第1章　緒論

1.1　研究動機

隨著科技的日新月異，電腦處理速度突飛猛進，消費性電子產品的發明帶領人類邁向更便捷的生活模式。與這些電子產品溝通的方法，除了有一般電視、電話、與電腦鍵盤的按鍵模式，還有指紋機、手寫板等觸控模式，聰明的人類思考並尋求其他種更簡易的溝通方法。其中利用「語音」輸入，已成為劃時代的研究議題。因為語言是亙古以來人類仰賴彼此溝通、了解最自然快速的重要工具，目前已知世界上有多達數千種不同的人類語言，如果再加上動物界其他聲音(如海豚發出的聲音、火車經過的聲與下雨聲等)，這些種種聲音，都可當作辨識的圖案(Pattern)或碼(Code) 。如果我們能直接透過語音操作電子設備，且電腦能夠理解我們的要求，做適當的處理，將能節省許多人力和時間。

語音辨識的研究發展，從 1952 年美國貝爾實驗室發展的獨立數字辨識(Isolated-Digit Recognition)，之後隨著演算法、電腦速度的進步，由數字單詞、關鍵詞擷取(Keyword Spotting)[Wilpon *et al.* 1990]演進到口語對話系統(Spoken Dialogue)[Hazen *et al.* 2002]、大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)以及其他語音檢索(Speech-based Information Retrieval)的應用，可見語音辨識研究的蓬勃發展。

目前語音辨識遇到的重要問題是：語音辨識的正確率並沒有達到 100%，這也是多年來語音辨識專家、學者致力研究的核心重點。另一方面，英語是邁入國際化必要學習的重要語言，世界各國爭相將英語加入到國人必要學習的第二外語，然而非英語系國家在學習其他語言，可能因擁有第一語言的發音特性或習

慣，故在學習英語上會產生不同的發音腔調或變異，本論文初步探討台灣腔英語之連續語音辨識的情況與發音變異。

1.2 語音辨識流程

語音辨識流程簡單的說，將輸入的語音訊號，輸出成對應的文字或語音。然而若要達成此目的，則需經多重複雜步驟，如前端處理(Front-End Processing)、聲學比對(Acoustic Matching)與語言解碼(Linguistic Decoding)等細部運算。其中在聲學比對與語言解碼部分，需準備使用聲音語料訓練過的聲學模型(Acoustic Model)、使用文字語料訓練過的語言模型(Language Model)，以及經前端處理轉換過的語音特徵，以產生最相符對應的辨識文句輸出。基本語音辨識流程如圖 1-1 所示：

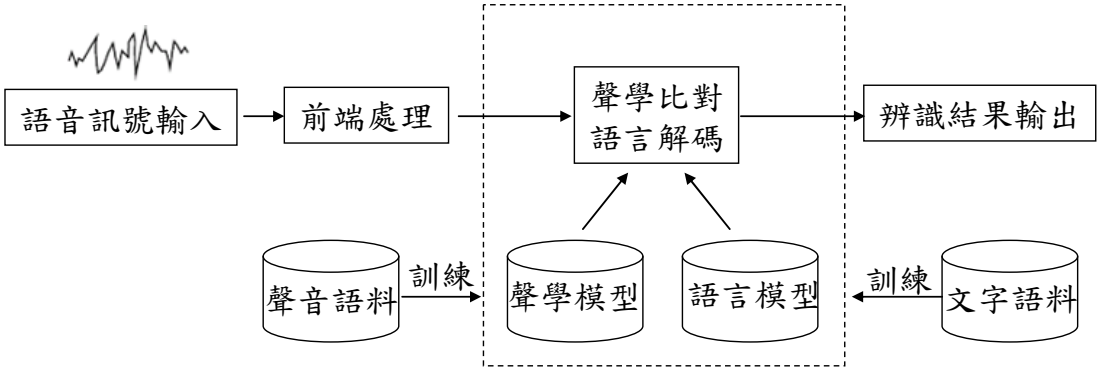

圖 1-1　基本語音辨識流程圖

以數學式來看[Jelinek 1999]，一段輸入的語音訊號段落以 O 表示，其中 O 可表示成語音特徵向量序列 $o_1, o_2, ..., o_T$，經由語音辨識器(Recognizer)辨識成對應的文字詞序列以 $\hat{W}$ 表示，為一連串詞 $w_1, w_2, ..., w_m$ 組成，語音辨識的過程即為找出

