_n(x) = \cos(n \arccos x)$$

經推導可以得到以下相同意義的多項式：

$$T_0(x) = 1$$
$$T_1(x) = x$$
$$T_2(x) = 2x^2 - 1$$
$$T_3(x) = 4x^3 - 3x$$
$$T_4(x) = 8x^4 - 8x^2 + 1$$
$$\cdots$$
$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \qquad n \geq 1$$

下圖為 x 值在[-1,1]區間時，$T_0(x)$ 到 $T_5(x)$ 的值：



當 x 值在[-1,1]的區間時 Chebyshev 多項式 $T_n(x)$ 有 n 個根(即有 n 個 x 使得 $T_n(x)=0$)，這些 x 的值如下式所示：

$$x=\cos(\frac{\pi(k-1/2)}{n}) \quad k=1,2,...,n$$

Chebyshev 多項式在向量觀點為兩兩互相垂直，亦即存在著正交(orthogonal)的關係，因此可以用來當成基底以表示一段基頻的軌跡(pitch contour)，以下是其 Chebyshev 多項式彼此之間的關係：

If $x_k(k=1,2,...,m)$ *are* m *zeros of* $T_m(x)$ *and if* $i,j<m$, *then*

$$\sum_{k=1}^{m}T_i(x_k)T_j(x_k)=\begin{cases}0, & i\neq j\\ m/2, & i=j\neq 0\\ m, & i=j=0\end{cases}$$

由【16】我們可以得知以下這個定理：

If $f(x)$ *is an arbitrary function in the interval [-1,1], and if N coefficients* $c_j, j=0,1,...N-1$, *are defined by*

$$\begin{aligned}c_j&=\frac{2}{N}\sum_{k=1}^{N}f(x_k)T_j(x_k)\\&=\frac{2}{N}\sum_{k=1}^{N}f[\cos(\frac{\pi(k-1/2)}{N})]T_j(\cos(\frac{\pi j(k-1/2)}{N}))\end{aligned}$$

Then the approximation formula

$$f(x)\approx\left[\sum_{k=0}^{N-1}c_kT_k(x)\right]-\frac{1}{2}c_0$$

藉由這個定理，我們可以將一段基頻軌跡用 Chebyshev 係數 $c_j, j=0,1,...N-1$ 來表示，經由實驗我們發現 N=6 時即有不錯的效果。

3.3.6 分群法則：K-means

分群法(clustering)【17】通常使用在將資料分類成群的應用上，主要的目的即是將資料分成一群一群(group)，讓相似的資料可以叢聚在一起，並將每個群取一代表點來代表整個群組，如此一來便可達到資料縮減的目的，也可以減輕計算量；在聲調辨識的實驗中我們將採用 K-means 分群法，期許找出每個聲調最有代表性的代表點。

K-means 分群法屬於分割式分群法(partitional clustering)的一種，在演算法一開始即給定分群的叢聚數目，然後藉著自動化的反覆修正，達到分群的目的，K-means 演算法的步驟概述如下：

1. 隨機選取 K 個起始點，分別視為 K 群的群中心
2. 對每一個資料 **x**，找其最接近之群中心，並將 **x** 加入該群，隨即重新計算該群的群中心(該群中原有的資料點加上 **x** 後的平均向量)
3. 對每一個資料點，檢查目前與其最接近的群中心是否和他群組分配一致，如果不是，則回到步驟 2。

3.3.7 分類法則

由本章前面幾節所述，我們將每一個聲調以兩種參數化表示法(Orthogonal Expansion 及 Chebyshev polynomial fitting)表示，因此每一聲調經由訓練之後會有兩組碼本，分別定義為 $CodeBook_O(j), j = 1,...4$ 及 $CodeBook_C(j), j = 1,...4$ ，在聲調分類方面，我們將欲測試的語料求取其特徵參數後與之前訓練出來的碼本進行 1-Nearest Neighbor 的分類，即會得到 $MinDist_O(j), j = 1,...4$ 及 $MinDist_C(j), j = 1,...4$ ，*MinDist* 代表測試語料跟每一聲調代表點中最小的距離，

