第2章　說話驗證

「說話驗證」可以視為英文語音評分的防火牆。在許多情況下，我們不僅關心測試語料的內容，更關心語料的內容有多麼的可靠，當我們可以針對不同的評分語音產生對應的數值，並依此而對該評分語音內容的正確性做出判斷，就是所謂的說話驗證(Utterance Verification)【1】。

2.1 驗證系統簡介

我們實作的驗證系統主要是基於語音辨識的精神，將目前語音處理常運用到的技術應用於其中，再建立合理的說話驗證系統。

簡單的說話驗證流程如圖 2-1 所示，當驗證系統接收到語音訊號後，分別對每個音素進行語音辨識，之後再依辨識結果的機率值排名並配合驗證機制給予最後的可信度值。

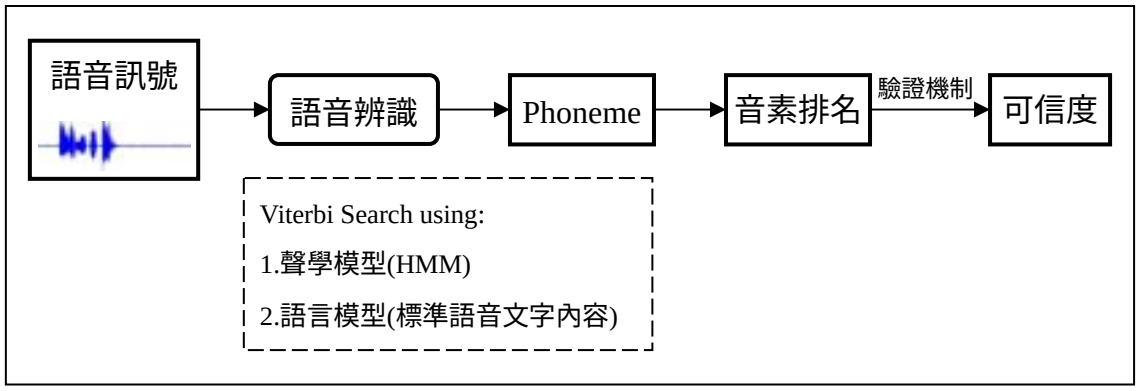

圖 2-1　說話驗證系統流程圖

2.2 語音訊號切割

在設計說話驗證系統時，由於此時還不能確定評分語音和標準語音的內容是否為相同，因此我們希望在一個可以容忍的範圍內，利用辨識技術盡可能地將代表語言模型的辨識網路(Recognition-Net)展開至盡頭，以便針對能夠切割出時間區段的各音素進行排名比對。

將語音訊號切割後，如果評分語音的內容和標準語音相當類似，則經由切割後產生音素的數量將接近甚至等同於標準語音音素的數量；相反地，若一句評分語音訊號中只有前 n 個音素和標準語音相同，後幾個字的發音則是完全不同，則經由語音辨識後產生的音素數量也就大約等於 n。舉例來說，如果標準語音為「I am a good student」、評分語音為「I am smart and cute」，由於兩語音只有前兩個單字「I」、「am」，也就是前三個音素「**AY**」+「**AE**」+「**M**」相同，則經由語音辨識後，在評分語音中所能切割出來的音素數量大約是 3。

這裡我們所使用的語音辨識技術，是採用 Pruning 的方式，將語音盡可能地依序切割出每一個音素，對於沒有切割出來的音素，我們將其可信度值設為 0，如此一來可以增加驗證系統的區別性，使與標準語音內容完全不相同的評分語音其可信度值變得相當低。這也就是我們為什麼不使用 Forced Alignment 強迫切割出每一個對應於標準語音音素的最主要原因。

圖 2-2 為兩個語音經由語音訊號切割後產生的不同結果。上半部的語音內容等同於標準語音內容，因此切割出來的音素很完整，而下方的語音內容只和標準語音的前半段內容相同，因而辨識程式將樹狀網路展開至節點 **UW** 之後，就無法再繼續切割下去。

