第3章　英文語音訊號切割

「語音訊號切割」提供一個將標準語料及評分語料切出正確發音區段的方法。以預先訓練好的兩種英文發音聲學模型當作比對標準，經由語音辨識技術，依不同的母語提供最適合的聲學模型來切割出正確的音素發音區段。以下章節將分成「聲學模型的訓練」和「利用語音辨識來進行語音訊號切割」這兩部份來介紹。

3.1 聲學模型訓練

實作語音訊號切割之前，我們必須先產生聲學模型，才能針對各種不同的語音進行切割動作。本論文中我們設計了兩種不同的聲學模型：一個是臺灣人口音的聲學模型；一個是外國人標準語音的聲學模型。

3.1.1 語料取得

首先針對母語為英文的聲學模型，我們使用 TIMIT 語料來加以訓練。語料內容為 2342 句平衡語料，由 438 位男性、192 位女性，共 630 人錄製，每人分配錄製 10 句，故共有 6300 句語音；依 TIMIT 的建議取其中 4620 句、總容量為 440 Megabytes、所有語料長度總和約為 3 小時 49 分 10 秒的語音訊號做為母語為英文的聲學模型訓練，另外 1680 句、總容量為 161 Megabytes、所有語料長度總和約為 1 小時 23 分 51 秒的語音我們用來做為外在測試檔(Outside Test)。

另一方面針對母語為國語的聲學模型，我們請 33 位學生，其中包含了 23 位男性、10 位女性，依 TIMIT 的資料錄製 7026 句平衡語料，我們取其中的 4684 句、總容量為 482 Megabytes、所有語料長度總和約為 4 小時 11 分 3 秒的語音做為母語為中文的聲學模型訓練，而另外的 2342 句、總容量為 226 Megabytes、所有語料長度總和約為 1 小時 57 分 43 秒的語音做為外在測試檔。

上述語料的音訊格式皆為 PCM；取樣頻率為 16 kHz；位元解析度為 16 bits。

3.1.2 聲學模型設計

英文中每一個音節可能由一個或數個音標所組成，而每一個音標都會對應到一個音素，目前我們使用的聲學模型是不考慮聲調與重音的。

在 TIMIT 語料中提出了 62 個音素，由於華人對於一些音素不像外國人念得那麼準確，再加上訓練語料的不足，如果我們減少訓練 Model 的個數，則可以使每個 Model 的訓練語料取樣數目增多。有鑑於上述兩個原因，因此我們將原先 TIMIT 設計的 62 個音素刪減成 40 個音素。在本章中我們使用的聲音單元即為每個音素一對一對應的單一單元(Unit-Gram)。舉例來說，“ school ”這個單字，其 KK 音標為〔*skul*〕，以我們設計的音素來表示，就是「**S**」＋「**K**」＋「**UW**」＋「**L**」，此聲音單元將視為語音的最小單元，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(Hidden

Markov Model, HMM)。

表 3-1 是我們所設計的 40 個聲學模型與 KK 音標對照表：

模型	音標	模型	音標	模型	音標	模型	音標	模型	音標
AA	*Z*	**D**	*D*	**IH**	*I*	**OW**	*o*	**TH**	*L*
AE	*G*	**DH**	*F*	**IY**	*I*	**OY**	*W*	**UH**	*U*
AH	*O*	**EH**	*A*	**JH**	*P*	**P**	*p*	**UW**	*u*
AO	*R*	**ER**	*S*	**K**	*k*	**R**	*r*	**V**	*v*
AW	*aU*	**EY**	*E*	**L**	*l*	**S**	*s*	**W**	*w*
AY	*aI*	**F**	*F*	**M**	*m*	**SH**	*B*	**Y**	*y*
B	*b*	**G**	*G*	**N**	*n*	**SIL**	*sil*	**Z**	*z*
CH	*Q*	**HH**	*H*	**NG**	*E*	**T**	*t*	**ZH**	*N*

表 3-1　40 個聲學模型與 KK 音標對照表

我們使用以下三種原則來對 TIMIT 的 62 個 Model 做刪減的動作，各 Model 後面刮弧內的英文字其底線部份即表示該 Model 的發音。

1. 替換：將發音相似的音素使用一個 Model 代替。

例如：**AXR** (butter〔*X*〕) → **ER** (bird〔*S*〕)

NX (winner〔*n*〕) → **N** (noon〔*n*〕)