具有最大事後(Maximum A Posteriori, MAP)機率的詞序列，也就是代表O最有可能的對應輸出文句$\hat{W}$，可表示成式(1-1)：

$$\begin{aligned}\hat{W} &= \arg\max_{W} P(W|O) \\ &= \arg\max_{W} \frac{P(W)p(O|W)}{p(O)} \\ &= \arg\max_{W} P(W)p(O|W)\end{aligned} \tag{1-1}$$

$p(W|O)$為給定語音段落O時，詞序列W的事後機率。經過貝氏定理(Bayes Theory)轉換，可表示成$P(W)$、$p(O|W)$與$p(O)$。其中$p(O|W)$代表聲學模型(Acoustic Model)產生語音段落O的機率密度函數(Probability Density Function, PDF)，直接估測語音段落O發生在詞序列W對應的聲學模型相似度(Likelihood)；$P(W)$代表語言模型(Language Model, LM)產生詞序列W的機率，用於評估詞序列W於自然語言的合理性，可視為詞序列W的事前機率。語言模型輔助解決聲學上之混淆，使得最後選出的詞序列$\hat{W}$能夠符合該語言特性。另一方面，$p(O)$代表語音O之事前機率密度，因為對某句語音O進行辨識，每條詞序列都同除以$p(O)$，故可忽略。本論文使用連續密度隱藏式馬可夫模型(Continuous Density Hidden Markov Model,CDHMM)[Rabiner *et al.* 1989]作為聲學模型、*N*-連(*N*-gram)模型作為語言模型。

在語音辨識過程中，聲學比對負責將音素及語句中每個可能的段落做比對，計算其相似度；語言解碼使用維特比動態規劃搜尋(Viterbi Dynamic Programming Search) [Viterbi 1967]，對聲學相似度和語言機率進行解碼以便找出機率最大的可能詞序列。然而搜尋過程會隨著模型愈複雜，搜尋空間也呈現指數成長，故為了降低搜尋複雜度，本論文利用兩階段的搜尋來完成：第一階段進行聲學比對，並使用較低階的語言模型來搜尋，以產生詞圖(Word Graph)，第二階段在詞圖上使

用較高階的語言模型進行重新搜尋(Rescoring) [Ortmanns *et al.* 1997]。

1.2.1 特徵擷取 (Feature Extraction)

語音訊號前端處理，目的是擷取出合適的語音特徵參數，目前廣為人知的特徵參數有：梅爾倒頻譜係數(Mel Frequency Cepstral Coefficients, MFCC)[Davis *et al*. 1980]，線性預測係數(Linear Prediction Coefficients, LPC)[Gray *et al*. 1973]與感知線性預測係數(Perceptual Linear Prediction Coefficients, PLPC)[Hermansky 1990]等。本論文之特徵擷取實驗以梅爾倒頻譜係數(MFCC)為基礎，其擷取參數流程如圖 1-2：

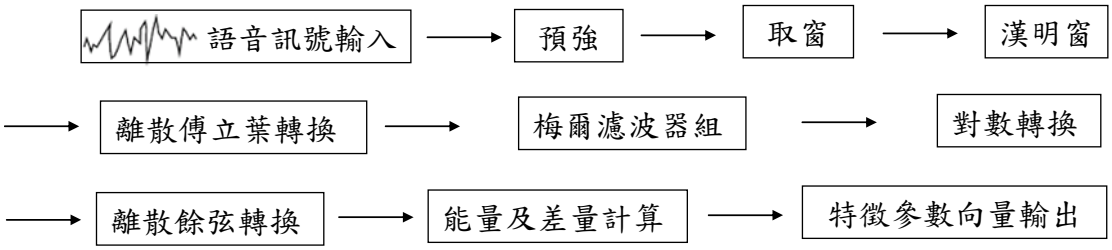

圖 1-2　梅爾倒頻譜係數之前端特徵擷取步驟

以下概述梅爾倒頻譜係數語音前端擷取過程如下：

1. 預強(Pre-emphasis)：

   因語音在空氣中傳送，或於發聲過程中聲門(Glottal)壓抑高頻部分能量，導致高頻的能量會快速遞減，故預強的功能是模擬人耳外聽道的功能，在時域上強調高頻能量。

$$H(z) = 1 - \alpha \cdot z^{-1} \tag{1-2}$$

式(1-2)可用來表示預強，其中 $H(z)$ 為高通濾波器在Z轉換(Z-Transform)的表示。實作上可在時域上處理，如式(1-3)，其中 $s(n)$ 為第 n 個採樣點，$\hat{s}(n)$ 為第 n 個採樣點經預強後的值。α 為預強參數，本論文設定為0.975。

$$\hat{s}(n) = s(n) - \alpha \cdot s(n-1) \tag{1-3}$$

2. 取框(Framing)：

在時域上觀測語音訊號的波形變化為十分迅速且無一定規則，但於頻域上觀察，則可以發現短時間(20ms~40ms)的情況下頻譜具有週期性的改變，所以在語音辨識的前處理，會假設語音訊號為短時間穩定(Short Time Stationary)，所以每隔一小段時間對語音訊號取一音框(Frame)，為了讓音框與音框之間聯繫著關係，所以音框與音框間會重疊(Overlap)一小段時間，此動作稱為取框(Framing)。本論文設定一個音框長為20ms，音框間重複為 10ms。