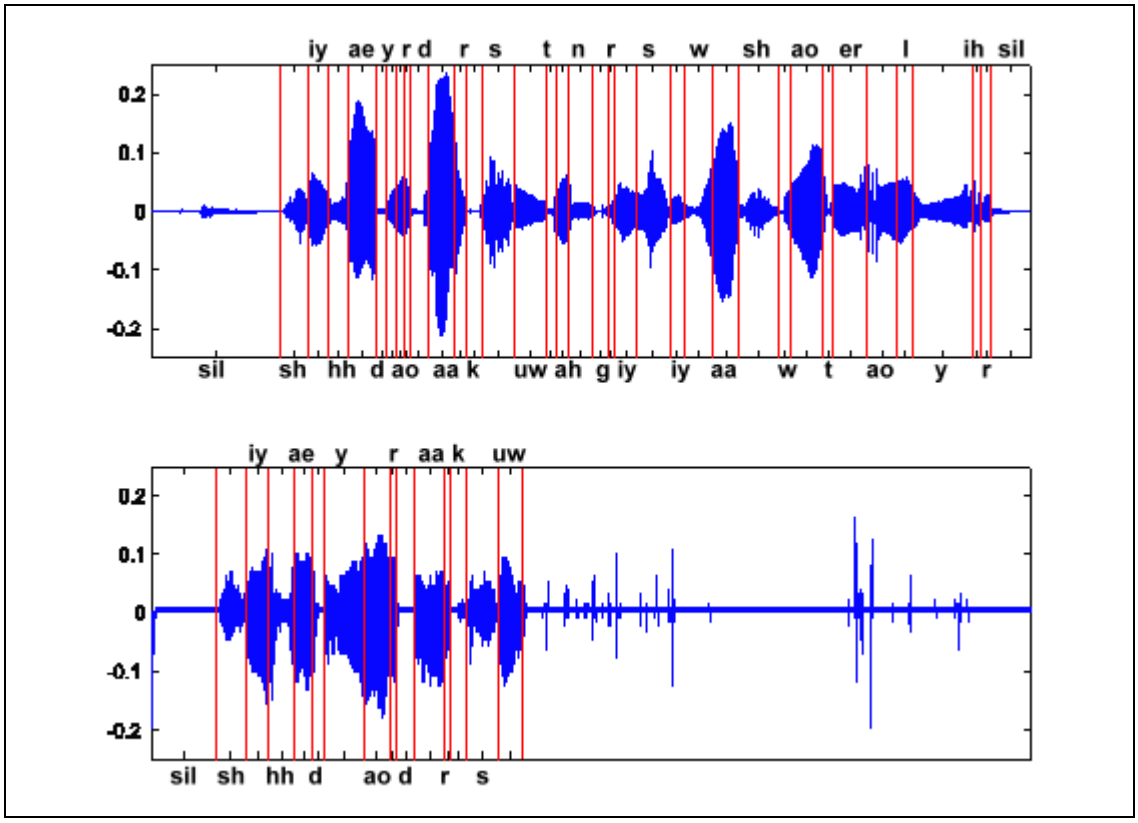

圖 2-2　說話驗證的語音訊號切割比較圖

2.3 驗證機制建立

經由語音訊號切割得到音素時間區段後，首先針對每個音素以 39 個 Phone Models 來比較對數機率並做排名的動作，之後再經由設計的驗證機制得到可信度值。我們將在以下小節中逐一介紹。

2.3.1 音素排名

對於語音訊號中已經切割好的每個音素，我們利用語音辨識分別建立 39 個辨識網路，並對每個音素計算求出相對於這 39 個辨識網路所得到的對數機率，這裡要注意的是，建立的 39 個辨識網路並不包含「**SIL**」，也就是代表無聲(Silence)的 Model。最後再將對數機率排序後即可得到如圖 2-3 的機率分佈：

