第1章 序論

「霹靂遊俠」是70年代一部非常熱門的電視影集。其中最令大家印象深刻的是主角李麥克的伙計-霹靂車，它不但擁有極高的人工智慧，並且還能直接接受主人的語音指令。的確，語音是人類最自然的溝通方式之一，人們也一直希望電腦能聽得懂人類的語言。過去由於電腦軟硬體技術的限制，加上對個人電腦使用者來說，利用鍵盤及滑鼠來當作輸入裝置也不會有太多不便，因此，電腦距離達到許多科幻電影或小說裡面的語音辨識功能，似乎相當遙遠。

如今隨著科技的發展，電腦技術的發展也跟著突飛猛進，消費性電子產品及許多可攜性裝置跟著融入了我們的生活中。為了節省體積及方便使用者攜帶，電子產品的體積越來越小是一定的趨勢。但鍵盤及滑鼠卻不可能跟著變小，否則會造成輸入不便，也不可能隨身攜帶它們，否則便失去了可攜性的便利。因此，過去利用鍵盤或滑鼠的輸入方式在可攜性裝置已經不太可行。許多研究學者便開始試著用語音來當作輸入裝置的一部份。過去數十年來，由於眾多學者專家的努力以及隱藏式馬可夫模型(Hidden Markov Model, HMM)[Rabiner 1989]的發展，自動語音辨識(Automatic Speech Recognition, ASR)的技術也獲得了很大的提昇。在今天，只要提供足夠的訓練資料，再利用一些標準的程序(如語音特徵參數擷取、聲學與語言模型訓練及建立辨識器等等)，便可以為某個特定的領域建立一套語音辨識系統，達到人類與電腦對話的初步夢想。但是，當我們將系統從實驗室環境移到真實世界應用時，由於環境的噪音、語者之間的差異以及通道效應等在真實環境才會遇到的問題，或是訓練資料與測試資料環境不匹配的情況，即使是目前最好的語音辨識系統，其效能依舊會大幅地下降。因此，要如何使語音辨識系統在真實環境使用，仍然有最好的效果，便成為了語音辨識研究裡一項極為重要的課題。此外，由於目前的語音辨識系統依然無法達到百分之百的正確，要如何讓系統能自動的判斷辨識結果的可靠性，對許多自動語音辨識相關應用，例如口語對話系統(Spoken Dialog System)[Hazen *et al*. 2002]、關鍵詞擷取(Keyword Spotting)[Wilpon *et al*. 1990]等，也很重要。除了

