第4章　英文語音評分

本章將介紹「英文語音評分」。在本論文中我們利用標準語音資料來進行一種較為主觀的評分方式，亦即存在一個標準語音，而評分的語音若和此標準語音的相似度愈高，則得到的分數也將愈高。

實作上我們主要使用圖樣比對(Pattern Matching)的方法，將評分的語音和標準語音資料逐音素地來做比較，以期找出評分語音和標準語音的差異程度，並藉此對評分語音進行評分。

4.1 評分系統簡介

本評分系統藉由語音訊號切割將標準語音切割出時間與發音的對應後，使用者再透過電腦麥克風錄製欲和標準語音評分的語音檔，語音長度為 5 秒、音訊格式為 PCM、音訊取樣頻率為 16 kHz、位元解析度為 16 bits、位元率為 256 kbps。經由說話驗證和語音訊號切割後，即可開始進行評分。

在這一章的英文語音評分又可細分為三大部份：

第一部份為特徵參數擷取(Feature Extraction)，在本論文中我們採用以下四個特徵，分別是音量強度曲線、基頻軌跡曲線、發聲急緩變化以及 HMM 對數機率差異。音量強度曲線代表聲音音量大小的變化趨勢；基頻軌跡曲線代表聲音音高

的起伏；發聲急緩變化代表說話的快慢節奏；HMM 對數機率差異則是使用語音相對於聲學模型的對數機率來代表說話發聲的差異性。經由特徵擷取後我們會針對特徵參數作正規化的動作。

第二部份為圖樣比對方法的設計，將評分的語音和標準語音資料逐音素地來做比較，以期找出評分語音和標準語音之間的差異程度。

第三部份則是評分機制的建立，經由我們設計的評分機制來對評分語料及標準語音之間的相似度評分。上述的三個部份將會在本章的各小節中逐一介紹。圖 4-1 為評分系統流程圖：

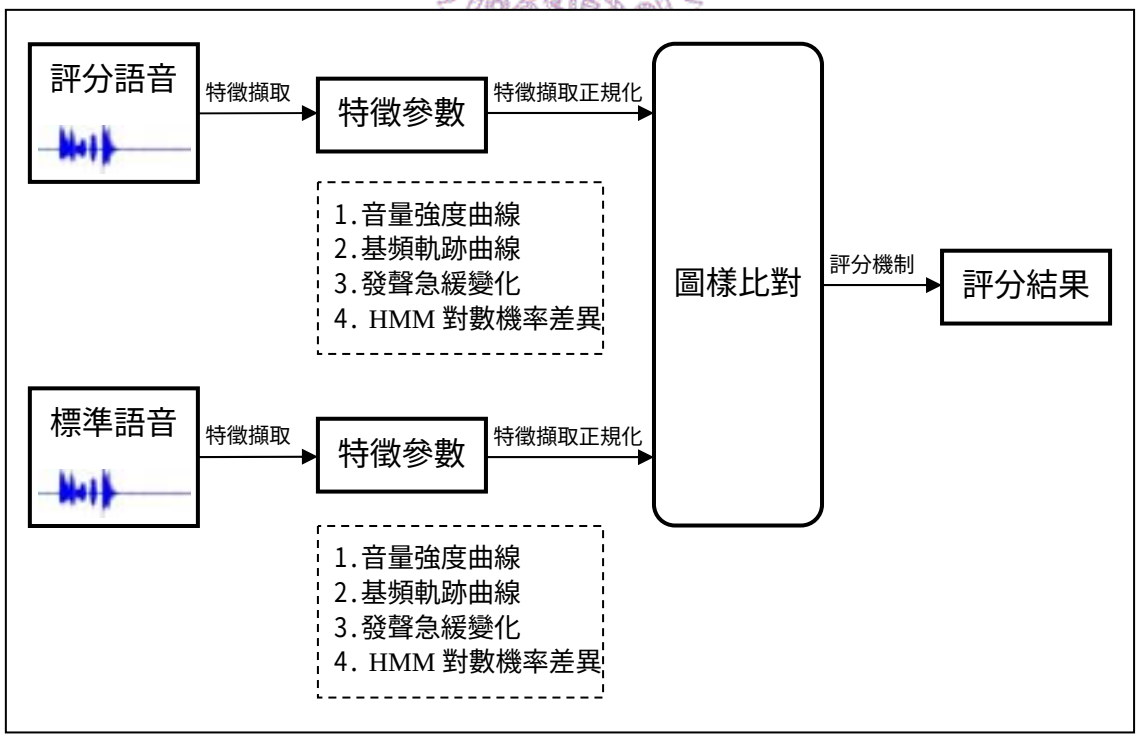

圖 4-1　評分系統流程圖

4.2 特徵參數擷取

聲音訊號是一種時變性(Time Varying)的訊號，其波形變化非常快速的，但是如果我們將觀察聲音訊號的時間單位縮小，此時可以發現，其波形變化反而相當緩慢，關於這種現象，我們稱其具有「短時間穩定」(Short Time Stationary)的性質【8】。有鑑於此，我們將觀察的單位稱為一個「音框」(Frame)，因而我們可以對聲音訊號以切割音框(Taking Frame)的方法進行各種特徵參數的擷取。