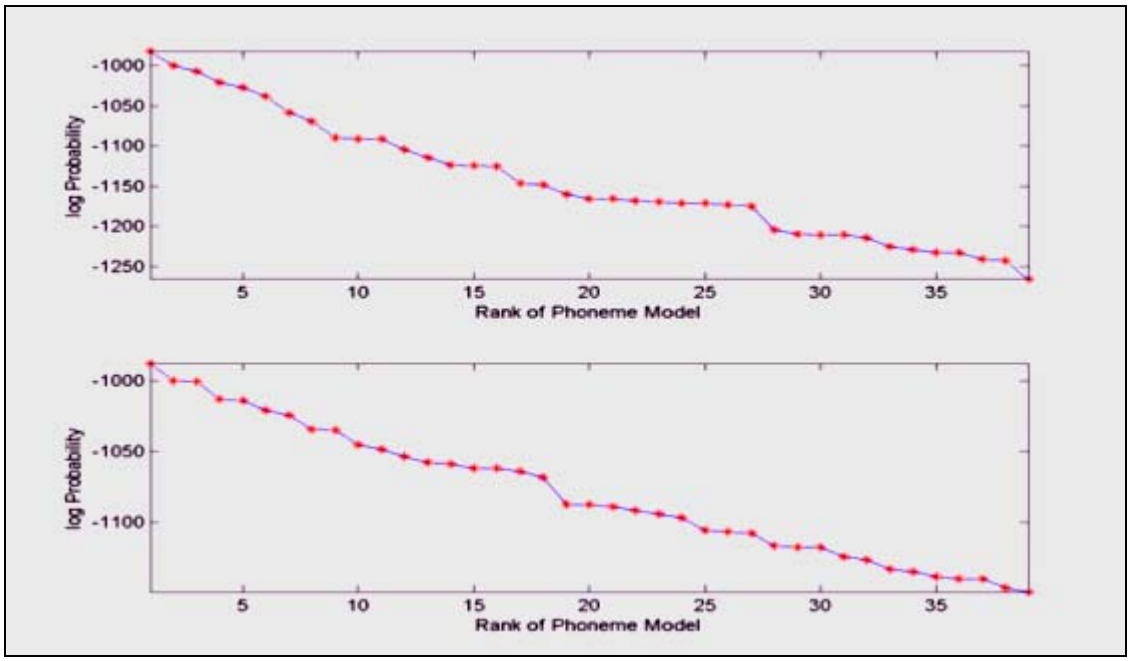

圖 2-3　音素機率排名

上下兩個機率分佈表示不同的音素經由辨識程式求得 39 個對數機率的結果，由該圖我們可以看出，對於不同的音素，即使排名同樣是第二名，可是和第一名的對數機率差距卻不相同，會造成這樣的原因在於有些音素的發音相似，而有些音素的發音差異則相當大【16】，因此我們對於上方圖中的音素，可解釋成其第一名和第二名 Model 的發音很接近，造成對數機率的差距相當小；而在下方圖中的音素，也許在我們 39 個 Models 中，只有一個 Model 的發音和該音素接近，因此更加突顯了其第一和第二名的對數機率差距。

2.3.2 驗證機制

經由語音訊號切割之後，產生的結果可能有兩種情況：一種是部份的語音訊號已經成功切割出時間區段的音素；另一種則是語音訊號的後半部可能沒有辦法切割出音素。而在這一節討論的驗證機制，主要是針對前者的情況，也就是如何將音素的排名正規化，得到一個合理的數值。

在 Sukkar 和 Lee 於 1996 年發表的論文【17】中提到，音素的對數機率差距和驗證系統的可信度值成反比，而音素相對於所有 Phone Model 的排名和可信度值成正比，基於以上的前提，我們將 Sukkar 和 Lee 所提出求取可信度值的式子改寫並以下列公式表示：

$$value_{pho} = \frac{2}{1+\exp\left(\alpha \cdot \left(Rank_{pho}-1\right) \cdot \frac{\log P_{Rank_{pho}}}{\log P_{Rank_1}}\right)}$$