2. 分解：將一個 Model 拆開成兩個以上的 Model 來組成。

例如：**EN** (button〔*Y*〕) → **AH** + **N** (〔*O n*〕)

ENG (Washington〔*I E*〕) → **IH** + **NG** (〔*I E*〕)

3. 刪除：將許多設定細微的暫停音素刪除。

例如：**PAU**、**EPI**

表 3-2 為 TIMIT 62 個 Models 修改前、後所對應的 40 個 Models 對照表。「╳」表示刪除；其它則表示經由修改後的狀況，如果前、後相同則表示未經修改。

修改前	修改後	修改前	修改後	修改前	修改後	修改前	修改後
AA	**AA**	**EH**	**EH**	**IY**	**IY**	**S**	**S**
AE	**AE**	**EL**	**AH L**	**JH**	**JH**	**SH**	**SH**
AH	**AH**	**EM**	**AH M**	**K**	**K**	**SIL**	**SIL**
AO	**AO**	**EN**	**AH N**	**KCL**	╳	**T**	**T**
AW	**AW**	**ENG**	**IH NG**	**L**	**L**	**TCL**	╳
AX	**AH**	**EPI**	╳	**M**	**M**	**TH**	**TH**
AX-H	**AH**	**ER**	**ER**	**N**	**N**	**UH**	**UH**
AXR	**ER**	**EY**	**EY**	**NG**	**NG**	**UW**	**UW**
AY	**AY**	**F**	**F**	**NX**	**N**	**UX**	**UW**
B	**B**	**G**	**G**	**OW**	**OW**	**V**	**V**
BCL	╳	**GCL**	╳	**OY**	**OY**	**W**	**W**
CH	**CH**	**HH**	**HH**	**P**	**P**	**Y**	**Y**
D	**D**	**H#**	**SIL**	**PAU**	╳	**Z**	**Z**
DCL	╳	**HV**	**HH**	**PCL**	╳	**ZH**	**ZH**
DH	**DH**	**IH**	**IH**	**Q**	**T**		
DX	**D**	**IX**	**IH**	**R**	**R**		

表 3-2　TIMIT 62 個 Models 修改前、後的關係對照表

3.1.3 特徵參數擷取

訓練聲學模型的第一步就是擷取訓練語料的特徵。我們將語音訊號經過特徵擷取，取出語音中的特徵，在此我們是採用 39 維的梅爾倒頻譜參數(Mel-Frequency Cepstral Coefficients, MFCC)【7】當作特徵參數。以下為擷取的基本流程：

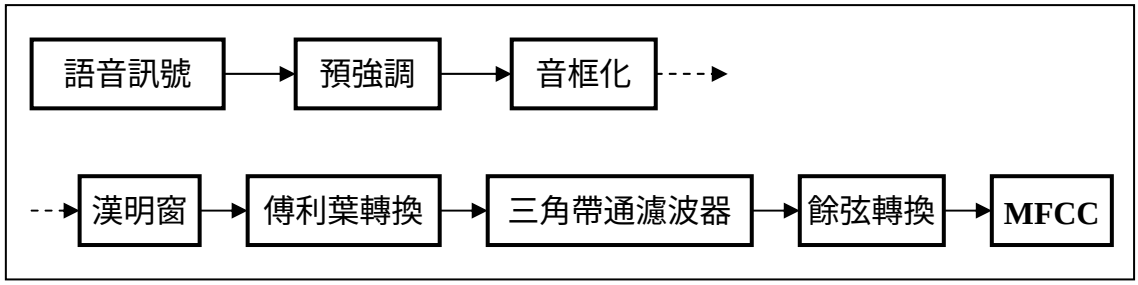

圖 3-1　梅爾倒頻譜參數擷取流程圖

為了補償語音訊號受到發音系統所壓抑的高頻部份，我們先將語音訊號經預強調(Pre-Emphasis)放大；之後取音框化，並對每個音框乘上漢明窗(Hamming Window)，以補償以音框為處理單位時，在邊緣所造成的訊號不連續現象；爾後對每個音框做快速傅利葉轉換(Fast Fourier Transform)，以求出每個音框的頻譜；再帶入一組 20 個三角帶通濾波器(Triangular Band-Pass Filter)即可求出每個頻帶的輸出對數頻譜 $m_j, j = 1,2,...,20$。經由前人研究【2】發現，人類對於低頻聲音的感知能力較強，對高頻聲音的感知能力較弱，所以設計三角帶通濾波器時就以低頻多取、高頻少取為原則；最後經由餘弦轉換即可求得 L 維的梅爾倒頻譜參數，其公式如下：