3. 漢明窗(Hamming Window)：

將時域的每個音框經離散傅立葉轉成頻域的訊號，但由於每個音框是固定時間點切割，所以音框左、右端的邊緣會造成訊號不連續現象，使得頻域上產生摺積效果，故在離散傅立葉轉換前會乘上一個漢明窗，特性在於主辦葉(Main Lobe)較寬，邊葉(Side Lobe)較窄，因此能有效壓抑訊號兩端，聚集中央部份特徵。漢明窗公式如式(1-4)，其中 α 為控制漢明窗之參數，本論文設定為 0.46。

$$w(n) = \begin{cases} (1-\alpha) - \alpha\cos\left(\dfrac{2\pi n}{N-1}\right) & n = 0,1,...,N-1 \\ 0 & otherwise \end{cases} \tag{1-4}$$

4. 離散傅立葉轉換(Discrete Fourier Transform, DFT)：

語音訊號在時域上的變化迅速且會隨著時間不斷改變，故不容易觀察出週期性的變化。為了找出語音訊號特性，故將語音訊號由時域轉成頻域，因為短時間內語音訊號在頻域上的能量分布是有規則性的，故一般可經由離散傅立葉轉換達成。

5. 梅爾頻率濾波器組(Mel-frequency Filter Bank)：

人耳對於聲音的高頻與低頻敏感度不同，在低頻部分人耳感受比較敏銳，而在高頻部分人耳的感受較不敏銳，因在耳蝸中不同位置的感受器連結到不同的反應神經，不同的反應神經代表不同的反應頻率，梅爾濾波器組主要是模擬人耳內部基底膜(Basiler Membrane)傳遞刺激到聽覺神經的過程。

6. 對數轉換(Logarithm)：

因人耳對於音強大小有不同解析度，為了保護人耳不受傷害，對音強小的聲音解析度較高，音強大的聲音有壓抑功能，使解析度較低，為模擬人耳此項功能，此將步驟 5 濾波器輸出取對數轉換。

7. 離散餘弦轉換(Discrete Cosine Transform, DCT)：

對數轉換後的梅爾頻率濾波器組的輸出，再經離散餘弦轉換成為梅爾倒頻譜係數，用意為降低維度間的關係，有助於隱藏式馬可夫模型在儲存共變異矩陣時資料的縮減，並可加快辨識效率，本論文之梅爾倒頻譜係數為 12 維。

8. 能量及差量計算(Log Energy and Time Derivatives)：

不同的音素(Phoneme)在能量(Energy)上的差異很大，由此可知能量為一重要的聲學特徵，一般會把能量與梅爾倒頻譜特徵結合。故於梅爾倒頻譜係數加上能量維後共 13 維，並加入各 13 維的一階與二階的差量計

算後，總共 39 維的 MFCC 語音特徵向量。

1.2.2 聲學模型 (Acoustic Model)

聲學模型的建立，首先對訓練語料中出現的各種語音訊號建立隱藏式馬可夫(Hidden Markov Model)之聲學模型，此模型中定義了四種參數：S、Π、A 與 B，其中 $S=\{s_1,...,s_n\}$ 表示每種模型中存有可能的狀態(State)，且此狀態中包含事件，這些事件可能為離散的事件或是連續事件，在連續事件中通常以高斯分布(Gaussian Distribution)來表示事件的分布情況，即為訓練語料中定義音素的分布情形，分布數目可能大於一，稱為高斯混合模型(Gaussian Mixture Model,GMM)。$\Pi=\{\pi_1,...,\pi_n\}$ 表示進入模型狀態的初始機率(Initial Probability)。$A=\{a_{ij}\}$ 代表任意狀態 s_i 與狀態 s_j 之間的轉換機率(Transition Probability)，$B=b_j(o_t)$ 代表語音特徵向量 o_t 在任意狀態 s_j 的觀測機率(Observation Probability)，如圖 1-3 代表單連(Monophone)音素 ax 對應的隱藏式馬可夫聲學模型：