這一小節將介紹我們在評分系統所用到的四個特徵參數，分別是音量強度曲線、基頻軌跡曲線、發聲急緩變化及 HMM 對數機率差異。

4.2.1 音量強度曲線

我們將取樣頻率為 16kHz 的語音訊號取音框化，音框大小為 512 點，約 32 毫秒，重疊(Overlap)部份為 170 點，約 10 毫秒，占每一個音框的三分之一，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：

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

4.2.2 基頻軌跡曲線

前人的研究提到許多關於求取基頻軌跡(Pitch Tracking)的技術，如 Wavelet、Autocorrelation【9】、AMDF(Average Magnitude Difference Function)【5】【8】等，而我們採用 AMDF 的方法，圖 4-2 為求取基頻軌跡曲線的流程圖：

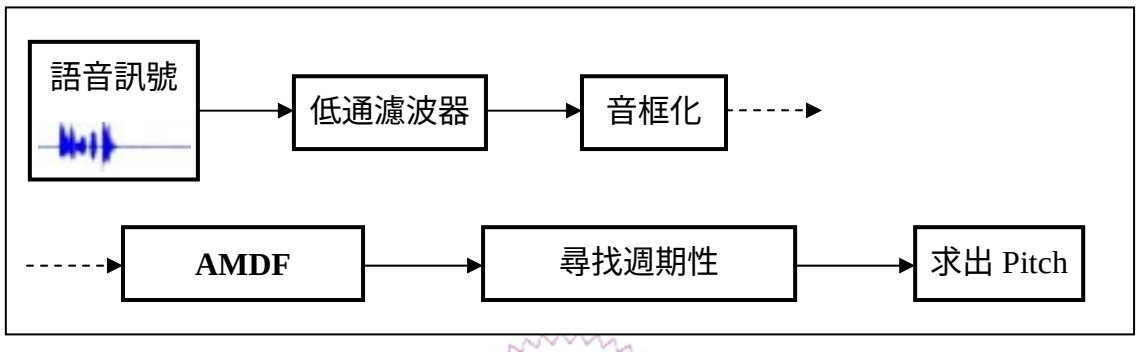

圖 4-2 基頻軌跡曲線擷取流程圖

我們在此簡介各步驟：

1. 低通濾波器 (Low-Pass Filter)

對於錄製好的語音訊號，為了除去高頻的雜訊和爆音，所以我們將語音訊號通過低通濾波器，希望能夠濾掉較高頻率的雜訊，而留下較低頻率的語音訊號。

2. 音框化 (Frame Blocking)

我們以 512 點當做一個音框的大小，另外為了避免音框間的變動過於劇烈，因此取音框時在兩兩音框間重疊 170 點。

3. AMDF

接著對每個音框做 AMDF，找出相似波型重覆出現的週期：

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

其中 M 是音框大小，τ 是平移量。

4. 尋找週期性並求出 Pitch

經過 AMDF 後，我們對此訊號尋找其週期性，即所謂的 Local Minimum，而後計算兩 Local Minimum 之間於時間軸上的距離就是聲音的週期，將其取倒數即為基頻。

將每個音框重覆步驟 3 及 4 之後即可得到整個語音訊號的基頻軌跡曲線。圖 4-3 為音量強度曲線及基頻軌跡曲線示意圖，上方為語音訊號；中間為音量強度曲線，下方為基頻軌跡曲線。