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

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

我們以 12 維梅爾倒頻譜參數與 1 維的對數能量，組成基本的 13 維特徵參數，再以這 13 維當作基礎，取其一階微分倒頻譜參數與二階微分倒頻譜參數，全部合起來總共是 39 維的梅爾倒頻譜特徵參數，如下圖所示：

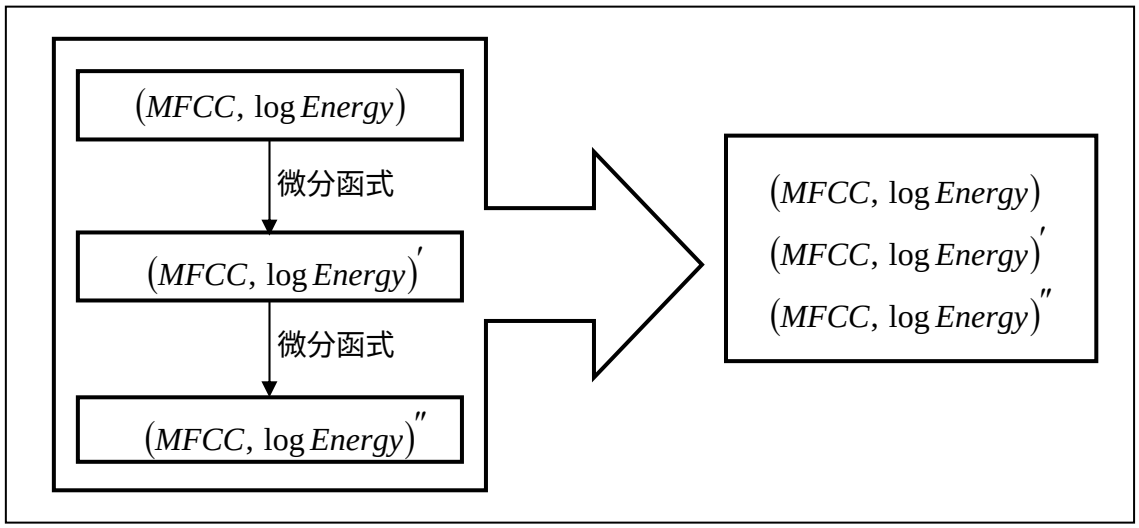

圖 3-2　39 維梅爾倒頻譜特徵參數示意圖

微分函式的意義在於梅爾倒頻譜參數相對於時間的斜率，也就是代表梅爾倒頻譜參數在時間上的動態變化程度。其公式如下：

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

上列公式中的 M 取 2，代表視窗寬度為 5 個音框，t 代表第幾個音框。

3.1.4 隱藏式馬可夫模型

本章英文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)【6】為基礎所訓練出來的。經由前人研究【2】【3】【4】，我們得知隱藏式馬可夫模型基本上是一種雙重且隨機的過程，而其稱為隱藏的主要原因在於其中有一組隨機過程是隱藏且無法看見，對應於語音時就如同人類在發聲的過程中其發聲器官，如聲帶、舌頭、口腔等，它們的變化沒有辦法從觀測的語音訊號序列中看出來。而另一組隨機過程則稱為觀測序列(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 的特性正好適用來描述語音的特性，我們可以把每個狀態看成是聲道(Vocal Tract)正處於某個發聲的狀態(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，或稱為高斯混合模型(Gaussian Mixture Model, GMM)。狀態觀測機率 $b_j(o_t)$ 定義為：

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

其中 $\#S$ 代表 Stream 數目；r_s 為 Stream 的權重，在本論文中設為 1；$\#M_s$ 代表 Stream 為 s 時，Mixture 的數目；w_{jsm} 及 G_{jsm} 則分別代表在狀態 j 下，Stream 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：

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

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。

在本論文我們使用 3 個 Stream，即 $\#S = 3$，Mixture 的數目有兩組，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 包含的 Mixture 個數，以第一組(6,2,2)為例，圖 3-3 為其示意圖：

