第4章 基礎實驗討論

本章實驗包含了三小節。4.1小節為建立外場受訪者聲學模型之實驗，此小節除了比較梅爾倒頻譜係數(MFCC)以及異質性線性鑑別分析加上最大相似度線性轉換(HLDA+MLLT)兩種語音特徵參數對語音辨識系統正確率的好壞之外，同時也探討最大化相似度訓練以及最小化音素錯誤訓練對於聲學模型訓練的影響。最後，則是嘗試在語言模型方面加入相同領域的語言模型訓練語料(請參照3.1.3小節)，來降低大詞彙連續語音辨識系統的錯誤率。4.2小節針對傳統信心度評估方法進行實驗討論;4.3小節則是討論關於前人應用信心度評估於降低詞圖搜尋錯誤率之實驗。

4.1 外場受訪者基礎實驗

4.1.1 最大化相似度(Maximum Likelihood, ML)訓練之實驗

此實驗的初始聲學模型之狀態高斯混合機率分佈均視為平均值等於0、標準差為1的標準常態分佈(Standard Normal Distribution)，利用HTK Tookit[Young *et al*. 2002]內建函數，根據MFCC和HLDA+MLLT兩種不同的語音特徵參數，各進行30次最大化相似度訓練。每間隔5次最大化相似度訓練後之聲學模型對於MATBN外場受訪者測試語料的自由音節辨識(Free Syllable Decoding)之錯誤率結果請參考表 4-1。而30次訓練之自由音節辨識錯誤率曲線請參考圖 4-1。接著，我們採用第30次的聲學模型作為大詞彙連續語音辨識系統的初始聲學模型，在執行詞彙樹複製搜尋及詞圖搜尋時，分別調整語言模型分數的權重，如式(4-1)中的β:

$$p(X \mid W)P(W)^{\beta} \tag{4-1}$$

觀察其對語音辨識系統錯誤率的影響。詞彙樹複製搜尋的結果可參考表 4-2及圖 4-2;而詞圖搜尋的結果可參考表 4-3及圖 4-3。

訓練次數	MFCC	HLDA+MLLT
5	67.89	66.03
10	67.40	65.56
15	67.02	65.44
20	67.11	65.34
25	67.04	65.23
30	66.80	65.27

表 4-1 外場受訪者:30 次最大化相似度訓練，每間隔 5 次之自由音節辨識錯誤率(%)

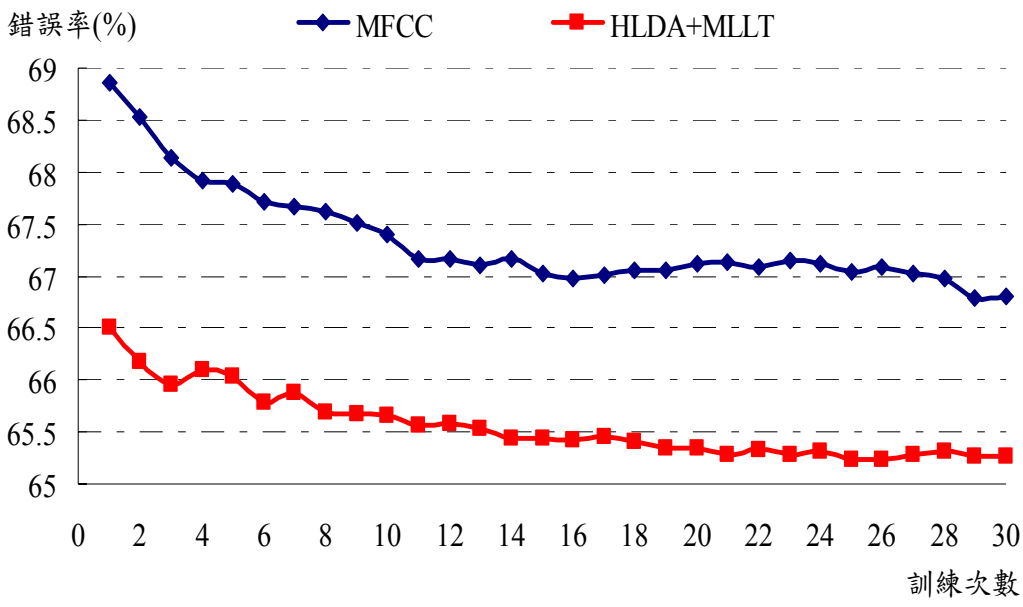

圖 4-1 外場受訪者:30 次最大化相似度訓練之自由音節辨識音節錯誤率曲線圖

語言模型權重	MFCC	HLDA+MLLT
5	62.60	57.70
6	61.65	57.11
7	61.60	56.74
8	62.00	56.85
9	62.17	57.30
10	62.42	57.52
11	63.20	58.13
12	63.47	58.70