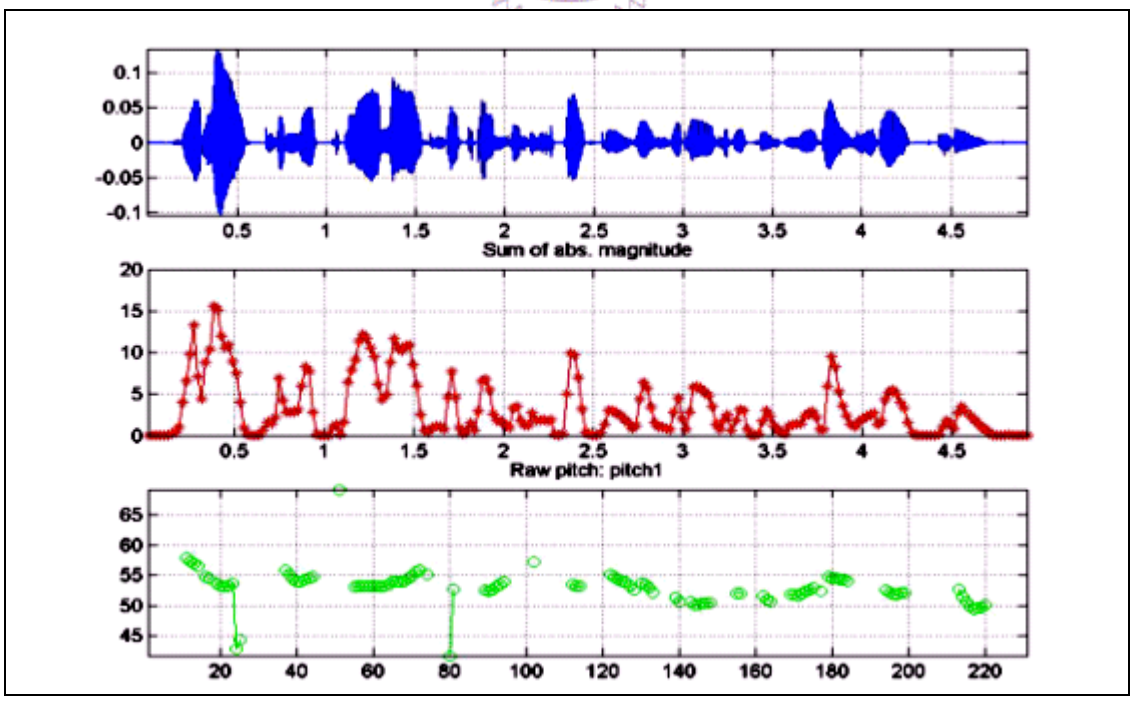

圖 4-3　音量強度曲線及基頻軌跡曲線示意圖

4.2.3 發聲急緩變化

發聲急緩變化代表說話的快慢、節奏【10】，我們使用 Forced Alignment 的方式對取樣頻率為 16 kHz、位元解析度為 16 bits 的語音訊號進行切割動作，即可得到每個音素的時間點。

上一段所提到的 Forced Alignment 就是將語言模型(HMM)限定為語音訊號所對應的文字。在實作上我們先於文件檔中寫下語音訊號的文字內容，在本評分系統中只有一句英文句子，接著再對此文件檔案進行標音，並建立辨識網路，如此一來產生的語言模型也就只有之前所寫入的英文句子。接下來我們強迫讓語音訊號與文字結果作對應(Match)，在透過維特比演算法之後就能將語音訊號的狀態序列依我們預先訂定的語言模型走出唯一的結果。圖 4-4 為擷取發聲急緩變化此項特徵的流程圖：

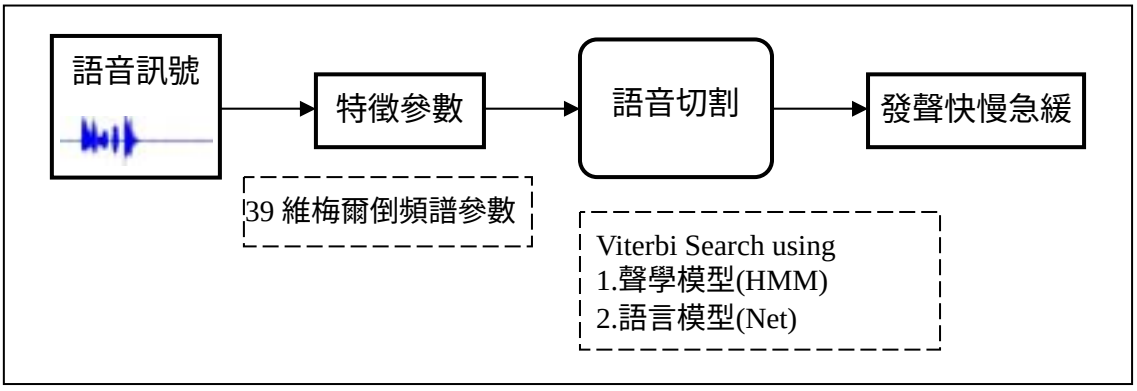

圖 4-4　發聲急緩變化擷取流程圖

4.2.4 HMM 對數機率差異

在透過維特比演算法求得發聲急緩變化這項特徵參數的同時，我們也可以得到每個音素對應於聲學模型的對數機率(HMM Log Probability)【10】【11】，而這個對數機率就是我們所說的 HMM 對數機率差異。

不過在細節上，HMM 對數機率差異和發聲急緩變化這兩項特徵參數擷取的過程還是有些許的不同，後者是使用兩種不同的聲學模型對標準語音及評分語音進行切割的動作。而在 HMM 對數機率差異這部份，由於經計算而得到的對數機率值是相對於所使用的聲學模型，因此為了讓兩個欲比較的語音有相同的比對基準，在這一部份我們針對兩個語音訊號使用相同的聲學模型來辨識以計算它們的對數機率。在第三章我們曾經提到，對語音訊號而言，由於使用 Native-Speaker 的聲學模型可以得到較高的辨識率，而高辨識率也會使音素的對數機率值較為準確，因此為了求取 HMM 對數機率差異這項特徵參數，我們是以 Native-Speaker 的聲學模型當作比較的基準。圖 4-5 為 HMM 對數機率差異擷取流程圖：