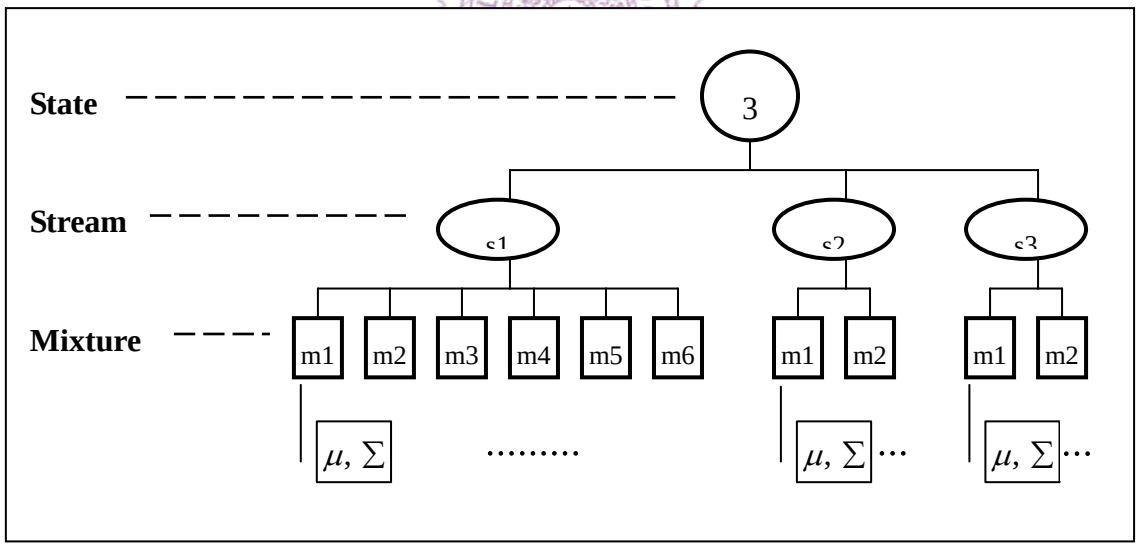

圖 3-3　State, Stream, Mixture 關係示意圖

3.2 語音辨識及語音訊號切割

我們之前不斷地強調，語音辨識對於本論文評分系統最主要的目的，就是希望透過針對不同母語背景所訓練好的聲學模型，再搭配一些方法而後達成準確的語音訊號切割。

在這一章所提到的語音訊號切割和第二章「說話驗證」使用的語音訊號切割技術和目的不完全相同。說話驗證系統中，我們希望在一個可以容忍的範圍內，利用辨識技術以 Pruning 的方式盡可能地將辨識網路延展至盡頭，因為這個時候我們還不能確定評分語音和標準語音的內容是相同的；但是在本章中，由於評分語音訊號已經通過說話驗證系統的考驗，因此我們可以確定語音訊號內容的正確性，在這樣的前提下，我們希望能用強迫對應(Forced Alignment)【6】的方式將語音訊號切割成各個音素的時間區段，以利評分機制的運作。

因此在這一章，語音訊號切割的主要目標即是希望能夠將連續的英文語音句子，其中包含了標準語音和評分的語音，切割成獨立的音素，如此一來我們才可以針對每一段句子中的音素和標準語音中的每一個音素做比較，而進一步完成評分的動作。

3.2.1 語音訊號切割流程

一個完整的語音訊號切割流程包含了辨識的前處理動作、語音的端點偵測

(Endpoint Detection)、特徵向量擷取、語音辨識等。

在前處理的過程中，我們必須讓辨識程式了解欲辨識的內容，因此首先將欲辨識的資料寫入文字檔中，接著再使用內含 127102 個單字的 CMU 字典(Dictionary from Carnegie Mellon University)來對該文字檔內的各個單字進行標音的動作。標完音後依標音結果建立對應的辨識網路。透過前處理動作產生的辨識網路可提供辨識程式完成語音訊號切割的功能。以下為前處理的基本流程圖：