$\exp(x)$表示 e^x，即自然對數的 e 的 x 次方；$Rank_{pho}$ 和 $\log P_{Rank_{pho}}$ 分別表示該音素在 39 個 Models 中的排名及對數機率值；1 表示第一名；α 為我們調整的參數值。由此公式可得知，當某音素相對於 39 個 Models 的排名為第一名時，該音素的可信度值為 1。

圖 2-4 表示對於「**SH**」這個 Model 藉由上述的公式可將其對應於 39 個 Models 所產生的對數機率及名次換算成可信度值。從圖中可以看出，當名次在第 10 名左右時，可信度值已經降至 0.2 了。

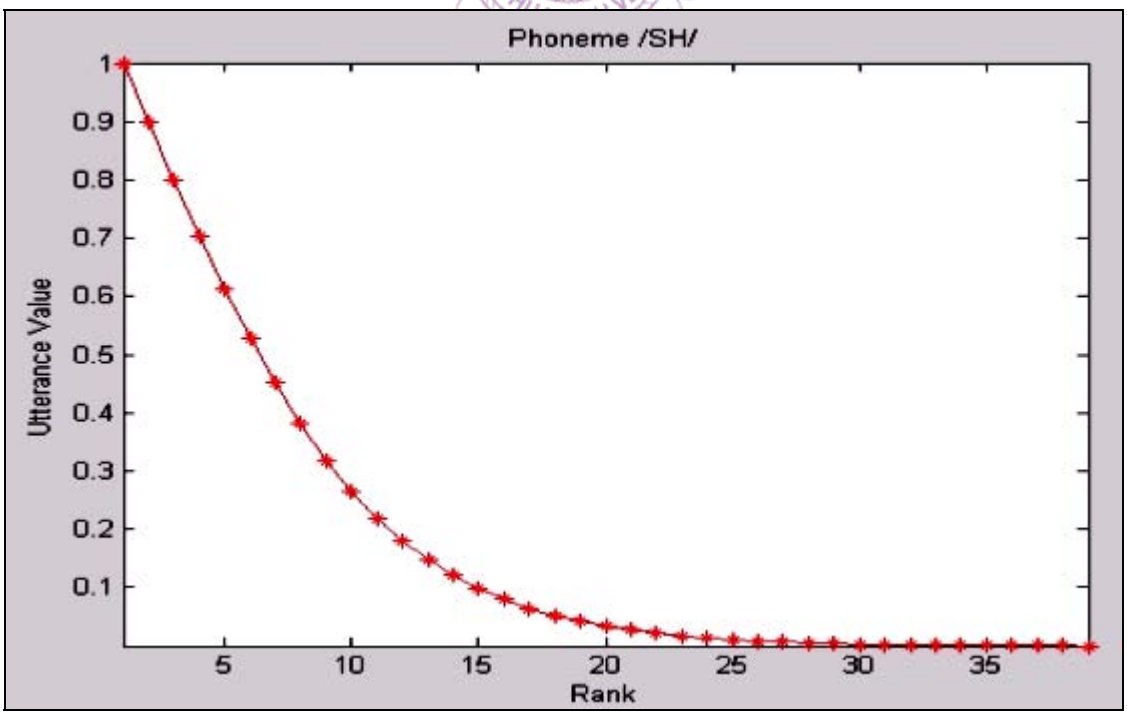

圖 2-4　音素 **SH** 的排名與可信度值的關係

另外由於音素間發音的差異性，因此我們在評斷可信度值時，不能單純地以排名來做比較。舉例來說，音素「**OW**」〔*o*〕和「**S**」〔*s*〕比對完 39 個 Models 後同樣都得到第二名的結果，但是對於「**OW**」而言，其第一名是「**AO**」〔*R*〕，而「**S**」音素的第一名是「**T**」〔*t*〕，則我們可以很明顯地看出「**OW**」和第一名的對數機率差距較小，也因此可信度值應該要比較高才合理。因此在上述公式中，我們將排名的差異再乘上對數機率的比例差異，如此一來就會使得每個音素的可信度值受到排名及對數機率的影響。最後經由計算得到的可信度值介於 0 和 1 之間。