表 4-2 外場受訪者:不同的語言模型權重，經詞彙樹複製搜尋後之字錯誤率(%)

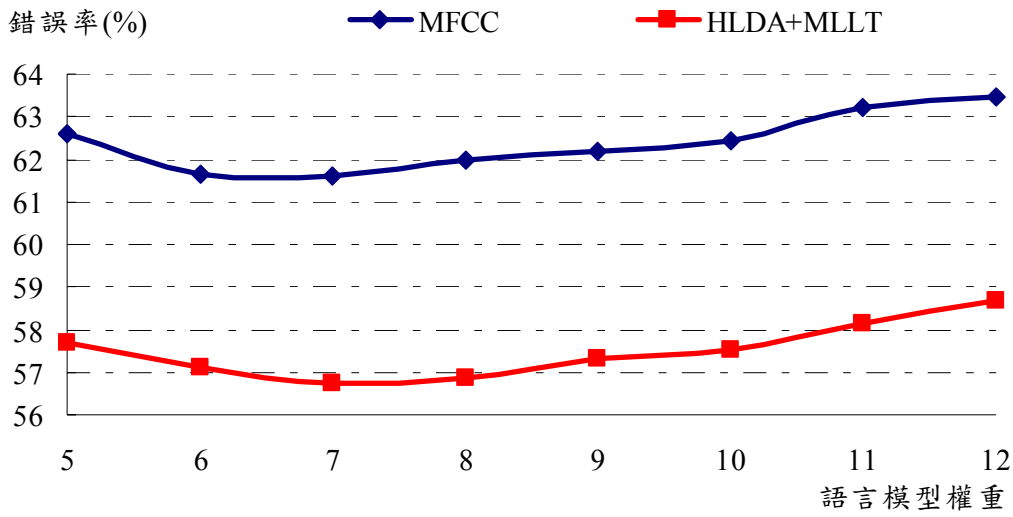

圖 4-2 外場受訪者:不同的語言模型權重，經詞彙樹複製搜尋後之字錯誤率曲線圖

語言模型權重	MFCC	HLDA+MLLT
5	60.73	56.11
6	59.71	55.21
7	59.57	55.50
8	60.14	55.75
9	60.35	55.69
10	60.73	55.77
11	61.09	56.40
12	61.20	57.00

表 4-3 外場受訪者:不同的語言模型權重，經詞圖搜尋後字錯誤率(%)

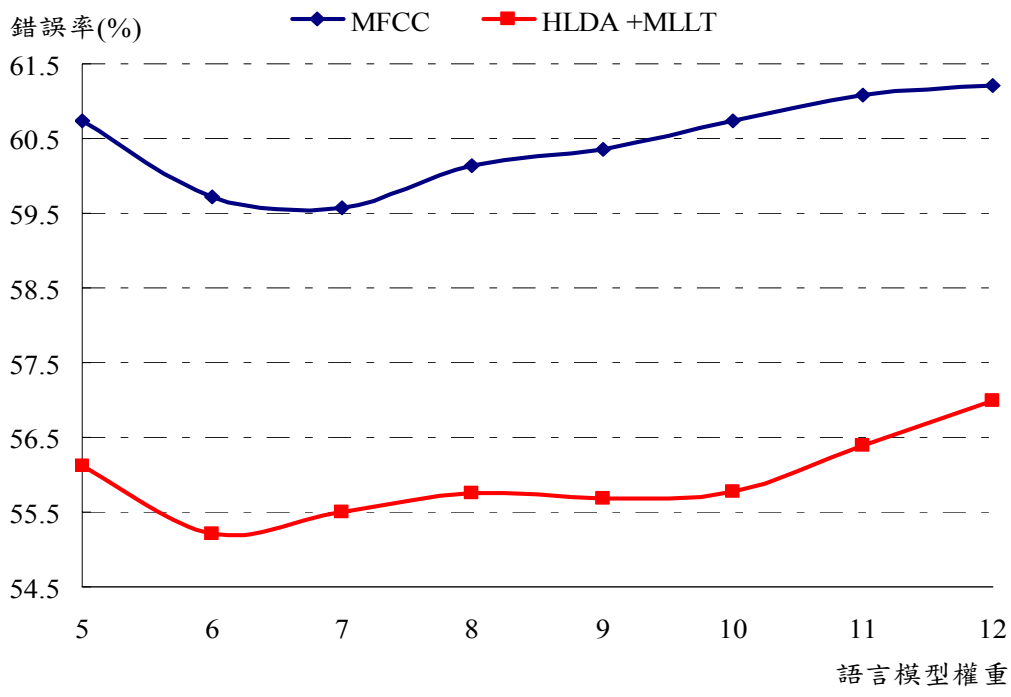

圖 4-3 外場受訪者:不同的語言模型權重，經詞圖搜尋後之字錯誤率曲線圖

【實驗討論】

在自由音節辨識實驗中，不論採用是何種語音特徵參數，最大化相似度訓練大概都在 15 次訓練次數之後呈現飽和(Saturation)。15 次之後的訓練，其錯誤率曲線會呈現上下振盪的情況，辨識錯誤率已無法有進一步明顯的下降。