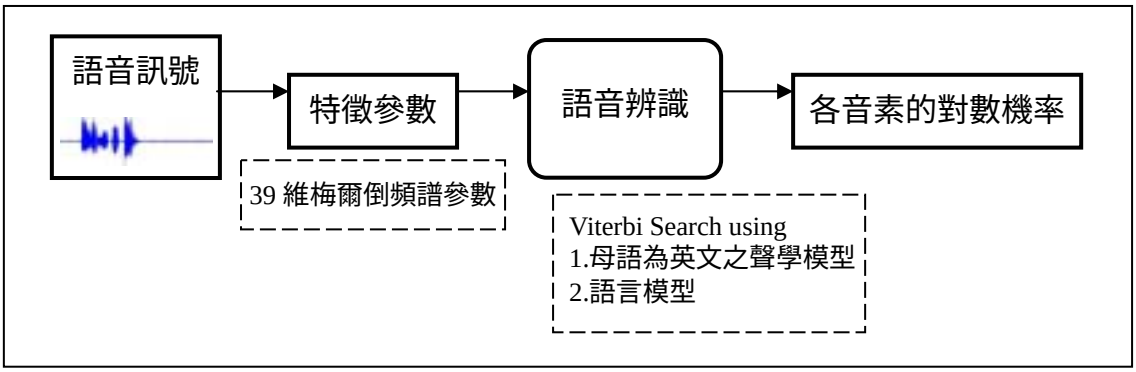

圖 4-5　HMM 對數機率差異擷取流程圖

4.3 特徵參數正規化

在特徵參數擷取的過程中，由於每個人聲調的高低或是錄音環境的差異，造成特徵參數或多或少受到了一些影響。在這一小節，我們提出了三個方法：內插法(Interpolation)、線性縮放(Linear Scaling)、線性平移(Linear Shifting)，以期能將特徵參數正規化。

4.3.1 內插法

由於經過語音訊號切割後評分語音和標準語音的音素長度不一定一樣，因此我們使用了一維內插法，可以在解析度盡量不失真的前提下，將音量強度曲線及基頻軌跡這兩個特徵參數作伸長或縮短的動作，以求與標準語音的音素時間長度一致，此方法可以有效解決特徵參數長短不一的問題。

4.3.2 線性縮放

即使是同一個人用同樣的音量對著麥克風錄音，還是有可能因為麥克風的不同而造成語音訊號的音量大小有所差異。我們定義標準語音的音量強度曲線為 $aveMag_1(n)$，評分語音的音量強度曲線為 $aveMag_2(n)$，另外也假設錄音環境的差異存在一倍數 θ 的關係，則我們希望能找出一個倍數讓兩曲線的誤差 $\bar{e}$ 愈小愈好，

假設

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

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

兩曲線存在 $\vec{B}\theta + \vec{e} = \vec{A}$ 的關係，由 Least-Squares Estimator【12】我們可以得知以下的結果：

$$\theta = \left(\vec{B}^T \vec{B}\right)^{-1} \vec{B}^T \vec{A}$$

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：

$$aveMag_2^{'}(n) = \vec{B}\theta = aveMag_2(n) \cdot \theta, \qquad n = 0,1,...,N-1$$

4.3.3 線性平移

每個人的聲調高低不盡相同，一般而言平均女性的聲調頻率在 200 Hz 左右，略高於平均男性的 150 Hz，由於我們的英文語音評分系統在聲調這部份著重於句子的重音、抑揚頓挫、以及音調起伏的趨勢，因此有必要對聲調高低的差異作一平移(Shifting)【13】，以解決基頻軌跡的差異。

我們假設標準語音及評分語音的基頻軌跡分別以 $f_1(x)$、$f_2(x)$ 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 $f_2^{'}(x)$，其公式為：

$$diff = \left(\frac{1}{N} \sum_{k=0}^{N-1} f_2(k) - \frac{1}{N} \sum_{k=0}^{N-1} f_1(k) \right)$$

$$f_2^{'}(x) = f_2(x) - diff$$

4.4 圖樣比對方法設計

本節將介紹如何針對我們所使用的四個特徵參數：音量強度曲線、基頻軌跡曲線、發聲急緩變化、HMM 對數機率差異，來設計圖樣比對【14】的方法，以找出標準語音和評分語音之間的差異程度。

4.4.1 音量強度曲線比對方法

音量強度曲線的比對方法如圖 4-6 所示，假設標準語音的音量強度曲線為 Mag_1，評分語音的音量強度曲線為 Mag_2，則我們以 Mag_1 的音素長度為基準，使用內插法調整 Mag_2 中每個音素的長度，以解決兩特徵長度不一的問題。我們使用音素的長度當成基準的原因在於比較兩個語音差異時是對每個音素來做比較。