當計算出句子所有成功切割的音素可信度值之後，利用每個音素的時間長度占句子時間長度的百分比作為權重，即可推導得出一句語音訊號的可信度值。以下是設定的公式：

$$value_{sen} = 100 \cdot \sum_{n=1}^{N} \frac{len(pho_n)}{len(sentence)} \cdot value_{pho_n}$$

N 為一單字中評分音素的數量；$len(x)$ 表示 x 的時間長度。至於有些單字可能其中的一些音素沒有辦法經由語音訊號切割產生，對於這些音素，我們就直接將其 $value_{pho}$ 設為 0；最後乘上常數 100 代表我們將說話驗證系統的結果定義在 0 至 100 之間。

2.4 說話驗證實驗結果

我們將一句輸入的語音經由辨識程式進行語音訊號切割和音素排名後，可以針對該句語音訊號求出對應的可信度值，但是這樣還是沒有辦法從數字來決定這一句話的內容可靠性，因此我們設計了一個實驗，期待從實驗中求出說話驗證系統的門檻值(Threshold)。

假設語音訊號透過驗證系統得到的可信度值高於門檻值，則我們稱「此句語音訊號的內容和標準語音訊號的內容相同」這句話是相當可靠的，也就表示我們可以放心地讓這句語音訊號經由之後英文評分系統的各個步驟繼續進行評分的動作；相反地，若語音訊號透過驗證系統得到的可信度值比門檻值還要低，則表示這句話和標準語音的內容不相同，因此我們也就停止讓兩句不相同的語音進行後續的評分動作。

在接下來的實驗中，對於實驗語料我們分兩部份來蒐集：

1. Correct

取 168 句說話內容相同的語音訊號當作標準語音內容，這部份語音檔案的總容量為 17.6 Megabytes，所有語音長度總和約為 9 分 10 秒。

2. Incorrect

取 168 句內容不等於標準語音內容的語料，這部份語音檔案的總容量為 15.8 Megabytes，所有語音長度總和約為 7 分 31 秒。其中一部份內容和標準語音完全不

相同；另一部份則是語音訊號內容「部份相同」於標準語音內容。在此我們定義一句話中若存在連續 3 個單字以上和標準語音內容相同，但並不是完全相同，即為「部份相同」。

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；所有的實驗語料皆為單聲道。

接著將上述兩部份各 168 句的實驗語料經由說話驗證系統得到對應的可信度值，而後再統計、分析這些可信度值即求得驗證系統的門檻值。圖 2-5 為求取門檻值的實驗結果分佈圖，橫軸為可信度值的範圍，縱軸為可信度值處於該範圍內的語音訊號個數。