在調整語言模型權重的實驗中，不論是採用詞彙樹複製搜尋或詞圖搜尋、在何種語音特徵參數，其錯誤率最低的語言模型權重皆約為 6 或 7 左右，顯示背景語言模型跟外場受訪者測試語料的領域(Domain)似乎有些不同。此外，詞圖搜尋由於用到更高階的語言模型，因此其錯誤率能比詞彙樹複製搜尋更降低一些。在往後的實驗，外場受訪者的詞彙樹複製搜尋語言模型權重將固定設 7，詞圖搜尋的權重則將設為 6。

4.1.2 最小化音素錯誤(Minimum Phone Error, MPE)訓練之實驗

在上一小節的實驗中，由於最大化相似度訓練很快就達到了飽和狀度。因此，接下來我們便再對經過30次最大化相似度訓練後的聲學模型進行10次的最小化音素錯誤訓練[Povey 2004]，觀察能否對聲學模型有所幫助。實驗結果可參照表 4-4，而對應的自由音節辨識錯誤率及字錯誤率曲線請參照圖 4-4及圖 4-5。

	自由音節辨識錯誤率		詞彙樹複製搜尋字錯誤率	
訓練次數	MFCC	HLDA+MLLT	MFCC	HLDA+MLLT
1	66.31	64.64	60.57	55.45
2	66.30	64.56	59.29	54.98
3	66.69	64.31	58.55	54.86
4	66.90	64.45	58.32	54.55
5	67.02	64.54	58.19	54.26
6	67.27	64.56	57.89	54.03
7	67.33	64.65	58.02	53.90
8	67.33	64.81	57.97	53.87
9	67.33	65.05	58.17	53.89
10	67.58	65.10	58.09	54.24

表 4-4 外場受訪者:10 次最小化音素錯誤訓練之音節與字錯誤率(%)