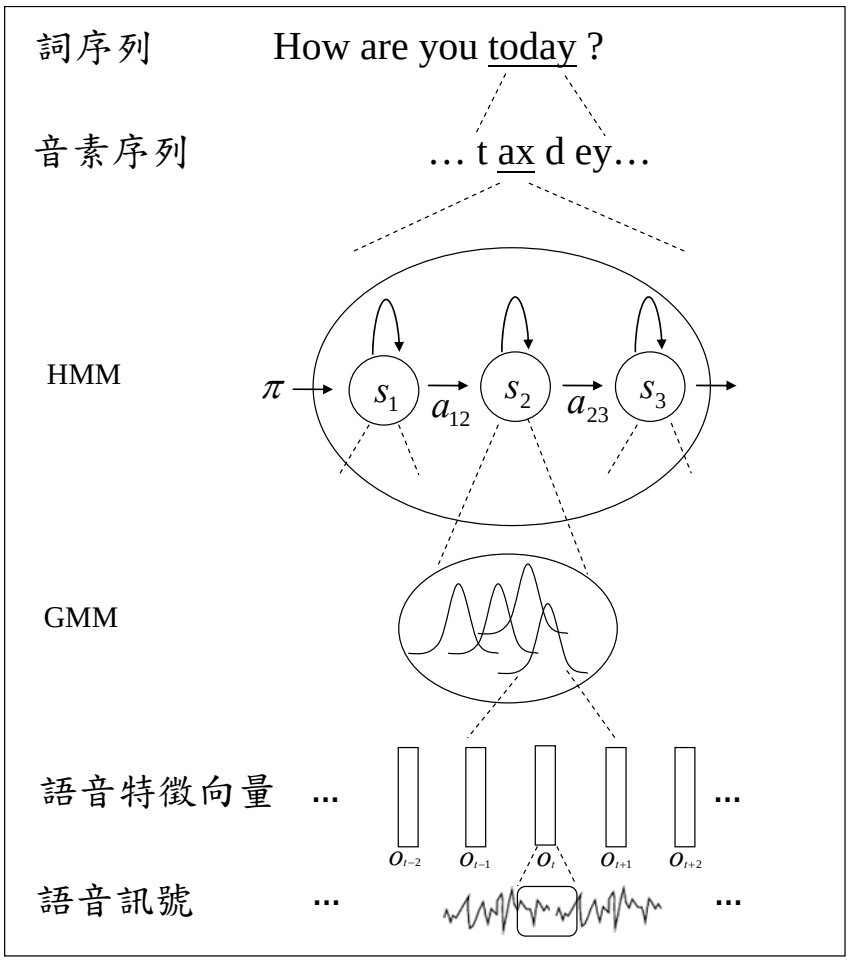

圖 1-3　單連音素 ax 之隱藏式馬可夫聲學模型

式(1-1)中 $p(O|W)$ 代表聲學模型分數，是假設經前端處理的語音訊號特徵向量序列 O（O 是經由一連串語音特徵向量 $o_1,o_2,...o_T$ 組成），特定文句 $W=w_1,w_2,...w_m$ 所對應的聲學模型下出現的機率。本論文使用英文訓練語料來訓練單連音素(Monophone)、二連音素(Biphone)、三連音素(Triphone)之聲學模型。聲學模型是利用最大化相似度訓練法(Maximum Likelihood Estimation, MLE) [Bahl *et al*. 1983]，配合使用波氏重估(Baum-Welch Re-estimation)演算法(又稱前向後向演算法, Forward-Backward Algorithm)[Baum 1972]訓練而得。我們可使用訓練好的單連音素、二連音素、三連音素之模型來構成詞或文句的對應聲學模型。

1.2.3 語言模型 (Language Model)

式(1-1)中 $P(W)$ 為文句 W 的語言模型分數。一篇文章出現某個詞的機率，可能與過去(History)出現的詞有關，因此式(1-1)的 $P(W)$ 可進一步利用連鎖率(Chain Rule)表示成式(1-5)：

$$\begin{aligned} P(W) &= P(w_1, w_2, ..., w_m) \\ &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)...P(w_m \mid w_1, w_2, ..., w_{m-1}) \\ &= P(w_1)\prod_{i=2}^{m} P(w_i \mid w_1, w_2, ..., w_{i-1}) \end{aligned} \tag{1-5}$$