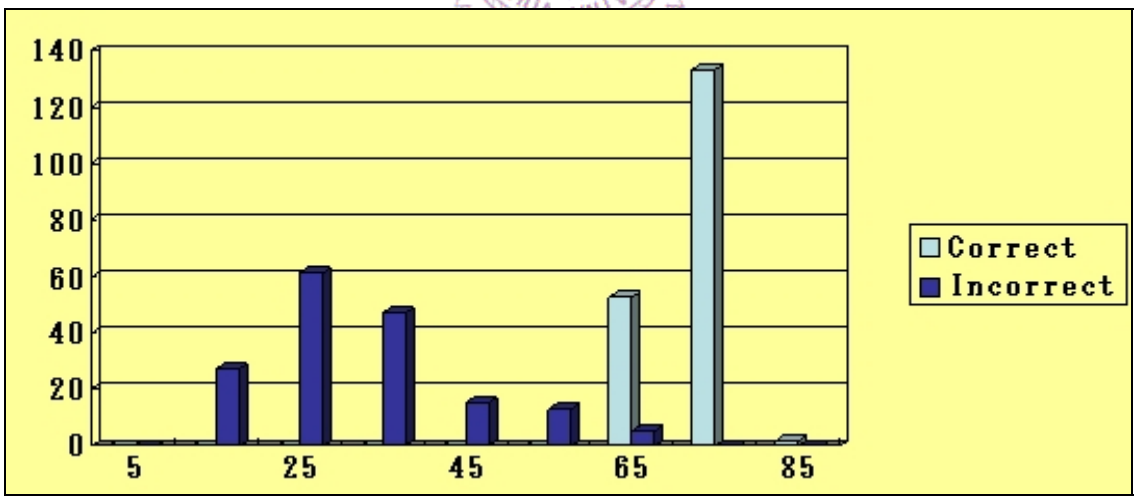

圖 2-5　說話驗證求取門檻值實驗結果分佈情況

圖 2-6 則是表示在這兩組語料中，正確語料被拒絕(False Rejection －Type Ⅰ Error)及錯誤語料被接受(False Acceptance － TypeⅡ Error)的 ROC 關係圖