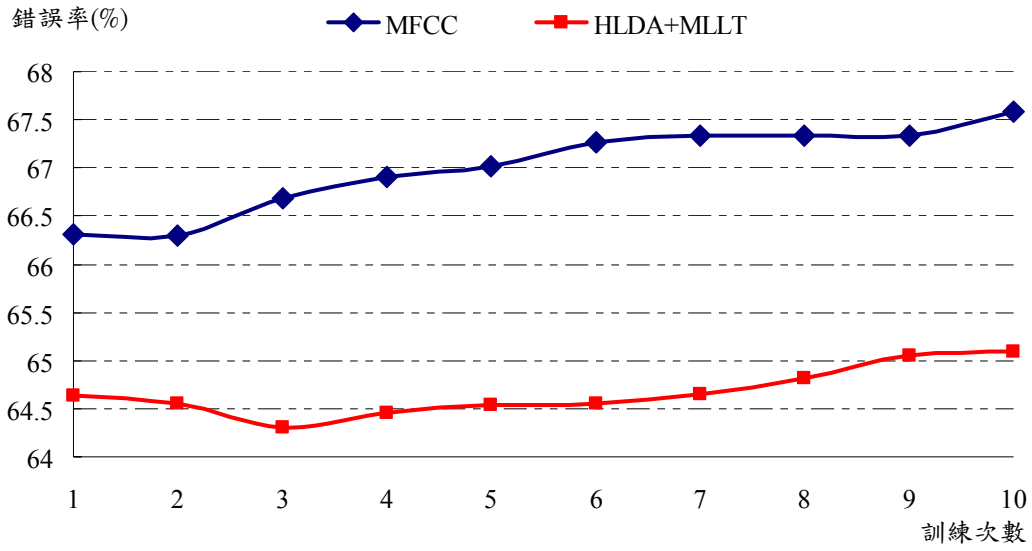

圖 4-4 外場受訪者:10 次最小化音素錯誤訓練之音節錯誤率曲線圖

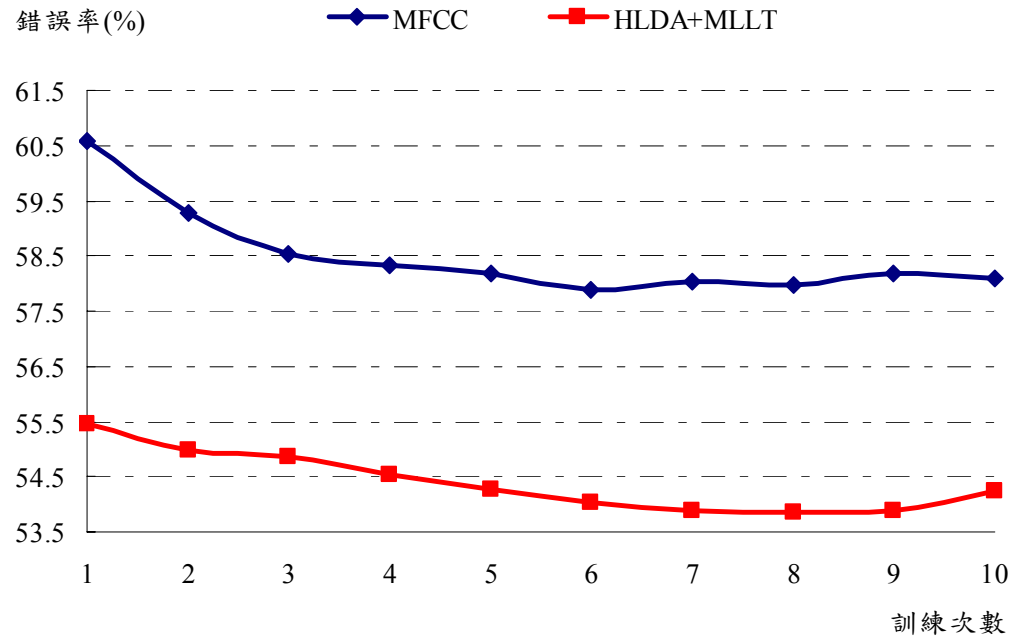

圖 4-5 外場受訪者:10 次最小化音素錯誤訓練之詞彙樹複製搜尋之字錯誤率曲線圖

【實驗討論】

在自由音節辨識實驗中，最小化音素錯誤訓練法則其作用似乎比較不明顯(其音節錯誤率約降低0.5%~1%左右)，而且很快就因為過度訓練(Over-training)而帶來錯誤率的上昇，對於辨識率較低的初始聲學模型尤其顯著，其原因可能出自辨識率較低的聲學模型無法提供較多的鑑別資訊。

而在詞彙樹複製搜尋方面的錯誤率似乎就較無自由音節辨識的問題。大致上來說，不論是何種語音特徵參數，都有2%~3%的字錯誤率絕對下降(Absolute Reduction)。可能是因為最小化音素錯誤訓練法則除了在訓練聲學模型本身會有效果之外，也會讓聲學模型在訓練時同時考慮到與語言模型結合的影響，而使得字錯誤率有較佳的結果。

在以上兩小節的實驗中，不論對最大化相似度或最小化音素錯誤訓練來說，HLDA+MLLT的語音特徵參數在自由音節辨識的音節錯誤率或詞彙樹複製搜尋的字錯誤率都較MFCC的語音特徵參數有明顯的下降，因此有關於外場受訪者實驗中，我們將採用HLDA+MLLT的語音特徵參數。另外，使用HLDA+MLLT語音特徵參數的最小化音素錯誤訓練聲學模型於詞圖搜尋的錯誤率(WG:CHAR表示)如表 4-5所示。由於在經第8次訓練後的字錯誤率最小，因此本論文最後採用第8次最小化音素錯誤訓練的聲學模型為外場受訪者的聲學模型，而在外場記者語料部份，則是之前的實驗，採用經由150次最大化相似度訓練及10次最小化音素錯誤訓練後的聲學模型做為記者語料的聲學模型, 而語音特徵參數也是使用HLDA+MLLT。

訓練次數	1	2	3	4	5
WG:CHAR	54.92	54.37	53.85	53.21	52.79
訓練次數	6	7	8	9	10
WG:CHAR	52.38	52.30	52.29	52.47	52.55

表 4-5 外場受訪者:10 次最小化音素錯誤率訓練詞圖搜尋之字錯誤率(%，語音特徵參數為 HLDA+MLLT)