若再進一步於式(1-5)中，假設目前的詞與過去出現過的詞無關，稱為單連詞(Unigram)，如式(1-6)所示。

$$P(W) = P(w_1)\prod_{i=2}^{m} P(w_i \mid w_1, w_2, ..., w_{i-1}) \approx \prod_{i=1}^{m} P(w_i) \tag{1-6}$$

如果目前的詞和過去緊鄰出現過的 N-1 個詞有關、和其他詞無關，稱為 N 連詞(N-gram)，如式(1-7)所示。

$$\begin{aligned} P(W) &= P(w_1)\prod_{i=2}^{m} P(w_i \mid w_1, w_2, ..., w_{i-1}) \\ &\approx P(w_1)\prod_{i=2}^{m} P(w_i \mid w_{i-N+1}, ...w_{i-2}, w_{i-1}) \end{aligned} \tag{1-7}$$

式(1-7)中機率模型可利用大量的文字訓練語料，利用最大化相似度估側法(MLE)，訓練而得。而如果有些詞沒有出現在訓練語料中，可用模型平滑化(Smoothing)技術，如 Katz [Katz 1987]、Kneser-Ney [Ney *et al*. 1994]與 Witten-Bell [Witten *et al.* 1991]等語言模型平滑化方法，使得在訓練語料中有出現的 N 連詞折扣(Discount)部分次數給在訓練語料中未出現過的 N 連詞，以解決機率值為 0 的情況[Chen and Goodman 1999]。

圖 1-4 是表示以詞「today」為現在目前出現的詞，往前看零個詞、一個詞(you)及二個(are、you)緊鄰出現過詞的情況：

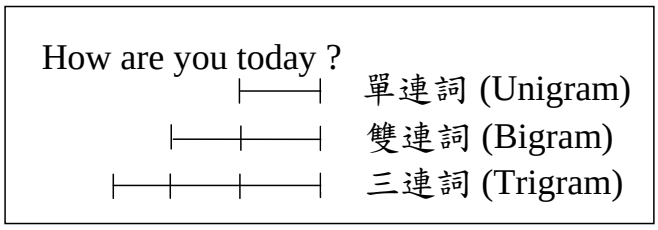

圖 1-4　以詞「today」為例的 *N*-gram 語言模型

1.2.4　語言解碼 (Linguistic Decoding)

依式(1-1)尋找最佳詞序列，備好所需之聲學模型、語言模型以及對應詞典(Lexicon)，使用維特比動態規劃搜尋法(Viterbi Dynamic Programming Search) [Viterbi 1967]找出輸入的語音訊號對應的最佳詞序列。由於搜尋空間會隨詞典與語言模型模型複雜度(例如詞二連、詞三連語言模型)大小呈指數成長，因此在搜尋時，通常會透過路徑裁減(Pruning)技術，停止搜尋機率較低的詞序列路徑，以降低計算複雜度與記憶體使用量。

1.3　研究內容

本論文研究內容，主要為：

1. 建立英文詞內(Intra-word)單連音素、二連音素、與三連音素(Triphone)狀態分享(State-tying)之聲學模型，觀察不同的聲學模型對語音辨識率的影響。

2. 探討台灣腔英語(EAT)、美國之音(VOA)兩套英文語料之連續語音辨識；於辨識器各階段中，分別使用鑑別性特徵擷取法、增加聲學模型之高斯混合數、

調適背景語言模型來提高語音辨識率。

3. 利用模糊矩陣尋找台灣腔英文發音變異，基於觀察之變異情況來修改訓練聲學模型狀態分享規則問題條件，重新修改狀態分享列表。另一方面，於語音辨識搜尋階段修正語音向量在隱藏式馬可夫模型的狀態之觀測機率，以改善系統辨識率。

4. 探討非監督式聲學模型訓練，首先對大量語料進行語音辨識，並使用語料及經辨識後自動轉寫文字(Transcription)資訊重新訓練聲學模型。

1.4 論文大綱

本論文大綱如下：

第二章　回顧現階段國外研究機構發展英文大詞彙連續語音辨識系統概況，以及聲學模型音素單位相似度測量方法。

第三章　介紹本論文實驗所用之音素、詞典與語料設定，以及所用之台師大大詞彙連續語音辨識系統。

第四章　詞內單連、二連、三連音素聲學模型訓練與兩組英文實驗語料之基礎實驗結果。

第五章　改善英文連續語音辨識之討論：使用鑑別式特徵擷取方法、語言模型調適方法、應用模糊矩陣於聲學模型訓練與語音辨識，最後討論非監督式最大化相似度之聲學模型訓練之方法與實驗結果。

第六章　結論與未來展望。