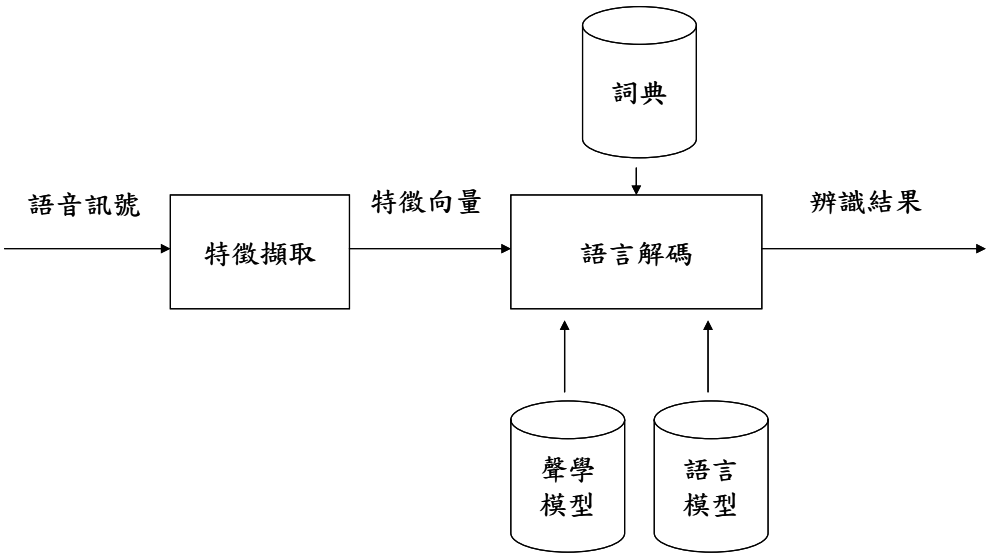

圖 1-1 語音辨識流程圖

利用信心度評估(Confidence Measures)找出辨識錯誤，並進一步決定是否要請使用者重新輸入外。更甚者，可以利用信心度評估修正辨識結果，提高系統的辨識正確率。而要如何達到上述之目標，也是目前亟待研究的方向。

1.1 語音辨識之流程

一般我們是以統計式的方法來實行自語音辨識[Jelinek 1999]。其數學式子可以寫成

$$W^* = \arg\max_{W} P(W \mid X) \tag{1-1}$$

其中 W 為一個語言 $\overline{\mathbf{W}}_{\Sigma}$ 中可能形成的任一詞序列(Word Sequence)，而 X 則代表輸入的語音訊號。統計式的語音辨識其方法很直覺，它希望在給定一段語音訊號後，從許多可能的詞序列中，找出一條詞序列，使得其在 X 出現時的機率是最大的。以人類的聽覺來說，就好像是要我們的大腦找出一條聽起來最有可能的詞序列。目前語音辨識系統的流程大致如圖 1-1所示，當一段語音進來時，語音辨識系統會先將語音訊號做特徵擷取的動作，接著針對所形成的語音特徵向量序列(Feature Vector Sequence)找出一條最符合的詞序列。而在這個步驟中，通常會將式(1-1)利用貝式定理展開

$$P(W \mid X) = \frac{p(X \mid W)P(W)}{p(X)} \tag{1-2}$$

其中，$p(X \mid W)$及$P(W)$便是分別代表經由圖 1-1中的聲學模型(Acoustic Model)及語言模型(Language Model)產生的分數。而這兩個模型一般是經由統計求得，也就是分別假設一組機率模型，利用訓練語料來估測其機率分佈(Probability Distribution)。以下分別簡介圖 1-1的每個模組運作方式。

1.1.1 特徵擷取(Feature Extraction)

這個部份主要是擷取語音訊號中比較重要的參數，一般較常用的語音特徵參數為梅爾倒頻譜係數(Mel-frequency Cepstral Coefficients, MFCC)[Davis and Mermelstein 1980]，其擷取步驟如圖 1-2所示。在取此特徵的時候，我們會將語音資料切割成一連串部份重疊的音框(Frames)，每一個音框最後表示成由13維的梅爾倒頻譜係數加上其一階與二階的時間軸導數(Time Derivatives)所組成的特徵向量。其中取一階與二階時間軸導數的原因主要是為了能獲得語音特徵在時間上(Temporal)的相關資訊。