(Receiver Operator Characteristic)。

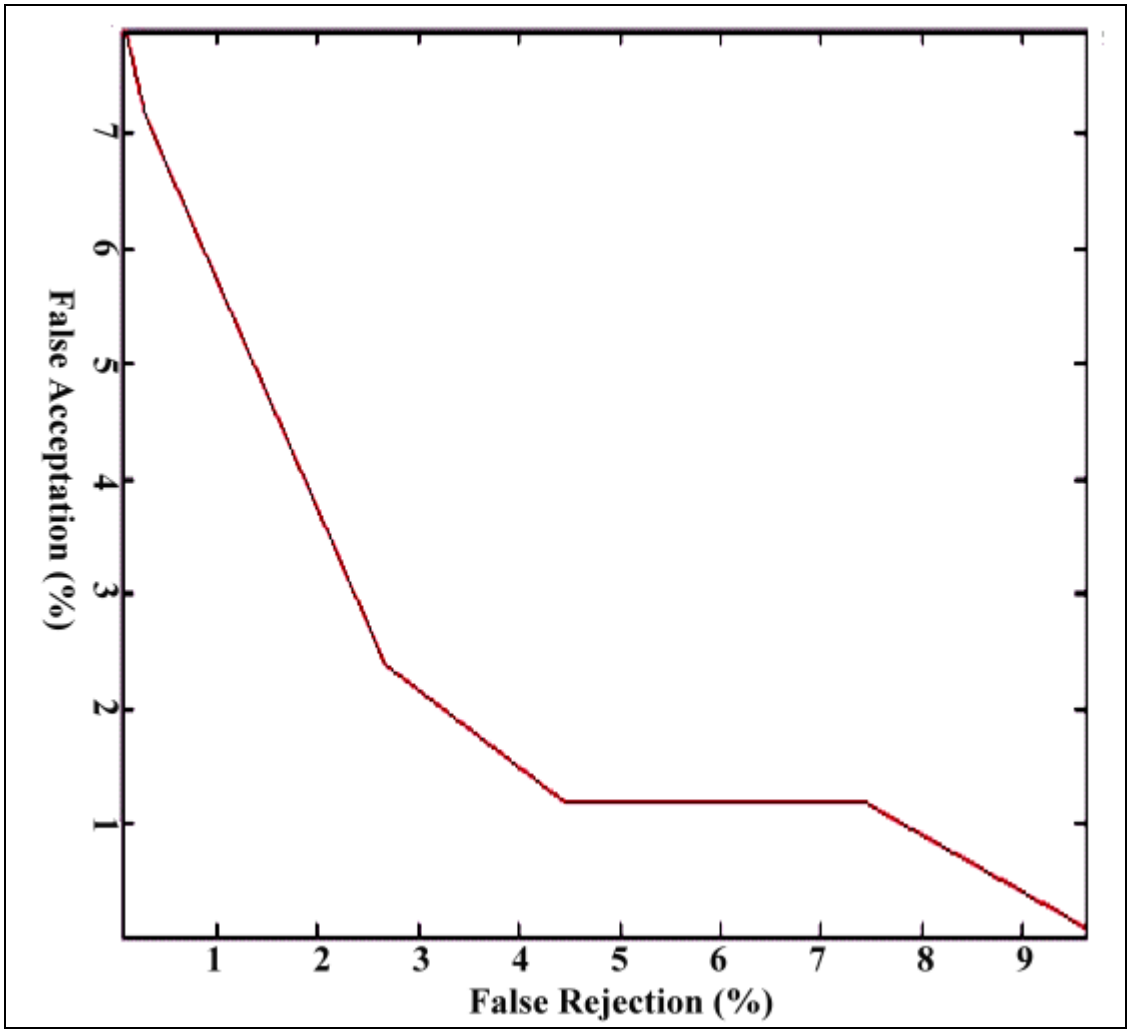

圖 2-6　False Rejection 與 False Acceptance 的 ROC 關係圖

我們以「正確拒絕的語音數加上錯誤接受的語音個數其值為最小」做為尋找門檻值的前提。根據實驗結果，我們發現 Correct 中的語料其最小可信度值為 63.21，而在 Incorrect 可信度值大於 60 的語料中最接近 63.21 的可信度值為 61.59，因此我們將說話驗證系統的門檻值設定成 62.40(即兩者的平均)，如此則正確接受率為 100%；正確拒絕率為 0%；錯誤接受率為 1.19%；錯誤拒絕率為 98.81%。

經由上述實驗計算求出門檻值後，我們另外準備一組內含 Correct 及 Incorrect 各為 168 句的語料，其中 Correct 語料的總容量為 14.3 Megabytes、所有語料長度總和約為 7 分 27 秒，Incorrect 語料的總容量為 17.2 Megabytes、所有語料長度總和約為 8 分 57 秒。將這些語料經由以門檻值為 62.40 的驗證系統後即可觀察其正確率。實驗結果分佈如圖 2-7 所示。

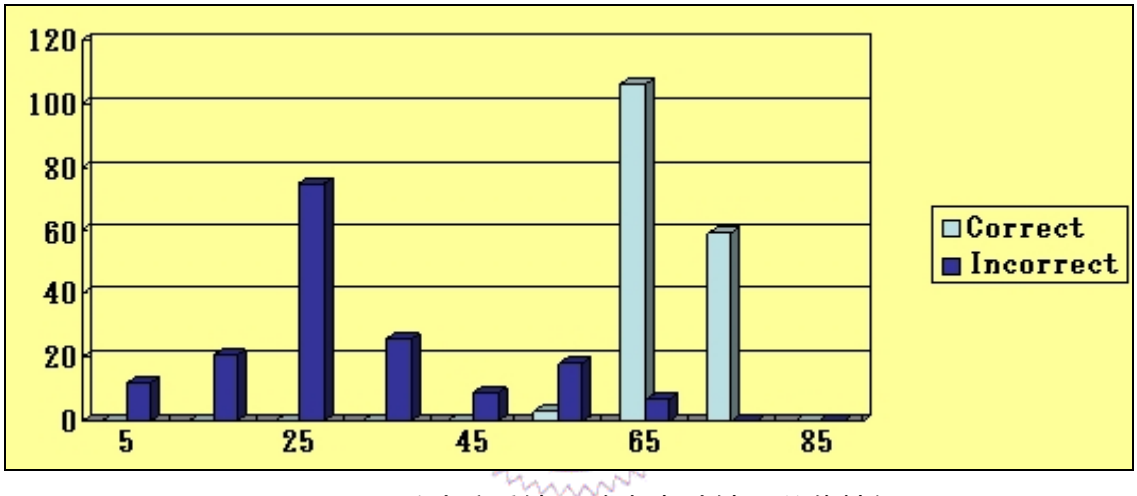

圖 2-7　說話驗證系統正確率實驗結果分佈情況

根據實驗結果，以門檻值 62.40 來決定一句語音是否和標準語音內容相同，則我們可以得到正確接受率為 92.86%；正確拒絕率為 7.14%；錯誤接受率為 0.60%，錯誤拒絕率為 99.40%。