接下來再經由線性縮放解決麥克風的差異性，即可將兩兩長度相同的音量特徵相互比較，得到其差異程度。

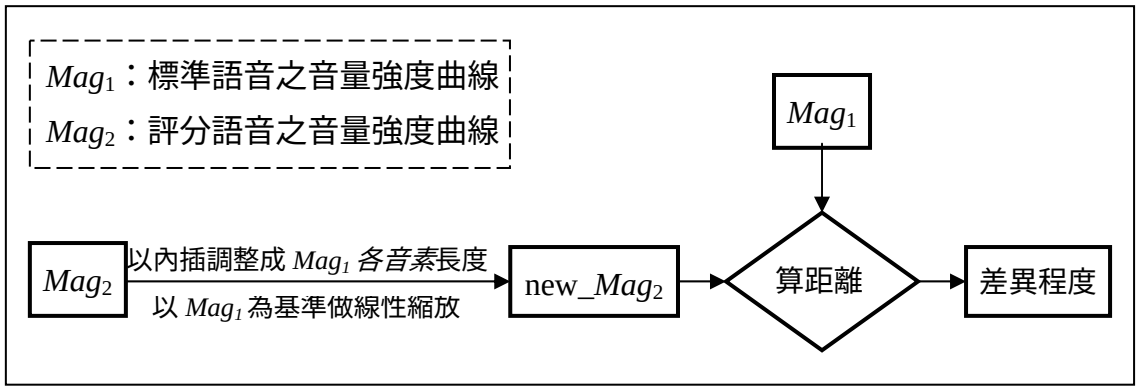

圖 4-6　音量強度曲線比對流程圖

4.4.2 基頻軌跡曲線比對方法

基頻軌跡曲線的比對方法如圖 4-7 所示，假設標準語音的基頻軌跡曲線為 Pit_1，評分語音的基頻軌跡曲線為 Pit_2，則我們以 Pit_1 的音素長度為基準，使用內插法調整 Pit_2 中每個音素的長度，以解決兩特徵長度不一的問題。

接下來再經由線性平移解決每個人音高不一致的差異後，即可將兩兩長度相同的基頻軌跡特徵相互比較，得到其差異程度。

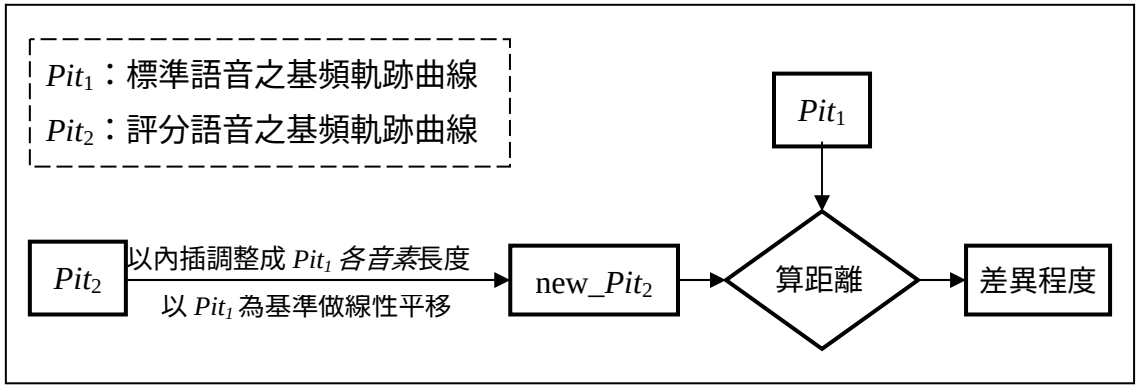

圖 4-7　基頻軌跡曲線比對流程圖

4.4.3 發聲急緩變化比對方法

圖 4-8 為發聲急緩變化的圖樣比對流程圖。此項特徵參數的圖樣比對方法較為單純，在經過 Forced Alignment 的語音訊號切割之後，我們可以得到各個音素的時間區段，接下來不必經由特徵參數正規化的步驟，即可對每個音素作直接的比對而得到差異程度。