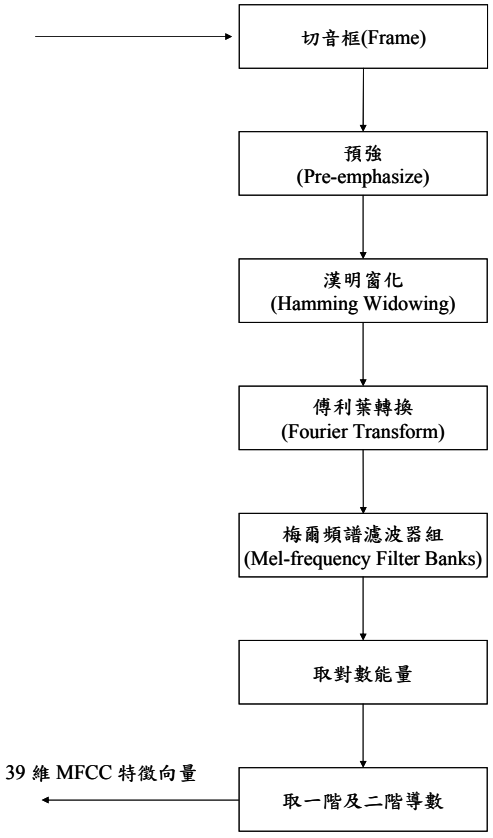

圖 1-2 梅爾倒頻譜係數特徵擷取步驟

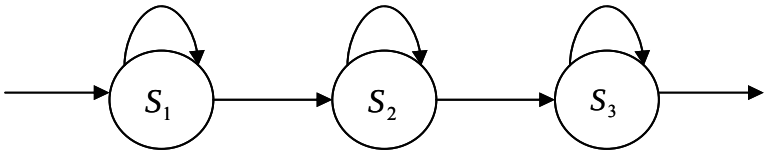

圖 1-3 隱藏式馬可夫模型範例

1.1.2 聲學模型(Acoustic Model)

為了處理語音訊號在時域上的變化，一般而言都是使用由左至右(Left-to-right)的隱藏式馬可夫模型(Hidden Markov Model, HMM)[Rabiner 1989]來作為聲學模型。圖 1-3 便是一個具有三個狀態(State)的HMM模型，每個狀態中都會有對每個音框所形成的語音特徵參數向量之觀測機率分佈(Observation Probability Distribution)。另外，每個狀態也有相對應的狀態轉移機率(State Transition Probability)，用來控制下一個時間點要停留在自己或是轉移到下一個狀態。根據語音特徵參數是連續或非連續的值，HMM每個狀態中的觀測機率估測方式可分為離散型(Discrete)、半連續型(Semi-continuous)及連續型(Continuous)三種[Huang *et al*. 2001]，目前的語音辨識系統主要都是連續型或半連續型為主。就連續型而言，為了減少估算觀測機率的參數量，以及因為任何機率分佈理論上皆可以由多個高斯分佈(Gaussian Distributions)來逼近的特性，一般而言都是使用高斯混合分佈(Gaussian Mixture Distribution)來近似此機率分佈。而連續型與半連續型主要的差別在於連續型每個狀態擁有自己的高斯分佈，半連續型則會有共用高斯分佈的情況。

由於一個中文音節(Syllable)是由一個聲母(INITIAL)及一個韻母(FINAL)組成，22 個聲母及 38 個韻母構成約 400 個音節。基本上，我們只要為每個聲母及韻母建立屬於它的聲學模型，便可以辨識所有的中文音節。本論文共使用了 38 個韻母模型，但在聲母模型的部份則是考慮不同種類的右邊相連韻母對其發音特性會造成不同的影響，而將 22 種聲母再細分成 112 種聲母模型，亦即聲母部份採用右相關聯模型(Right-context-dependent Model, RCD);另外，我們加入一個靜音(Silence)模型來估測語音訊號中靜音部份。為了讓聲學模型有更精確的估算能力，除了要有足夠的訓

練資料之外，還需要有好的訓練方法，較常見的有最大化相似度訓練法則[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]、最大化交互資訊(Maximum Mutual Information, MMI)[Bahl *et al.* 1986]、最小化分類錯誤(Minimum Classification Error, MCE)[Juang and Katagiri 1992]以及新近被提出的最小化音素錯誤(Minimum Phone Error, MPE)[Povey 2004]訓練法則等。

1.1.3 語言模型(Language Model)