經由實驗我們可以發現若定義以下分類器結合(classifier combination)的方法，將可達到最佳的分類效果：

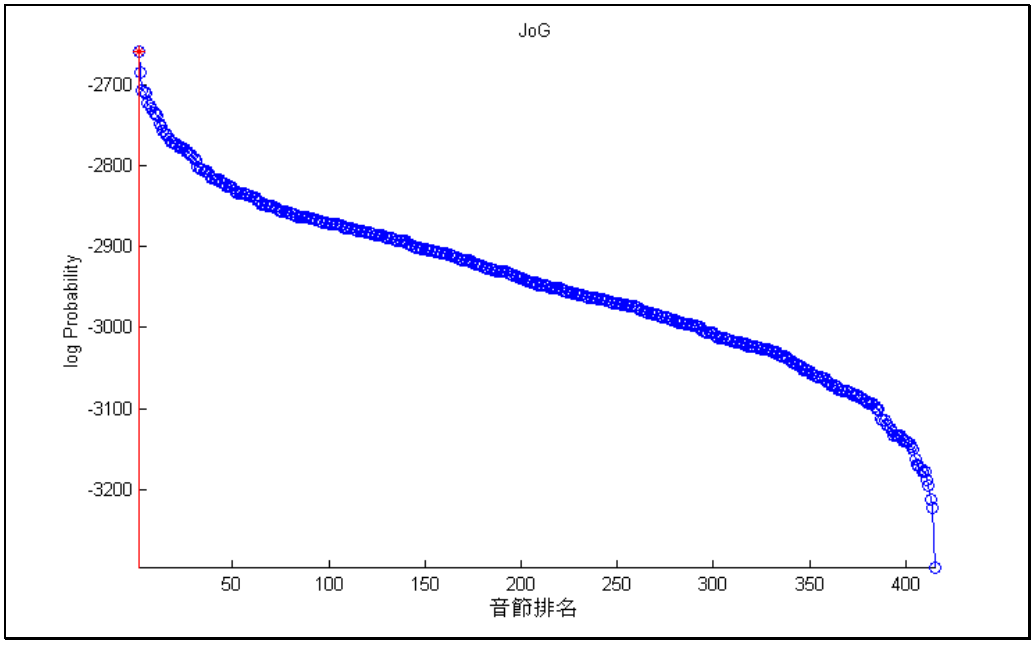
$$
\begin{aligned}
&ComDist(j) = \\
&\quad MinDist_O(j) / \max(MinDist_O) \\
&\quad + MinDist_C(j) / \max(MinDist_C) \quad , j = 1,...4
\end{aligned}
$$

經距離的正規化後加總找其 *ComDist* 距離最小的類別，此類別所屬的聲調即為聲調辨識的結果。

3.4 評分機制

「利用 HMM 及音高資料的評分」評分機制的設計主要是針對聲音及聲調辨識的結果加以評分；由於我們將語音訊號切成一個一個的單音節，所以我們對每個單音節評分，整句語音的分數即是所有單音節分數的平均。

每個音節在聲音辨識方面，我們是利用 Viterbi Search 找出此音節分別跑所有 415 個音節的 Net 的對數機率，即跑第 1 個音節的 Net 之後得到 1 個對數機率、跑第 2 個音節的 Net 之後得到第 2 個對數機率，依此類推，我們將機率排序之後會得到如下圖的機率分佈：



x 軸代表音節依對數機率的排名，從第 1 名到 415 名，y 軸是代表對應的對數機率，由此圖我們可以看出對數機率從第 1 名到第 60 名下降的斜率較急促，第 60 名到第 350 名左右的下降斜率就比較緩慢，350 名之後下降的斜率又急速向下；此對數機率的分佈相當合理，原因如下推論：前面幾十名的音節大多相差一兩個模型(Model)，跟正確音節差異性相較於五六十名之後的音節會較大，因為經由實驗我們發現五六十名之後的音節幾乎所有模型(Model)皆跟正確音節擁有的模型(Model)完全不同。

我們定義在排名兩百名之後的音節由於跟測試語音差異性太大了，所以我們只給 20 分，在第 1 名到第 200 名的我們則依比率給分，評分區間從 100 到 20，第 1 到 200 名的音節圖示如下：