4.1.3 相同領域與背景語言模型線性插補實驗

由於外場受訪者是屬於偏即性口語對話語料，單靠中央通訊社的新聞背景語料似乎有所不足，因此，本論文額外使用了外場受訪者聲學訓練語料的人工轉寫文字檔共2,002句及從漢語連續口語對話語音語料庫抽出的1,791句，合併成3,793句的相同領域語言模型訓練資料。將得到的相同領域語言模型與中央通訊社背景語言模型做線性插補，而線性插補的公式如式(4-2)所示:

$$P(W) = \alpha \cdot P_{BG}(W) + (1-\alpha)P_{InDmain}(W) \tag{4-2}$$

其中α代表背景語言模型$P_{BG}(W)$的權重(其值介於0~1之間)，$P_{InDomain}(W)$則是代表相同領域語言模型的分數。觀察對詞圖搜尋錯誤率的的影響，其實驗結果可參考表4-6。其中MATBN_IV_LM代表單單只使用外場受訪者聲學訓練語料的文字檔進行線性插補，而MCDC_MATBN_IV_LM則代表將外場受訪者聲學模型訓練語料的文字檔與漢語連續口語對話語音語料庫合併後進行線性插補。而詞圖搜尋字錯誤率曲線圖可參考圖 4-6。

插補權重α	MABN_IV_LM	MCDC_MATBN_IV_LM
0.1	50.91	50.85
0.2	50.41	50.36
0.3	50.35	50.16
0.4	49.72	50.71
0.5	49.72	49.60
0.6	49.73	49.69
0.7	49.75	49.56
0.8	50.06	50.21
0.9	50.48	50.38
1	52.40	52.19

表 4-6 相同領域語言模型與背景語言模型做線性搜補之詞圖搜尋字錯誤率(%)

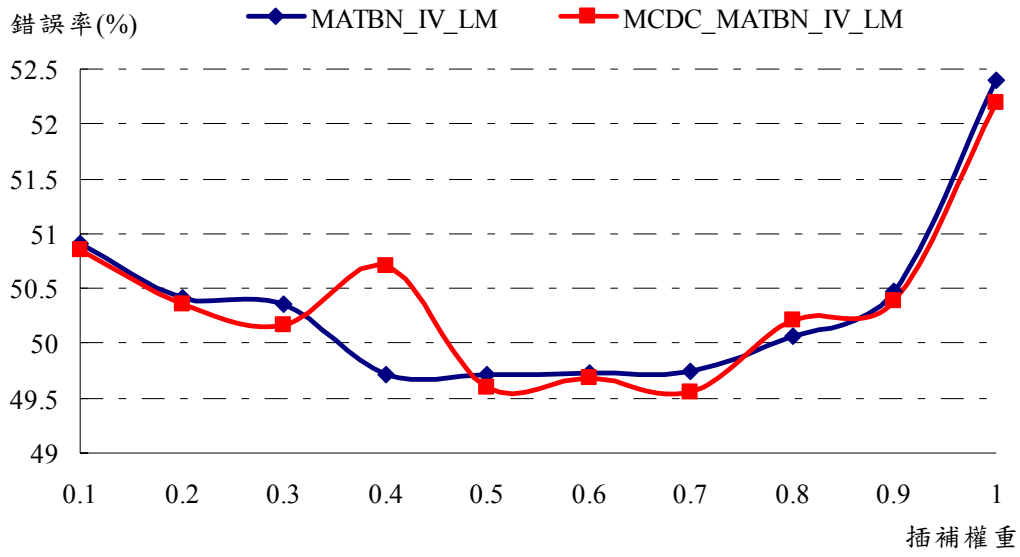

圖 4-6 相同領域語言模型與背景語言模型做線性插補之詞圖搜尋字錯誤率曲線圖

【實驗討論】

將相同領域的語言模型與背景語言模型做線性插補後，使得新的語言模型能更符合外場受訪者的情況，不論是單單使用外場受訪者聲學訓練語料的文字檔或是將外場受訪者聲學訓練語料的文字檔與漢語連續口語對話語音語料庫合併後進行線性插補，對減低詞圖搜尋的字錯誤率皆有一定的效果，根據表 4-6可得知最好的結果為使用兩套文字語料合併的效果最好，當插補權重α設為0.7時，字錯誤率為49.56%。

4.2 傳統信心度評估之實驗

本小節分為兩部份:4.2.1小節討論以特徵為基礎之信心度評估;4.2.2小節則為討論事後機率之信心度評估。而信心度評估的對象分別為外場記者與受訪者兩套語料庫其語音辨識系統字錯誤率基礎實驗結果中個別的最佳辨識詞序列中之每一個詞。其中MATBN_R代表外場記者語料，而MATBN_IV則代表受訪者語料。這兩套語料其信心度評估之基礎實驗結果(Baseline)可參照表 4-7。