由於聲學模型本身只能辨識某一段語音訊號發的是何種音節(Syllable)序列，無法確認其對應的詞(中文有許多同音詞)，且句子中詞跟詞的連接其實存在概略的規則，因此便需要有語言模型的存在。由於語言模型的機率分佈是離散型的，在估計語言模型的機率時，並不使用機率密度分佈函數，而是直接估測個別詞序列的機率資量函數 $P(w_1, w_2, \ldots w_N)$，其中 $w_1, w_2, \ldots, w_N$ 為此詞序列所包含的詞。但對整個詞序列的估測參數會隨著詞數量成指數成長，因此會遭遇資料稀疏(Data Sparseness)的問題。為了解決此問題，我們會先將語言模型的式子展開成機率的連乘積，再利用 n-1階的馬可夫假設(n-1 Order Markovian Assumption)做簡化的動作，如式(1-3)所示:

$$P(W) = P(w_1, w_2, \ldots, w_N) = \prod_{k=1}^{N} P(w_k \mid w_{k-1}, w_{k-2}, \ldots, w_{k-n+1}) \qquad (1\text{-}3)$$

其中 N 為詞的個數，$w_{k-1}, w_{k-2}, \ldots, w_{k-n+1}$ 則是 w_k 的歷史詞序列，式(1-3)便是常見的 n-連(n-gram)語言模型表示法。一般為了方便起見，以及減少參數量的複雜度，常使用詞雙連(Bigram)及詞三連(Trigram)兩種模型(也就是分別使用一階及二階的馬可夫假設)。如同聲學模型，語言模型也需要有大量的文字語料來做為訓練之用。n-連語言模型的訓練方法有最大化相似度估測法(Maximum Likelihood Estimation, MLE)及最大熵值法(Maximum Entropy, ME)[Rosenfeld 1996]等，另外為了處理某些詞可能在訓練語料沒有出現的問題，通常會搭配如Katz Smoothing[Katz 1987]及Kneser-Ney

Smoothing[Ney *et al*. 1994]等語言模型平滑技術，對這些估測機率原本為零的部份加以平滑化處理。

1.1.4 語言解碼(Lingusitc Decoding)

在依式(1.2)尋找最佳詞序列時，由於分母的部份並不會影響最後詞序列排名的結果，因此實作上常將分母的部份省略。有了這項前提之後，再將每個音節比對語句中每一個音框，找出一條最佳的詞序列。而為了有效率的求解，一般是使用維特比動態規劃搜尋(Viterbi Dynamic Programming Search)[Viterbi 1967]。此外，由於搜尋空間會隨著詞典大小成指數成長，因此，在搜尋時，通常會透過搜尋路徑裁減(Pruning)技術來停止繼續尋找一些機率較低的詞序列，以減低其計算複雜度及記憶體使用量。

1.2 現階段語音辨識研究內容

目前在語音辨識的研究中，強健性語音辨識(Robust Speech Recognition)主要在探討如何使得語音辨識系統在真實充滿各種噪音的環境底下仍有一定的辨識效果，而除了維護語音辨識系統的效能之外，信心度評估則是應用於準確地判斷語音辨識系統的結果正確性，在1.2.1及1.2.2小節將分別簡介強健性語音辨識及信心度評估的研究。

1.2.1 強健性語音辨識

此方法主要為降低環境噪音或不同語者等因素對語音訊號的影響，或是找出比較具有鑑別性的語音特徵，使得語音辨識的正確率不會因環境的噪音因素而有所降低。大致來說，可再細分為兩個方向:

(i) 語音強化(Speech Enhancement)

這類方向主要是提昇語音訊號本身的品質，希望藉由乾淨語音及噪音不同的統計特性，想辦法將受噪音影響的聲音訊號還原成乾淨語音。常見的技術有頻譜消去法(Spectral Subtraction)[Boll 1979]及維爾濾波器(Wiener Filter)[Wiener 1949]等。