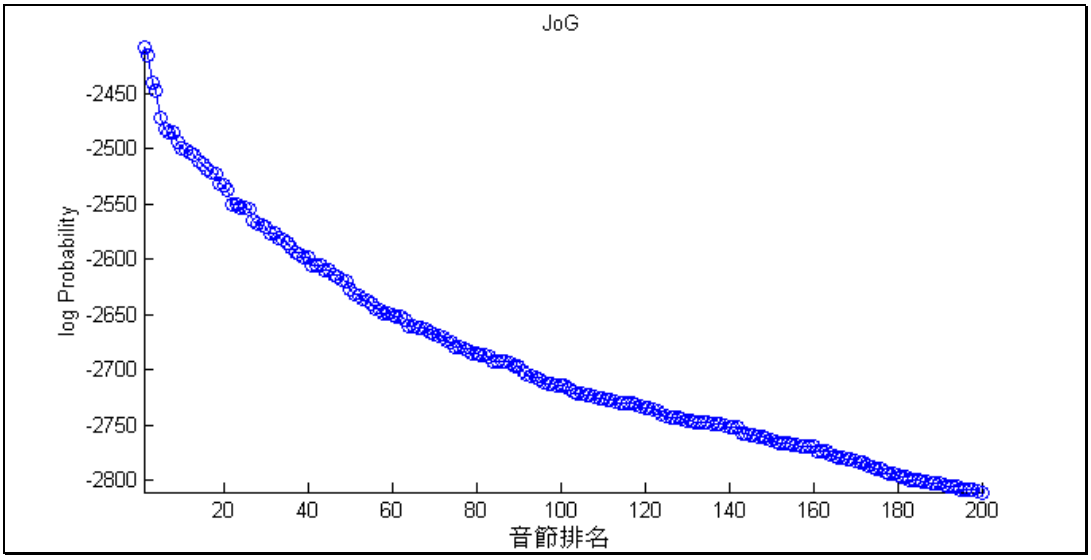


圖 3-11　音節排名與對數機制示意圖

評分示意圖如下：

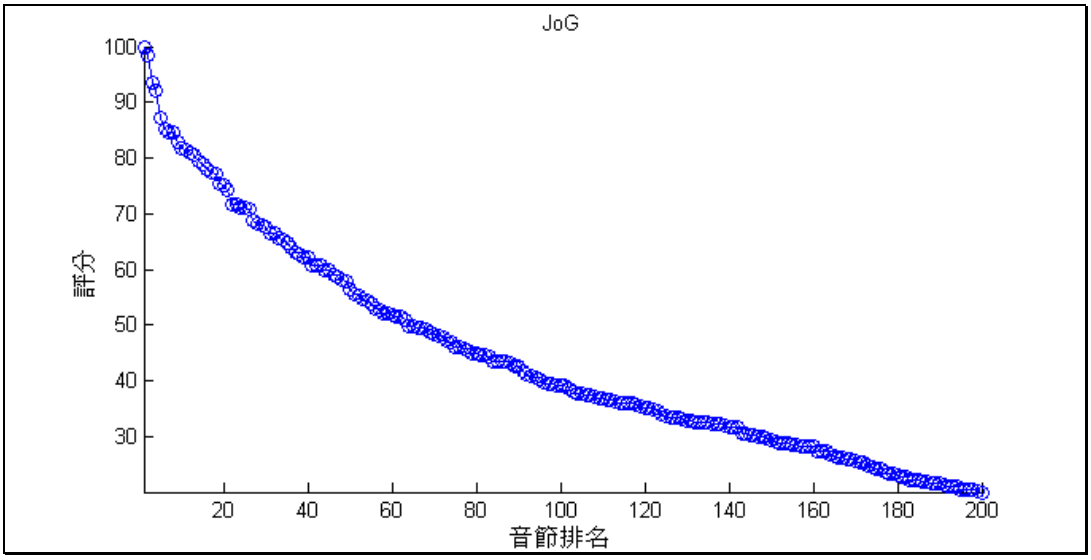


圖 3-12　音節排名與分數關係示意圖

每個音節在聲調辨識方面，我們是將這個音節的基頻軌跡轉成特徵之後再比對聲調模型，若正確聲調的排名落在前兩名的話，即代表此音節的聲調過關；若排名落在後兩名的話，我們就會對此音節的分數扣分。

第4章 結論與展望

本論文的研究主題是「語音評分」，包含「利用標準語音資料的評分」以及「利用 HMM 及音高資料的評分」兩個部分，並經由實作英語評分系統以及唐詩語音評分系統分別應證理論的可行性。

「利用標準語音資料的評分」所運用到的技術，包含特徵參數的擷取、圖樣比對方法的設計以及評分機制的建立三大部分，藉由英語評分系統的設計讓我們可以整合這些技術成為應用；經由實驗我們可以得到三個特徵的權重，分別是 8.5%、16.7%及 74.8%，由此可知梅爾倒頻譜參數代表的重要性最高，其次是基頻軌跡，最後是音量強度曲線。