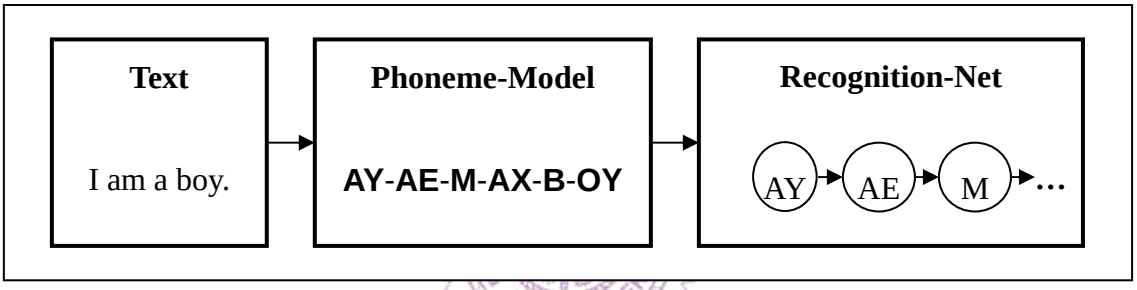

圖 3-4　語音訊號切割前處理流程示意圖

完成前處理動作後，我們可繼續進行主要語音訊號切割的流程，首先將一語音訊號經過端點偵測後再經由特徵擷取，取出語音中的特徵，在此我們採用 3.1.3 節所提到的 39 維梅爾倒頻譜參數當作特徵參數，然後將這些特徵參數透過聲學模型(隱藏式馬可夫模型)及語言模型(辨識網路)，利用維特比演算法(Viterbi Algorithm)即可找出最相似的音素，並得知各音素的時間區段。圖 3-5 為完成前處理後的語音訊號切割流程圖：

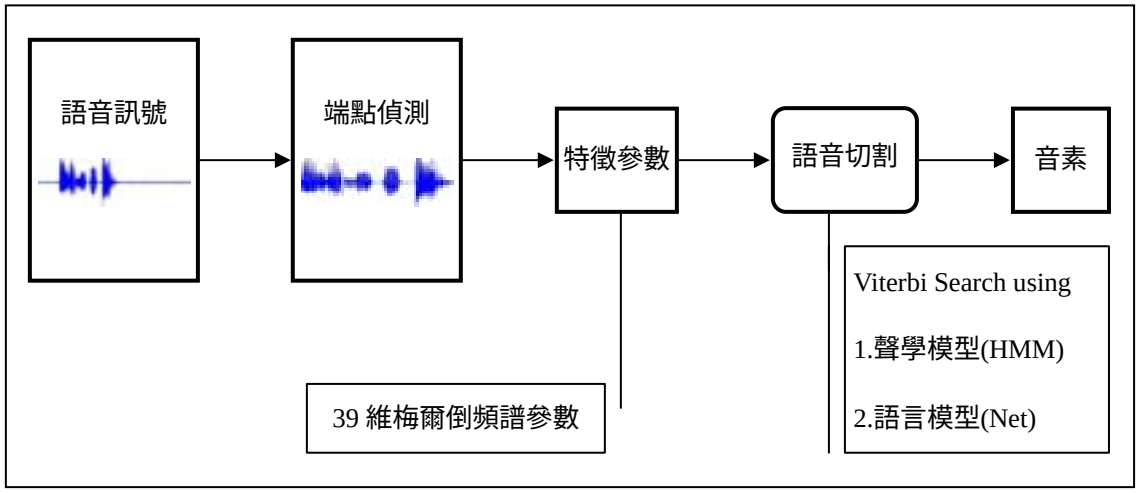

圖 3-5　語音訊號切割流程圖(前處理完成後)

我們將在接下來的小節中逐一介紹端點偵測和維特比演算法等語音辨識中重要的核心技術。

3.2.2 端點偵測

透過端點偵測，我們可以將語音訊號除去前段與後段的非有聲訊號，所產生的語音訊號可使辨識程式減少一些不必要的影響，如原先語音訊號前、後端的雜訊、靜音，而確實地達成其辨識的功能。

在端點偵測的理論中，前人提過諸多方法，而其中比較常拿來使用的是能量值與過零率(Zero-Crossing Rate)判斷法【5】，本論文就是採取這個方法來處理端點偵測。此方法利用兩個能量值的臨界值(Threshold)和過零率的另一個臨界值來斷定一段語音訊號的起迄點，這兩個能量值有一大一小，在此我們假設大的值為 L1，小的值為 L2，當能量大於 L1 時，則可以確定這一段的聲音訊號是有聲訊號而非

雜訊或是靜音；而當能量小於 L2 時，則可以判斷這段聲音訊號並非是有聲訊號，在此我們先暫時假設在這個 L2 以下的訊號是可以刪除的，最後經過過零率的把關下，假如被暫時假定刪除的 L2 訊號它的過零率值大於某個設定的臨界值，那麼這段訊號就不應該被刪除。另外假如訊號的能量值是大於 L2，但是小於 L1 的部分，那麼此訊號被 L2 所分割的起訖點便是一段語音的起迄點。