MATBN_R	MATBN_IV
24.52	51.97

表 4-7 外場記者與受訪者信心度評估之信心度錯誤率(%)基礎實驗結果

4.2.1 以特徵為基礎之信心度評估實驗

在此實驗中，我們主要分別使用2.1小節所提到的聲學穩定度(AS)及候選詞假設密度(HD)兩種傳統較常使用的特徵，或是將此兩種特徵合併使用(HD+AS)，搭配2.1小節所提到的自然貝氏分類器來對外場記者與受訪者兩套語料庫做信心度評估。其實驗結果可參照表 4-8。

	MATBN_R	MATBN_IV
HD	23.34	33.08
AS	21.25	31.74
HD+AS	21.08	30.50

表 4-8 以特徵為基礎之信心度評估之信心度錯誤率(%)

【實驗討論】

由實驗結果可得知，不論是聲學穩定度或候選詞假設密度在兩套語料皆能有效的降低信心度錯誤率，當進一步採用自然貝氏分類器的假設合併此兩項預估特徵時，更能降低信心度錯誤率，相較於表 4-7，表 4-8的最佳的結果對外場記者及受訪者測試語料分別有14.03%，41.31%的信心度錯誤率相對下降(Relative Reduction)。

4.2.2 事後機率之信心度評估實驗

我們針對 2.2.2 小節中式(2-17)，也就是以一般傳統的事後機率為辨識詞的信心度，以及 2.2.3 小節中的式(2-18)至(2-20)三種不同的信心度評估進行實驗比較。在計算事後機率的時候，由於語言模型的分數為介於 0~1 之間的值，但聲學分數的區間則是 0 至無窮大。所以我們通常在計算傳統的事後機率時，會使用一個權重$\kappa(\kappa>1)$來拉近聲學與語言模型之間分數的比例，如將式(2-17)修改為:

$$P(a:[w_a;s_a,e_a]\,|\,\Psi^X)=\frac{\sum_{\{\overline{W}:[w^n;s^n,e^n]_{n=1}^N\}\in\Psi^X,a\subset\overline{W}}\left\{\prod_{n=1}^{N}p(x_{s^n}^{e^n}\,|\,w^n)^{1/\kappa}\cdot P(w^n\,|\,h^n)\right\}}{\sum_{\{\overline{W}':[w^m;s^m,e^m]_{m=1}^M\}\in\Psi^X}\left\{\prod_{m=1}^{M}p(x_{s^m}^{e^m}\,|\,w^m)^{1/\kappa}\cdot P(w^m\,|\,h^m)\right\}} \tag{4-3}$$

$1/\kappa$代表壓縮聲學模型分數，使之分數區間能較接近語言模型的分數區間。在本實驗中，首先試著調整κ的值，使得式(4-3)在測試語料的信心度錯誤率為最低。實驗結果可以參照表 4-9，對應的信心度錯誤率曲線圖可見圖 4-7及圖 4-8。

κ	MATBN_R	MATBN_IV
5	24.13	37.46
6	24.31	36.38
7	24.18	34.18
8	23.65	32.92
9	22.80	32.45
10	22.32	32.55
11	22.18	31.31
12	22.43	31.31

表 4-9 使用不同κ值計算事後機率之信心度錯誤率(%)