第二部分「利用 HMM 及音高資料的評分」主要是提供另一種語音評分的方式，以預先訓練好的聲學模型及聲調模型當成標準答案，經由語音辨認技術的使用，找出測試語音跟模型間的差異程度，並配合評分機制給與評分；主要分為「語音辨識」、「聲調辨識」以及「評分機制」包含三個主體； 「語音辨識」以隱藏式馬可夫模型當作聲學模型，以樹狀網路當作語言模型，利用 Viterbi Search 找出最相似的狀態組合，即可辨識出語音的內容； 「聲調辨識」利用 Orthogonal Expansion 及 Chebyshev Approximation 來參數化基頻軌跡，並由 K-means 來訓練聲調模型，藉由聲調模型我們就可以找出測試語音最有可能的聲調； 「評分機制」主要利用語音辨識及聲調辨識的結果，即聲音在所有可能 415 個音裡面的排名以及聲調在四聲中的排名，以這兩個排名配合我們定義好的評分機制即可以針對測試語音在聲音和聲調兩個觀點的評分。

語音評分的運用相當廣泛，例如「利用標準語音資料的評分」由於運算量較小，未來可以運用在現在流行的週邊設備(device)中，例如手機裡的行動 KTV 系統及一般語言學習機的評分系統；「利用 HMM 及音高資料的評分」則可以運用在國語正音系統以及語料收集的篩選系統等等。

附錄一：Orthogonal Expansion 基底推導

由於基頻軌跡可以用多項式 $f(\frac{i}{N})=\sum_{j=0}^{3}a_j(\frac{i}{N})^j$， $i=0,1,...,N$ 來表示，因此我們定義其內積值(Inner Product)為

$$< f_a(\frac{i}{N}), f_b(\frac{i}{N}) >= \frac{1}{N+1}\sum_{i=0}^{N} f_a(\frac{i}{N})\cdot f_b(\frac{i}{N}) \quad , \quad N>2 \ ,$$

並利用 Gram-Schmidt Orthogonalization procedure 即可將 $\{1,\frac{i}{N},(\frac{i}{N})^2,(\frac{i}{N})^3\}$ 轉成

$$\{\phi_0(\frac{i}{N}),\phi_1(\frac{i}{N}),\phi_2(\frac{i}{N}),\phi_3(\frac{i}{N})\}$$

$$\phi_0=1$$

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

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

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

證明：

$$(\vec{x}_0,\vec{x}_1,\vec{x}_2,\vec{x}_3)=(1,\frac{i}{N},(\frac{i}{N})^2,(\frac{i}{N})^3)$$

$$\vec{y}_0=\vec{x}_0=1$$

$$ps.\ \|\vec{y}_0\|=\frac{1}{N+1}\sum_{i=0}^{N}1\cdot 1=\frac{N+1}{N+1}=1$$

$$\vec{y}_1=\vec{x}_1-\frac{<\vec{x}_1,\vec{y}_0>}{\|\vec{y}_0\|^2}\vec{y}_0=\frac{i}{N}-\frac{<\frac{i}{N},1>}{1}\cdot 1=\frac{i}{N}-\frac{1}{2}$$

$$ps.<\frac{i}{N},1>=\frac{1}{N+1}\sum_{i=0}^{N}\frac{i}{N}\cdot 1=\frac{1}{N+1}\cdot\frac{1}{N}\cdot\frac{N(N+1)}{2}=\frac{1}{2}$$

$$\vec{y}_2 = \vec{x}_2 - \frac{<\vec{x}_2, \vec{y}_0>}{\|\vec{y}_0\|^2}\vec{y}_0 - \frac{<\vec{x}_2, \vec{y}_1>}{\|\vec{y}_1\|^2}\vec{y}_1$$

$$= (\frac{i}{N})^2 - \frac{<(\frac{i}{N})^2, 1>}{1}\cdot 1 - \frac{<(\frac{i}{N})^2, \frac{i}{N}-\frac{1}{2}>}{\left\|\frac{i}{N}-\frac{1}{2}\right\|^2}\cdot(\frac{i}{N}-\frac{1}{2})$$

$$= (\frac{i}{N})^2 - \frac{2N+1}{6N} - (\frac{i}{N}-\frac{1}{2})$$

$$= (\frac{i}{N})^2 - \frac{i}{N} + \frac{N-1}{6N}$$

ps.