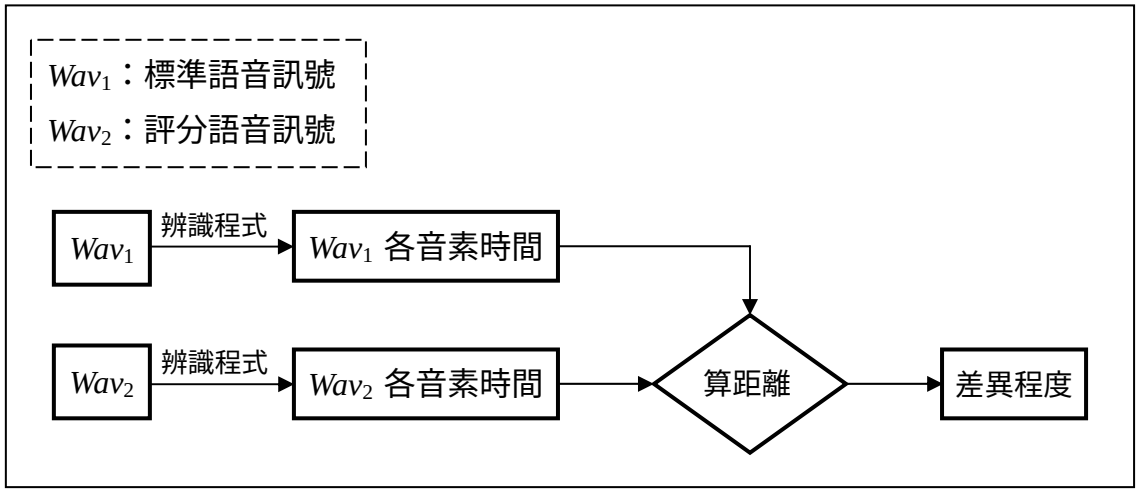

圖 4-8　發聲急緩變化比對流程圖

4.4.4 HMM 對數機率差異比對方法

對於 HMM 對數機率差異，我們先以 N-HMM(HMM trained from Native Speaker) 求出標準語音訊號及評分語音訊號中每個音素的對數機率，若對數機率值愈大，表示該音素的發音愈接近聲學模型。圖 4-9 為 HMM 對數機率差異比對的流程圖：

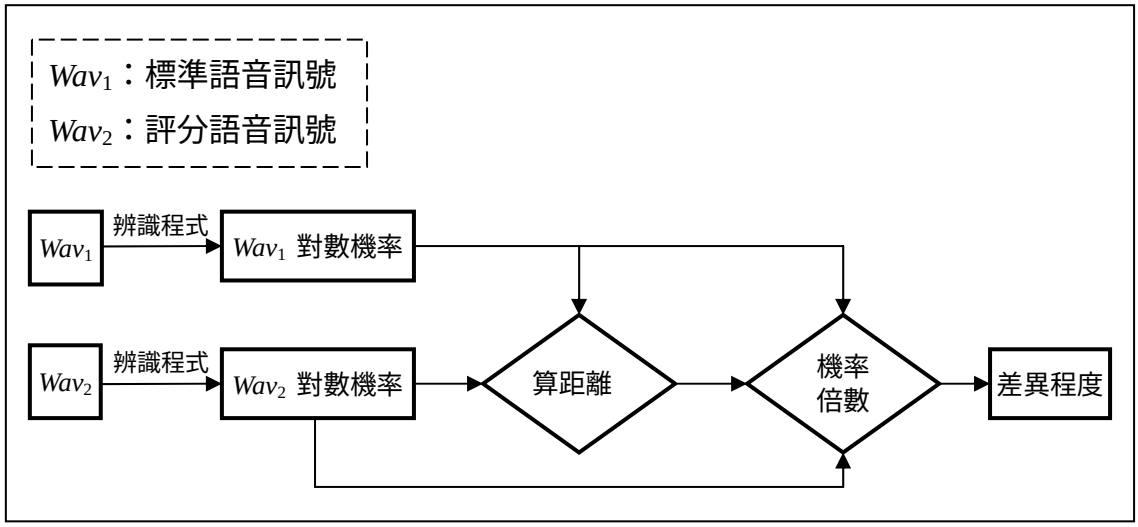

圖 4-9　HMM 對數機率差異比對流程圖

由於對數機率是相對的，因此如何比較兩個語音訊號 HMM 對數機率差異的相對值，就是我們這一小節所探討的重點。我們可以將標準語音的對數機率($Stand_{\log}$)和評分語音的對數機率($Evalu_{\log}$)分為四種情況來討論：

1. $Stand_{\log}$ 大，$Evalu_{\log}$ 大

兩語料的對數機率皆接近於聲學模型的發音，表示此兩語料的發音也極為相似，因此發音的差異小。

2. $Stand_{\log}$ 大，$Evalu_{\log}$ 小

其中標準語料的發音接近聲學模型，而評分語料的發音和聲學模型較不相似。兩語料的 HMM 對數機率差異特徵和對數機率的差異值成正比。

3. $Stand_{\log}$ 小，$Evalu_{\log}$ 大

和第 2 點類似，評分語料的發音接近聲學模型，而標準語料的發音和聲學模型較不相似。兩語料的 HMM 對數機率差異特徵也和對數機率的差異值成正比。

4. $Stand_{\log}$ 小，$Evalu_{\log}$ 小

此為較複雜的情況，兩個語音訊號的對數機率都小，表示這兩個語音的發音都不接近聲學模型。我們將語音訊號簡化為一維空間來探討：首先假設聲學模型為一維空間中的 0 點，如果兩語音相對於聲學模型是處於同方向的狀態，此時兩語音可視其同為正號或同為負號，因此兩語音的發音有可能是很接近的；反之若兩語音相對於聲學模型是處於反方向，則表示兩語音的發音相差很大。