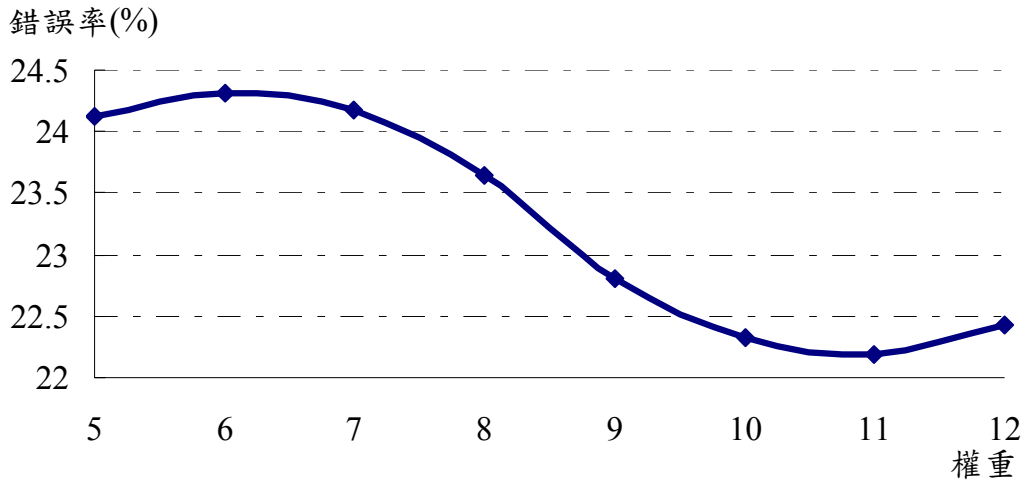

圖 4-7 外場記者:使用不同κ值計算事後機率所獲得的信心度錯誤率曲線圖

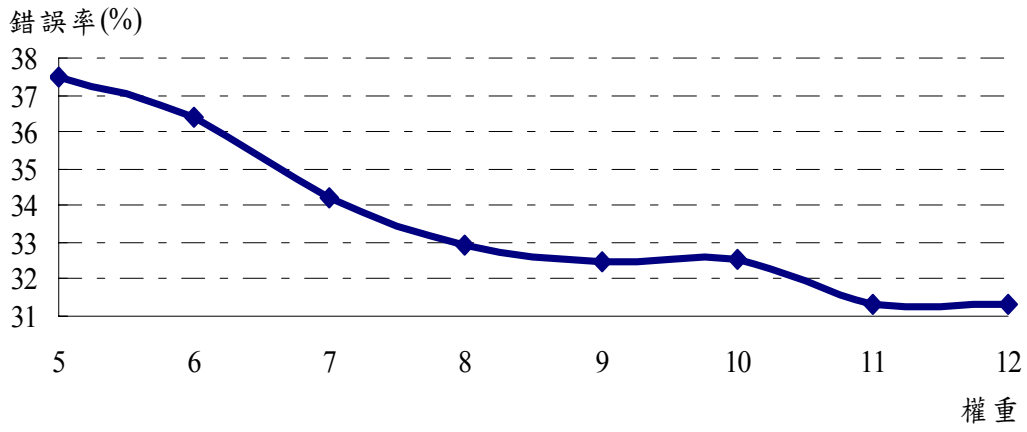

圖 4-8 外場受訪者:使用不同κ值計算事後機率所獲得的信心度錯誤率曲線圖

根據表 4-9的結果，我們進一步比較式(2-17)至(2-20)四種信心度評估之效果。實驗結果可參考表 4-10。其中Cnormal代表式(2-17)，也就是傳統事後機率的算法，而外場記者與受訪者在計算式(2-17)的事後機率時聲學分數權重κ根據表 4-9皆設為11，使其在評估語料有最低的信心度錯誤率。

	MATBN_R	MATBN_IV
Cnormal	22.18	31.31
Csec	21.47	32.32
Cmed	21.47	32.32
Cmax	21.47	32.32

表 4-10 不同的辨識詞事後機率方法之信心度錯誤率(%)

【實驗討論】

由實驗結果可得知，雖然Cmed、Csec及Cmax三種信心度評估雖然跟基礎實驗結果比較有相當的進步，但是與傳統的事後機率方法相比，並不一定會有較佳的效果。其主要原因可能在於外場受訪者測試語料本身的辨識率較低，造成當使用Cmed、Csec及Cmax時，大部份都是加大錯誤的詞信心度，反而造成信心度錯誤率因此升高。

4.3 信心度評估應用於降低詞圖搜尋錯誤率之實驗

此實驗主要分為兩個部份: 4.3.1小節討論關於運用傳統的事後機率於降低詞圖搜尋之錯誤率;4.3.2小節則是探討最小化音框錯誤率對於詞圖搜尋的影響。外場記者與受訪者字錯誤率的基礎實驗結果可參考表 4-11。

MATBN_R	MATBN_IV
20.79	49.56

表 4-11 外場記者與受訪者語料經詞圖搜尋後之字錯誤率(%)，此為基礎實驗結果

4.3.1 運用事後機率降低詞圖搜尋錯誤率之實驗

根據2.5.3小節中的式(2-43)，必須先求得詞圖中每個詞段的事後機率。我們根據4.2.2小節的實驗，聲學分數權重κ設為11去計算每個詞段的事後機率，再進行式(2-43)的詞圖搜尋，其實驗結果可參照表 4-12。

MATBN_R	MATBN_IV
20.68	47.60

表 4-12 外場記者與受訪者語料經事後機率詞圖搜尋後之字錯誤率(%)

【實驗討論】

由實驗結果可得知，如果事後機率的信心度估評在某個語料其信心度錯誤率下降越明顯，其運用在降低詞圖插尋錯誤率有較佳的效果。如在本實驗中，外場受訪者測試語料就有較佳的進步。

4.3.2 最小化音框錯誤詞圖搜尋之實驗

本實驗主要是探討2.5.4小節所討論的式(2-46)，其中α的部份我們嘗試設0~0.1的區間，間隔為0.01，來代表是否要強調長詞。實驗結果可參考表 4-13，而對應的錯誤率曲線請參考圖 4-9及圖 4-10。