$$<(\frac{i}{N})^2, 1>= \frac{1}{N+1}\sum_{i=0}^{N}(\frac{i}{N})^2\cdot 1 = \frac{1}{N+1}\cdot\frac{1}{N^2}\cdot\frac{N(N+1)(2N+1)}{6} = \frac{2N+1}{6N}$$

$$<(\frac{i}{N})^2, \frac{i}{N}-\frac{1}{2}>= \frac{1}{N+1}\sum_{i=0}^{N}(\frac{i}{N})^2\cdot(\frac{i}{N}-\frac{1}{2})$$

$$= \frac{1}{N+1}\cdot\frac{1}{N^3}\cdot\frac{N^2(N+1)^2}{4} - \frac{1}{N+1}\cdot\frac{1}{2N^2}\cdot\frac{N(N+1)(2N+1)}{6}$$

$$= \frac{N+1}{4N} - \frac{2N+1}{12N}$$

$$= \frac{N+2}{12N}$$

$$\left\|\frac{i}{N}-\frac{1}{2}\right\|^2 =< \frac{i}{N}-\frac{1}{2}, \frac{i}{N}-\frac{1}{2}>= \frac{1}{N+1}\sum_{i=0}^{N}(\frac{i}{N}-\frac{1}{2})^2$$

$$= \frac{1}{N+1}\sum_{i=0}^{N}[(\frac{i}{N})^2 - \frac{i}{N} + \frac{1}{4}]$$

$$= \frac{1}{N+1}\cdot\{\frac{1}{N^2}\cdot\frac{N(N+1)(2N+1)}{6} - \frac{1}{N}\cdot\frac{N(N+1)}{2} + \frac{N+1}{4}\}$$

$$= \frac{2N+1}{6N} - \frac{1}{2} + \frac{1}{4}$$

$$= \frac{4N+2-3N}{12N}$$

$$= \frac{N+2}{12N}$$

同理，我們可以推導 $\vec{y}_3$

$$\vec{y}_3 = \vec{x}_3 - \frac{<\vec{x}_3, \vec{y}_0>}{\|\vec{y}_0\|^2}\vec{y}_0 - \frac{<\vec{x}_3, \vec{y}_1>}{\|\vec{y}_1\|^2}\vec{y}_1 - \frac{<\vec{x}_3, \vec{y}_2>}{\|\vec{y}_2\|^2}\vec{y}_2$$

$$= (\frac{i}{N})^3 - \frac{<(\frac{i}{N})^3, 1>}{1}\cdot 1 - \frac{<(\frac{i}{N})^3, \frac{i}{N}-\frac{1}{2}>}{\left\|\frac{i}{N}-\frac{1}{2}\right\|^2}\cdot(\frac{i}{N}-\frac{1}{2})$$

$$- \frac{<(\frac{i}{N})^3, (\frac{i}{N})^2 - \frac{i}{N} + \frac{N-1}{6N}>}{\left\|(\frac{i}{N})^2 - \frac{i}{N} + \frac{N-1}{6N}\right\|^2}\cdot((\frac{i}{N})^2 - \frac{i}{N} + \frac{N-1}{6N})$$

$$= (\frac{i}{N})^3 - \frac{3}{2}(\frac{i}{N})^2 + \frac{6N^2-3N+2}{10N^2}(\frac{i}{N}) - \frac{(N-1)(N-2)}{20N^2}$$

$$\phi_0 = \frac{\vec{y}_0}{\|\vec{y}_0\|} = \frac{1}{1} = 1$$

$$\phi_1 = \frac{\vec{y}_1}{\|\vec{y}_1\|} = \frac{\frac{i}{N}-\frac{1}{2}}{\left\|\frac{i}{N}-\frac{1}{2}\right\|} = (\frac{12N}{N+2})^{1/2}\cdot(\frac{i}{N}-\frac{1}{2})$$

$$\phi_2 = \frac{\vec{y}_2}{\|\vec{y}_2\|} = \frac{(\frac{i}{N})^2 - \frac{i}{N} + \frac{N-1}{6N}}{\left\|(\frac{i}{N})^2 - \frac{i}{N} + \frac{N-1}{6N}\right\|}$$

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

ps.