但實際上我們採用 39 維梅爾倒頻譜參數當作辨識程式的特徵參數，而不是上一段所舉例的一維空間，因此若兩個對數機率值都很小的情況下，這兩個語音訊號的發音相近的可能性非常低。

由於上述的四種狀況，我們發現從標準語音和評分語音的對數機率差異值並沒有辦法決定其發音的相近與否，因此我們設計了機率倍數來修正對數機率的差異值。圖 4-10 為機率倍數的示意圖，當兩語音的對數機率絕對值皆小於 1050 時，機率倍數的變化趨勢較小；當兩語音的對數機率絕對值皆大於 1050 時，機率倍數的變化趨勢較大。

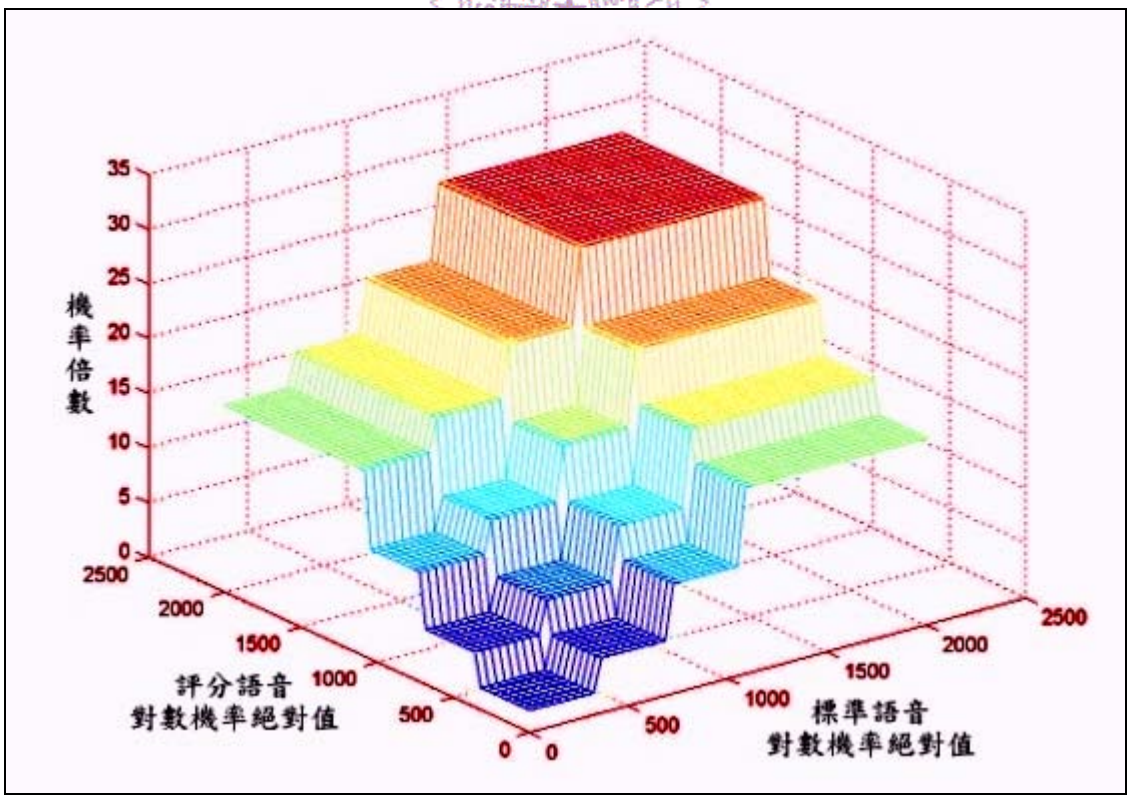

圖 4-10　機率倍數示意圖

關於機率倍數我們定義以下的公式：

$$Const = \begin{cases} \left\lceil \dfrac{|\log P|}{350} \right\rceil, & 0 \le |\log P| \le 1050 \\ 3 + \min\left(1, \left\lceil \dfrac{|\log P|}{1400} \right\rceil\right), & |\log P| > 1050 \end{cases}$$

$$factor_P = (Const_{stard})^2 + (Const_{Evaul})^2$$

$\log P$ 表示語音的對數機率。當算出標準語音和評分語音的 $Const$ 值後，再經由平方相加即可得到機率倍數 $factor_P$，將此機率倍數乘上兩語音訊號對數機率的差距就是我們發音特徵的差異程度。

4.5 評分機制建立

這一小節將介紹英文語音評分的評分機制。經由圖樣比對後，我們可以計算出標準語音和評分語音之間每個音素的差異程度，再藉由特徵(Feature)、音素、單字(Word)和句子(Sentence)等的評分機制，就可以得到最後評分的結果，細節將於以下各小節逐一介紹。