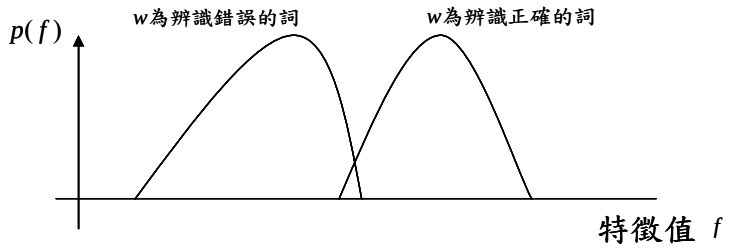

圖 1-4 一個預估參數範例

(ii) 強健性語音特徵(Robust Speech Features)

這類技術是以擷取語音訊號中較具有強健性的特徵為主要目的，使得擷取出來的特徵可以抵抗週遭的環境變化。常見的技術有倒頻譜平均消去法(Cepstral Mean Subtraction, CMS)[Atal 1974]、倒頻譜正規化法(Cepstral Normalization, CN)[Viikki and Laurila 1998]、統計圖等化法(Histogram Equalization, HEQ)[Korkmazsky 2004]等。

除了以上兩項技術之外，還可以利用鑑別性分析(Discriminant Analysis)方法來計算原始語音資料的一些相關統計資訊，將原本的語音特徵投影到新的特徵空間，以得到較具有鑑別性的特徵。較常見的方法有線性鑑別分析(Linear Discriminant Analysis, LDA)[Duda and Hart 1973]、異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)[Gales 1999]等。

1.2.2 信心度評估

信心度評估為本論文主要的研究重點，此研究方向的基本應用是給定語音辨識系統的輸出結果一個分數(輸出結果可以是針對整句詞序列，詞序列中的某個詞，或是音節等其它更小的單位，而給定的分數通常是介於0~1之間)，來判斷這個辨識結果的可靠度。舉例來說，信心度評估可以辨別每個辨識出來的詞它被辨識正確的機率有多高。如果依方法來分的話，大致上可分為三大類[Jiang 2005]：

(i) 以特徵為基礎(Feature-based)之信心度評估

此種方法通常都是利用在進行語音辨識的過程中可獲得的一些所謂的預估特徵(Predictor Features，包含聲學及語言等資訊)。而一個特徵要能被稱為是預估特徵，其特徵值對正確辨認詞及錯誤辨認詞所建立的機率密度函式(Probability Density Function, PDF)必須具有很大的鑑別性，如圖 1-4 所示。另一方面，每個預估特徵之間可利用某種方式結合，再配合不同的分類器，如有限向量機(Support Vector Machine, SVM)[Zhang and Rudnicky 2001]、自然貝氏分類器(Naïve Bayes Classifier)[Sanchis *et al*. 2004]或決策樹(Decision Tree)[Eide *et al*. 1995; Neti *et al*. 1997]等來決定辨識結果的正確性。

(ii) 發音確認 (Utterance Verification)

此種方法則是將信心度評估視為統計式的假設檢定(Hypothesis Testing)的一種問題[Rose *et al*. 1995]。在這個架構之下，通常會提出兩個互斥的假設:

$$\begin{aligned} H_0 &\text{ (虛無假設，Null Hypothesis) : } X \text{ 之辨認的結果為正確} \\ H_1 &\text{ (對立假設，Alternative Hypothesis) : } X \text{ 之辨認的結果為錯誤} \end{aligned} \tag{1-4}$$

其中 X 代表一段聲學觀測序列。然後，我們測試虛無假設及對立假設，以決定辨識結果之正確與否。而測試之方法則是使用相似度比例檢測(Likelihood Ratio Testing, LRT) :

$$\frac{p(X \mid H_0)}{p(X \mid H_1)} \underset{H_1}{\overset{H_0}{\gtrless}} \tau \tag{1-5}$$

τ 為事先設定的門檻值(Threshold)，而 $p(X \mid H_0)$ 及 $p(X \mid H_1)$ 一般來說可使用隱藏式馬可夫模型來做估算。如果計算出來的值大於門檻值，我們便相信辨識結果的正確性，否則，便認為辨識結果為錯誤的。