以下是經由上述的理論基礎，對於一個英文單音節做端點偵測的展示圖：

首先利用 L1 和 L2 臨界值法則，將有聲的部分留下，約略切在 0.77 秒與 1.32 秒左右，如圖 3-6 所示，其中上圖為聲波分布圖，下圖為 Log Energy 分布圖，下圖中兩條橫線代表 L1、L2，而兩條直線代表此單字音的起迄點：

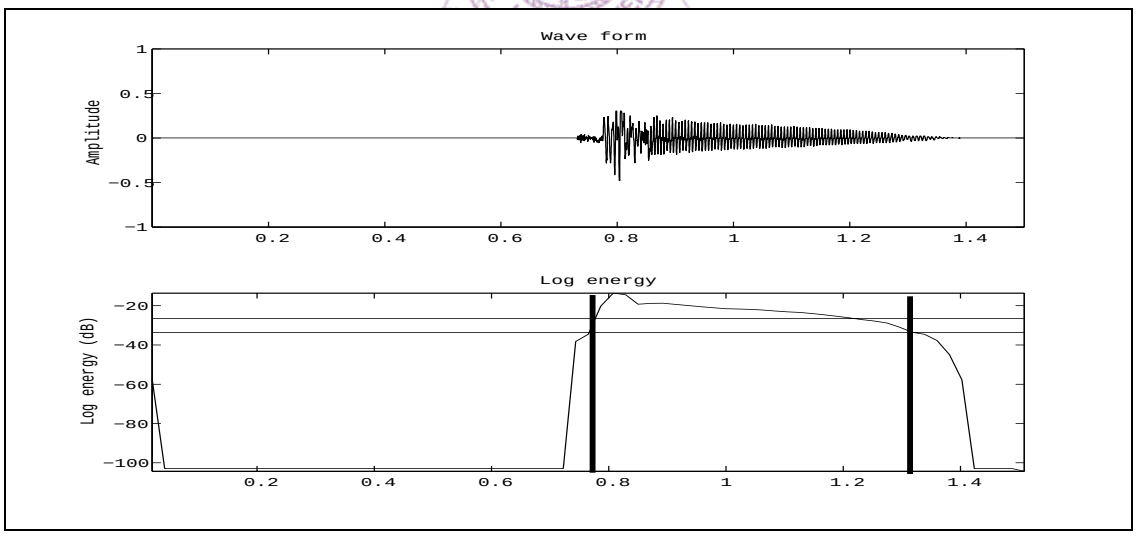

圖 3-6　以 L1 和 L2 來判斷單字音範圍

經由上述的切割法，可能會造成無聲子音的部分被刪除，這時可以利用過零率加以輔助判斷，假如被刪除部分是無聲子音，那麼它的過零率值是會大於某個

Threshold 值。此時要將此現象給補償回去。所以綜合第 1 步驟的結果，這個單音節的部分應該是在 0.75 秒至 1.32 秒之間。

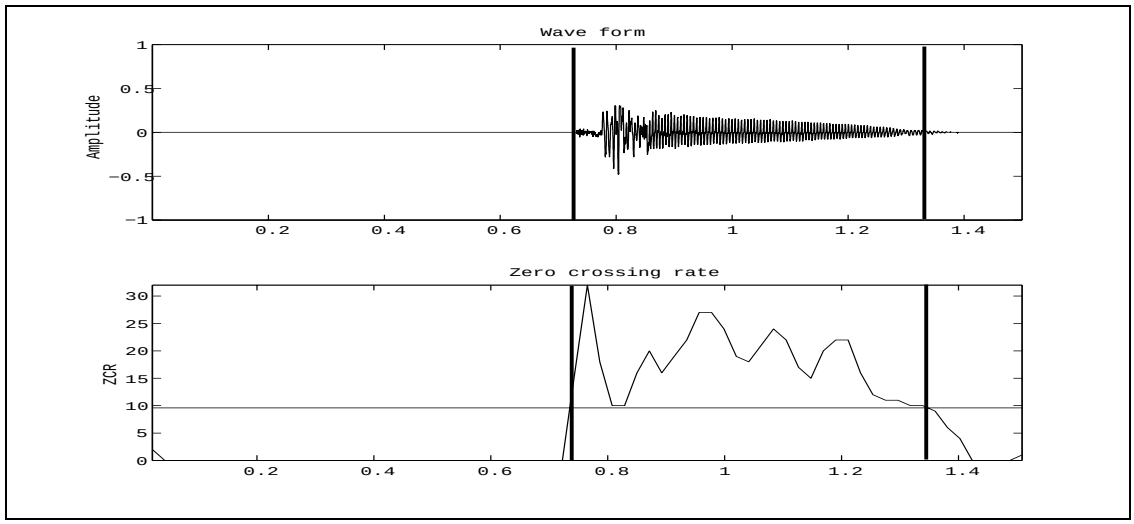

圖 3-7　以過零率來判斷單字音範圍

圖 3-7 的下方圖中橫線代表過零率的 Threshold 值,而兩條直線代表此單字音的起迄點。

3.2.3 維特比演算法

本論文所提出的語音辨識，其實是將語音訊號與之前已經訓練好的聲學模型做比對的動作；而要辨識一段經過特徵參數擷取後的語音訊號，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合適的。在此我們使用維特比(Viterbi)演算法來找出與觀測序列匹配的最佳狀態序列。

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，或稱為高斯混合

模型(Gaussian Mixture Model, GMM)。首先我們先介紹在實作上如何求狀態觀測機率$b_j(o_t)$，狀態觀測機率$b_j(o_t)$定義為：

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

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，介紹它的對數機率求取過程，Stream2 和 Stream3 亦同理。Stream1 的對數機率為

$$\log(w_1 G_1 + w_2 G_2 + ...)$$

依前人的研究【6】，我們可以轉換並簡化成

$$\log(w_1 G_1) + \log(w_2 G_2) + ...$$
$$= \log(w_1) + \log(G_1) + \log(w_2) + \log(G_2) + ...$$

而高斯函數如 3.1.4 節所述

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

因此

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

其中$\log\left((2\pi)^d |\Sigma|\right)$可依下式求得

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