4.5.1 評分機制 － 特徵

對於每個音素中特徵參數的分數，我們設定以下的公式【15】：

$$score_{fea} = \frac{100}{1 + a \cdot (dist)^{b}}$$

由這個公式我們就可以將兩音素間某個特徵的差異程度轉成 0 到 100 之間的分數，只要設定好兩組的 $dist$ 及對應的 $score_{fea}$，即可從中求出 a 和 b，接著所有的距離也將可以計算出對應的分數。

4.5.2 評分機制 － 音素

當計算出每個音素中四項特徵參數的分數後，利用四項特徵對於英文語音評分系統所占的權重加總後即可得到每個音素的分數。以下是設定的公式：

$$score_{pho} = w_1 \cdot score_{fea_1} + w_2 \cdot score_{fea_2} + w_3 \cdot score_{fea_3} + w_4 \cdot score_{fea_4}$$

w_1、w_2、w_3、w_4 分別代表四個特徵參數的權重。經由下一節的實驗，我們可以求出這四項權重，也可以由權重的比例得知四項特徵參數對於英文評分的重要性。

4.5.3 評分機制 － 單字

得知每個音素的得分後，以每個音素占單字的時間為權重，即可求出句子中每一個單字的分數，以下為設定的公式：

$$score_{word} = \sum_{n=1}^{N} \frac{len(pho_n)}{len(word)} \cdot score_{pho_n}$$

N 為一單字中評分音素的數量；$len(x)$ 表示 x 的時間長度。

4.5.4 評分機制 － 句子

由於單字的時間長短會影響人耳對於一句話的關注點，因此我們也是以單字的時間為權重來計算出一句語音訊號最後得到的分數。以下為定義的公式：

$$score_{sen} = \sum_{n=1}^{N} \frac{len(word_n)}{len(sentence)} \cdot score_{word_n}$$

其中 N 表示句子中單字的總數；$len(x)$ 表示 x 的時間長度。

4.6 英文語音評分實驗結果

我們得到四個特徵參數中各音素的差異程度，並依所佔的比例求出一個句子的平均差異程度後，即可代入以下的公式：

$$score = w_1 \cdot \frac{100}{1 + a_1 \cdot (dist_1)^{b_1}} + w_2 \cdot \frac{100}{1 + a_2 \cdot (dist_2)^{b_2}} + w_3 \cdot \frac{100}{1 + a_3 \cdot (dist_3)^{b_3}} + w_4 \cdot \frac{100}{1 + a_4 \cdot (dist_4)^{b_4}}$$

$a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4$ 為差異程度轉成分數的參數；w_1, w_2, w_3, w_4 為四個特徵參數的權重；而 $dist_1, dist_2, dist_3, dist_4$ 表示標準語音和評分語音訊號經由圖樣比對後得到四項特徵參數的差異程度。為了求得以上各個參數的值，我們設計了以下的實驗：

在語料的部份我們收集 200 組語料，每一組的語料分別包括一句標準語音和一句評分語音，其中標準語音的總容量為 24.7 Megabytes、所有語料長度總和約為 12 分 51 秒，評分語音的總容量為 35.8 Megabytes、所有語料長度總和約為 18 分 39 秒。接著請外語系的學生協助我們對每一句評分語音之於標準語音作主觀的評分，之後再統計實驗中每一句語音人為評分的平均分數；另外我們再將這 200 組的語料透過本論文的英文語音評分系統，則每組評分語音都會得到四個特徵對應的差異程度 $dist_1, dist_2, dist_3, dist_4$。收集了這些差異程度和對應的分數後，以四項特徵參數的權重 $w_1 = 0.25, w_2 = 0.25, w_3 = 0.25, w_4 = 0.25$ 為啟始值，使用 Down-hill Simplex Search，就可以找出 $a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4$ 和這四項特徵參數的權重

w_1, w_2, w_3, w_4。

經由上述的實驗，我們得到音量強度曲線的權重為 7.45%；基頻軌跡曲線的權重為 22.40%；發聲急緩變化的權重為 17.24%；HMM 對數機率差異的權重為 52.91%。

之後我們將外語系同學對於 200 句語音訊號之於標準語音的主觀評判，分成三個等級(Bad、Average、Good)；另外也把這些語音訊號經由英文語音評分系統計算出分數後，依分數的高低以 60 分和 80 分為基準劃分成三個等級。最後再統計每句語音訊號的人工評分和評分系統的評分後，就可以得到表 4-1 的結果：

人工評分 / 評分系統評分	Bad	Average	Good
低於 60 分	**28**	17	7
60 分 ~ 80 分	20	**26**	20
高於 80 分	10	11	**61**

表 4-1　人工評分和英文語音評分的關係對照表

其中橫軸表示人工評分的等級項目；縱軸表示評分系統評分的等級項目；表格中央的部份則是表示符合兩種評分等級的語音數目。

由表格中對角正相關的數據可得知，在經由 Down-hill Simplex Search 調整各參數之後，我們的英文評分系統所評判出來的分數和人工評分有一定的正相關性，約為 (28+26+61) / 200 = 57.50%。