(iii) 事後機率 (Posterior Probability)

在傳統的最大事後機率(Maximum a Posterior, MAP)語音辨識方法中，式(1-2)的事後機率 $P(W \mid X)$ 對詞序列而言其實可以算是一種很好的信心度評估準則。但

是如1.1.4小節所提到的，我們通常會省略分母項，造成語音辨識系統輸出的分數不再是介於0到1的值。即使不省略分母項，但由於語音訊號有無窮多種，所以要如何估測出 $p(X)$ 便變成了一個癥結所在。為了解決這個問題，先前學者的研究曾提出了下列兩種方式來求得近似解:

A. 填充化基礎(Filler-based)法:此類方法主要是需要另外一組填充模型(Filler Model)或背景模型(Background Model)，如全音素辨識(All-phone Recognition)[Young 1994]、全包式模型(Catch-all Model)[Kamppari and Hazen 2000]等。

B. 圖形化基礎(Graph-based)法:這類的方法主要是根據前向後向演算法(Forward-backward Algorithm)在詞圖上(Word Graph)上計算分母的值[Kemp and Schaaf 1997;Wessel *et al*. 2001]等，本論文將會於 2.2 小節更深入的探討此種方法。

1.3 本論文研究成果與貢獻

如之前所提到的，在過去的研究中，信心度評估主要是用來判斷語音辨識系統其辨識結果的正確性，再決定是否要接受其結果，應用範圍並不算廣泛。直至最近，開始有學者將信心度評估應用至別的研究領域，如使用事後機率增進語音辨識系統的正確率[Wessel *et al*. 2000]、非監督式(Unsupervised)聲學模型訓練[Wessel and Ney 2005; Chen *et al*. 2005]以及大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)系統的往前觀測(Look-ahead)法則[Afify *et al*. 2005]等。

在傳統判別辨識結果的正確性研究方面，本論文提出使用熵值資訊並結合以事後機率為基礎之信心度評估方法，在公視新聞外場記者及受訪者測試語料最佳實驗結果中，較傳統以事後機率為基礎之信心度評估有16.37%及12.00%的信心度錯誤率相對下降。而在探討信心度評估於提高語音辨識的正確率方面，吾人嘗試結合以梅爾倒頻譜係數及異質性線性鑑別分析搭配最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)兩種不同語音特徵參數所形成的詞圖資訊，並以最小化

音框錯誤率(Minimum Time Frame Error Decoding)[Wessel *et al*. 2001b]來降低辨識系統的字錯誤率，實驗結果顯示，在外場記者及受訪者測試語料中各獲得4.6%及4.8%的相對字錯誤率減少。最後，本論文嘗試在傳統以Levenshtein距離為成本函式(Cost Function)的最小化貝氏風險(Minimum Bayes Risk)[Goel and Byrne 2000]，適當的加入以特徵為基礎的信心度評估，經由*N*-最佳化詞序列重新排序(Reranking)實驗得知，相較於傳統單單使用Levenshtein距離為成本函式的最小化貝氏風險辨識法則而言，有少許的相對字錯誤率減少。

1.4 論文架構

本論文第二章將回顧過去有關於信心度評估的研究，主要討論過去有關如何計算信心度評估的方法，以及一些應用信心度評估至其它領域的研究。第三章則是介紹大詞彙連續語音辨識系統的基本架構、實驗語料的設定以及實驗評估方法。第四章則是有關信心度評估、使用事後機率增進辨識系統正確率的基礎實驗。第五章的部份則是除了探討如何結合熵值(Entropy)資訊與傳統信心度評估外，還討論關於結合了不同語音特徵參數所形成的詞圖後之最小化音框錯誤率(Time Frame Error)解碼以及結合了信心度評估的最小化貝氏風險(Minumum Bayes Risk)解碼之實驗。第6章則是結論與未來之展望，探討未來可繼續研究之方向。