另一部份$(x-\mu)^T\sum^{-1}(x-\mu)$亦可依以下的方法來求得

$$(x-\mu)^T\Sigma^{-1}(x-\mu)$$
$$= \begin{bmatrix} A & B & .. & .. \end{bmatrix}_{1\times 13} \cdot \begin{bmatrix} C & 0 & .. & 0 \\ 0 & D & 0 & .. \\ .. & 0 & .. & 0 \\ 0 & .. & 0 & .. \end{bmatrix}_{13\times 13} \cdot \begin{bmatrix} A \\ B \\ .. \\ .. \end{bmatrix}_{13\times 1}$$
$$= \begin{bmatrix} AC & BD & .. & .. \end{bmatrix}_{1\times 13} \cdot \begin{bmatrix} A \\ B \\ .. \\ .. \end{bmatrix}_{13\times 1}$$
$$= A^2C + B^2D + ...$$

其中共變異短陣$\begin{bmatrix} C & 0 & .. & 0 \\ 0 & D & 0 & .. \\ .. & 0 & .. & 0 \\ 0 & .. & 0 & .. \end{bmatrix}_{13\times 13}$的假設是每一維的維度和本身具相關性，和其他維則無相關性，所以採用對角化矩陣，如此一來可在反矩陣計算上可簡化其運算。

了解如何求取狀態觀測機率$b_j(o_t)$後，接下來我們介紹維特比演算法。首先假設觀察序列$\vec{O} = \{o_1, o_2, \ldots, o_T\}$的最佳狀態序列為$\vec{q} = \{q_1, q_2, \ldots, q_T\}$，並以$\delta_t(i)$代表從頭開始，直到時間點$t$的觀測值$o_t$為狀態$i$的最大機率，以下式表示：

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

由【7】我們可以得知

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

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

以下是維特比演算法【7】的步驟：

1. 初始 (Initialization)

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

$\psi_t(j)$用以回溯(Backtracking)

2. 遞迴 (Recursion)

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

3. 結束 (Termination)

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

4. 回溯 (State Sequence Backtracking)