α	MATBN_R	MATBN_IV
0	20.81	47.35
0.01	20.68	47.45
0.02	20.67	47.74
0.03	20.61	47.87
0.04	20.62	48.01
0.05	20.60	48.16
0.06	20.57	48.19
0.07	20.56	48.37
0.08	20.58	48.44
0.09	20.59	48.57
0.1	20.57	48.59

表 4-13 運用最小化音框錯誤率於詞圖搜尋之字錯誤率(%)

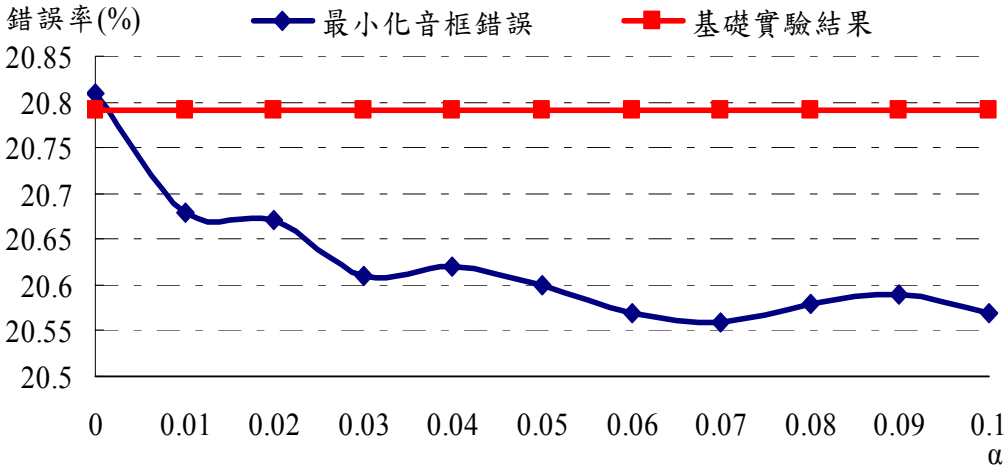

圖 4-9 外場記者:運用最小化音框錯誤率於詞圖搜尋之字錯誤率曲線圖

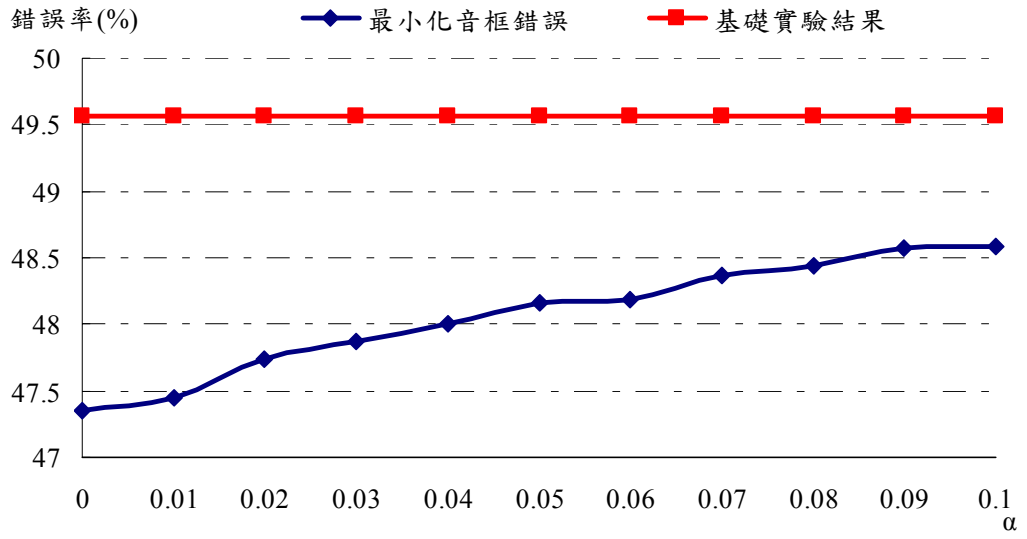

圖 4-10 外場受訪者:運用最小化音框錯誤率於詞圖搜尋之字錯誤率曲線圖

【實驗討論】

由實驗結果可得知，在外場記者的評估語料中，最好的結果相較於基礎實驗結果可以有 1.11%的相對進步。而在外場受訪者的部份則更可以有 4.56%的相對進步。此方法似乎在語音辨識系統的正確率較低的情況之下，有較大的進步的空間。另外，從實驗可觀察出外場記者的語料似乎有出現少許的長詞。而受訪者的部份，因為α越大，則錯誤率越高，有偏向於短詞較多的情形。