$$\left\|(\frac{i}{N})^2 - \frac{i}{N} + \frac{N-1}{6N}\right\|^2 = <(\frac{i}{N})^2 - \frac{i}{N} + \frac{N-1}{6N}, (\frac{i}{N})^2 - \frac{i}{N} + \frac{N-1}{6N}>$$

$$= \frac{1}{N+1}\sum_{i=0}^{N}[(\frac{i}{N})^2 - \frac{i}{N} + \frac{N-1}{6N}]^2$$

$$= \frac{(N-1)(N+2)(N+3)}{180N^3}$$

$$\phi_3 = \frac{\vec{y}_3}{\|\vec{y}_3\|} = \frac{(\frac{i}{N})^3 - \frac{3}{2}(\frac{i}{N})^2 + \frac{6N^2-3N+2}{10N^2}(\frac{i}{N}) - \frac{(N-1)(N-2)}{20N^2}}{\left\|(\frac{i}{N})^3 - \frac{3}{2}(\frac{i}{N})^2 + \frac{6N^2-3N+2}{10N^2}(\frac{i}{N}) - \frac{(N-1)(N-2)}{20N^2}\right\|}$$

$$= [\frac{2800N^5}{(N-1)(N-2)(N+2)(N+3)(N+4)}]^{1/2}$$

$$\cdot [(\frac{i}{N})^3 - \frac{3}{2}(\frac{i}{N})^2 + \frac{6N^2-3N+2}{10N^2}(\frac{i}{N}) - \frac{(N-1)(N-2)}{20N^2}]$$

ps.

$$\left\|(\frac{i}{N})^3 - \frac{3}{2}(\frac{i}{N})^2 + \frac{6N^2-3N+2}{10N^2}(\frac{i}{N}) - \frac{(N-1)(N-2)}{20N^2}\right\|^2$$

$$= \frac{(N-1)(N-2)(N+2)(N+3)(N+4)}{2800N^5}$$

備註：

$$\sum_{i=0}^{N} i = \frac{N(N+1)}{2}$$

$$\sum_{i=0}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}$$

$$\sum_{i=0}^{N} i^3 = \frac{N^2(N+1)^2}{4}$$

$$\sum_{i=0}^{N} i^4 = \frac{N(N+1)(2N+1)(3N^2+3N-1)}{30}$$

$$\sum_{i=0}^{N} i^5 = \frac{N^2(N+1)^2(2N^2+2N-1)}{12}$$

$$\sum_{i=0}^{N} i^6 = \frac{N(N+1)(2N+1)(3N^4+6N^3-3N+1)}{42}$$

參考文獻

【1】J.D., J.G.P, J.H.L.H, *Discrete-Time Processing of Speech Signals*, Prentice Hall, 1993.

【2】T.W. Parsons, *Voice and Speech Processing*, McGraw-Hill, 1986.

【3】葉佳慧,“以音符及節拍為主的音樂檢索系統”, 清華大學碩士論文, 民國 90 年

【4】Lawrence Rabiner, B.H Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.

【5】J.-S. R. Jang, C.-T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing*, Prentice Hall, 1996.

【6】高名揚,“以聲音內容為主的音樂資料庫檢索系統的加速方法”, 清華大學碩士論文, 民國 90 年

【7】方士豪,“雜訊及通道環境下語音辨認技術之研究”, 台灣大學碩士論文, 民國 90 年

【8】JULIUS T. TOU, RAFAEL C. GONZALEZ, *Pattern Recognition Principles*, Addison-Wesley Publishing Company, 1974.

【9】呂道誠,“不特定語者、國台雙語大詞彙語音辨識之聲學模型研究”,長庚大學碩士論文,民國 90 年

【10】楊永泰,“隱藏式馬可夫模型應用於中文語音辨識之研究”, 中原大學碩士論文, 民國 89 年

【11】陳柏琳,“中文語音資訊檢索—以音節為基礎之索引特徵、統計式檢索模型及進一步技術”, 台灣大學博士論文, 民國 90 年

【12】Steven Young, *The HTK Book version 3*, Microsoft Corporation, 2000.

【13】莊向凱,“國語語音資料庫之標音系統”, 清華大學碩士論文, 民國 88 年

【14】徐光輝,“國語語音資料庫 MAT-2000 上的聲調辨認研究”，清華大學碩士論文，民國 89 年

【15】王逸如,“對基週軌跡做向量量化之線性預估語音編碼”，交通大學碩士論文，民國 76 年

【16】Press, William H., *Numerical Recipes in C, The Art of Scientific Computing*, Cambridge University Press, 1992.

【17】許文豪,“圖形辨識概述與實作”，清華大學碩士論文，民國 89 年

【18】蔣昇倫,“經電話通道之國語連續 411 音節辨認”，交通大學碩士論文，民國 86 年

【19】謝宏坤,“語音說明中搜尋任意定義之關鍵詞的研究”，台灣科技大學碩士論文，民國 89 年