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

如此即可找出最佳狀態序列。

3.3 英文語音訊號切割實驗結果

經由前面兩個小節的內容，我們可以得到語音訊號切割後各音素的時間區段，本節將透過實驗的結果，即語音切割的正確率，來對這一章做個簡短的結論。

關於實驗語料的部份，我們使用了 1680 句母語為英文的語音檔案，其語料的總容量為 161 Megabytes、所有語料長度總和約為 1 小時 23 分 51 秒，以下我們簡稱為 N-Wave (Waves from Native-Speaker)；另外使用了 2342 句母語為國語的語音檔案，語料的總容量為 226 Megabytes、所有語料長度總和約為 1 小時 57 分 43 秒，以下簡稱為 T-Wave(Waves from Taiwanese)，來做測試，而這些語音檔案都是外在測試檔(Outside Test)，換句話說，這些語料並沒有被用來和訓練語料一起訓練出我們欲得到的聲學模型。實驗用的語料其音訊格式皆為 PCM；音訊取樣頻率為 16 kHz；位元解析度為 16 bits；位元率為 256 kbps。

在聲學模型這個部份，我們訓練出了兩個聲學模型：一個是由母語為英文的使用者所錄製的訓練語料產生的聲學模型，以下我們簡稱為 N-HMM(HMM trained from Native-Speaker)；另一個則是由臺灣人所錄製的訓練語料所產生的，以下我們簡稱為 T-HMM(HMM trained from Taiwanese)。

關於實驗的方式，我們分別對每一句語音訊號和已知的語音內容文字作 Forced Alignment，再由產生的結果對每個單字及音素判斷其時間區段的切割是否正確。

為了比較兩個聲學模型所產生的影響，我們對語料(N-Wave, T-Wave) 和聲學模型(N-HMM, T-HMM)作交叉實驗。表 3-2 列出音素切割正確率的實驗結果：

實驗方式 / 項目	N-Wave N-HMM	N-Wave T-HMM	T-Wave N-HMM	T-Wave T-HMM
實驗語料句子總數	1680	1680	2342	2342
實驗語料句子容量	161 MB	161 MB	226 MB	226 MB
實驗語料時間長度	1:23:51	1:23:51	1:57:43	1:57:43
實驗語料音素總數	58282	58282	81229	81229
切割後正確音素總數	58253	57142	77293	80230
音素時間正確率	**99.95%**	**98.04%**	**95.15%**	**98.77%**

表 3-3　英文語音訊號切割實驗結果

在判斷音素時間正確率的部份，對於 N-Wave 而言，由於所有的語料 TIMIT 都有提供標音檔，因此我們可直接將切割出來的時間點和標音檔作比較，假設辨識得出音素的時間區段和該音素在標音檔裡的時間區段前、後各相差在 0.1 秒以內，則我們稱此音素的時間為正確；對於 T-Wave 而言，因為這些語料並沒有經過人工進行標音的動作，因此我們在這麼龐大的語料中取一部份進行人工的判斷，只要對人耳而言該區段聽起來不會相差太多，則我們稱該音素的時間為正確。

由表中的實驗結果我們可以看到使用不同的聲學模型經由 Forced Alignment 對語音訊號作切割得到的音素時間正確率都非常地高，不過針對不同語言背景的語句使用對應的聲學模型作切割辨識還是有較好的正確性。

另外，表 3-3 則是 N-Wave、T-Wave 透過大詞彙辨識的方式，經由 N-HMM、T-HMM 所得出的辨識率；其中詞彙內容為 2342 句英文句子、共有 6233 個不重覆的單字，而我們採用 Mono-Phone 當做辨識的最小單元。

實驗方式 / 項目	N-Wave N-HMM	N-Wave T-HMM	T-Wave N-HMM	T-Wave T-HMM
實驗語料句子總數	1680	1680	2342	2342
實驗語料句子容量	161 MB	161 MB	226 MB	226 MB
實驗語料時間長度	1:23:51	1:23:51	1:57:43	1:57:43
辨識正確句子總數	1650	622	1997	1425
句子辨識率	**98.21%**	**37.02%**	**85.26%**	**60.85%**

表 3-4　英文語音辨識率

由表中的結果我們可以發現，對於相同語料，N-HMM 的辨識率皆高於 T-HMM，這就表示當我們以 N-HMM 為聲學模型來對語音訊號求取對數機率時，所得到的對數機率值其可信度會高於 T-HMM。因此在接下來的章節中，我們將會以 N-HMM 當作我們對數機率比對的標準。

另外，由於錄製 T-Wave 的語者其英文能力參差不齊，因此未來若能將語料合併訓練，相信對 Forced Alignment 的音素時間切割正確率和英文語音辨識的辨識率，都能有較好的結果。