第四章　基礎實驗結果

4.1　查表式統計圖等化法的深入探討

本小節針對查表式統計圖等化法（Table Based Histogram Equalization，THEQ），進行各方面的探討，最終目的是討論各因素對辨識率的影響。包括了(1)參考模型的來源，(2)表格大小的影響，(3)作用在能量維的改善程度，(4)以及作用在頻譜和倒頻譜上的優劣。

4.1.1 不同參考模型的比較

一般來說傳統的查表式統計圖等化法要建立表格時，參考模型的來源有兩種。第一種是參考全部訓練語料所累積的資訊來建立表格。第二種是參考標準高斯分佈來建立表格。不過無論是前者或後者，我們均須同時對於個別的訓練語料以及測試語料進行等化法，目的在於希望每個單位（Sentence-Wise）的語料，在機率密度函數上都有相似於參考模型的分佈。而本實驗吾人採用的是華語廣播新聞的資料庫，且辨識結果採用音節辨識率。表 4.1.1.1 是吾人針對不同參考模型建立出的表格所進行等化法後的實驗結果：

表 4.1.1.1	Baseline	Gaussian_Based	Training_Data_Based
3	53.35	55.61 / 4.24	54.89 / 2.89
6	53.64	56.47 / 5.28	56.11 / 4.60
9	54.19	56.62 / 4.54	56.31 / 3.91
12	54.22	56.69 / 4.56	56.30 / 3.84
15	54.63	56.95 / 4.25	56.56 / 3.53
18	54.64	57.83 / 5.84	56.55 / 3.50
AVG	-	4.79	3.71

表 4.1.1.1 針對不同參考模型建立出的表格所進行等化法後的實驗結果

第一行代表的是 HMM 模型訓練的次數。第二行代表的是原始語音辨識的結果。第三行代表的是利用標準高斯分佈建立出的表格後，所進行查表式統計圖等化法的結果。第四行則是利用所有訓練語料建立出的表格後，所進行查表式統計圖等化法的結果。而第八列代表的是兩種方法相對於 baseline 在不同訓練次數的模型下的平均改善程度。表格中除第二行外，每一格的數據所代表的意義是，(音節辨識率/相對於 baseline 改善程度的百分比)。

表 4.1.1.1 中我們可以清楚的觀察到，無論 HMM 模型訓練次數的多寡，若查表式統計圖等化法採用的是標準高斯分佈所建立的表格，效果均比利用所有訓練語料所建立出的表格來的好。而平均來說，相對於 Baseline 的結果，前者的表現更比後者多了超過 1 個百分點的改善。

4.1.2 不同表格大小的比較

上個小節吾人所討論的是，不同的參考模型對查表式統計圖等化法的影響。這個小節吾人想探討的是，表格大小對查表式統計圖等化法的影響。所謂表格的大小，指的是所建立出的表格中點數的多寡。吾人認為無論是何種資料來源所產生的機率密度空間分佈，當我們在該分佈上採樣的點越多，就越能充分的表達出該分佈的特性，這好比用 24-bit 的色彩和 256 色來顯示同一張圖片的結果優劣的道理是類似的。因此以下的實驗，吾人試圖在標準高斯分佈上，各採取 1000 點、

10000 點和 20000 點來建立表格，探討不同表格大小對查表式統計圖等化法的影響。本實驗吾人採用的也是華語廣播新聞的資料庫。表 4.1.2.1 是實驗結果：（註：4.1.1 節的實驗是採樣 1000 點的實驗結果）

表 4.1.2.1	1000 points	10000 points	20000 points
3	55.61 / 4.24	56.39 / 5.70	55.91 / 4.80
6	56.47 / 5.28	57.13 / 6.51	56.52 / 5.37
9	56.62 / 4.54	57.37 / 5.87	56.61 / 4.47
12	56.69 / 4.56	57.59 / 6.21	56.75 / 4.67
15	56.95 / 4.25	57.81 / 5.82	57.02 / 4.38
18	57.83 / 5.84	57.75 / 5.69	57.25 / 4.78
AVG	4.79	5.97	4.75

表 4.1.2.1 針對不同表格大小所進行等化法後的實驗結果

第一行代表的是 HMM 模型訓練的次數。第二行代表的是在標準高斯分佈上採樣 1000 點建立表格的實驗結果。第三行代表的是在標準高斯分佈上採樣 10000 點建立表格的實驗結果。第四行代表的是在標準高斯分佈上採樣 20000 點建立表格的實驗結果。而第八列代表的是三種實驗相對於 baseline 在不同訓練次數的模型下的平均改善程度。表格中每一格的數據所代表的意義是，(音節辨識率/相對於 baseline 改善程度的百分比)。

表 4.1.2.1 中我們可以清楚的觀察到，無論 HMM 模型訓練次數的多寡，大致上來說，使用採樣 10000 點的表格其表現是比較好的。以採樣點 1000 和採樣點 10000 作比較，採樣較多的確能讓辨識率較好。以採樣點 10000 和採樣點 20000 作比較，採樣更多的點卻沒有讓辨識率更好，可能的原因是表格的點數過多，以至於等化法作用過後的累積密度分佈不夠平滑。換句話說，相對於 Baseline 的結果，採樣點 10000 平均而言更比其他的採樣點數多了近 1.2 個百分比的改善。圖 4.1.2.1 是針對前兩小節對查表式統計圖等化法作交叉比對的比較圖，本實驗吾人只以實驗結果較好的採樣點 1000 與採樣點 10000 來作比較：

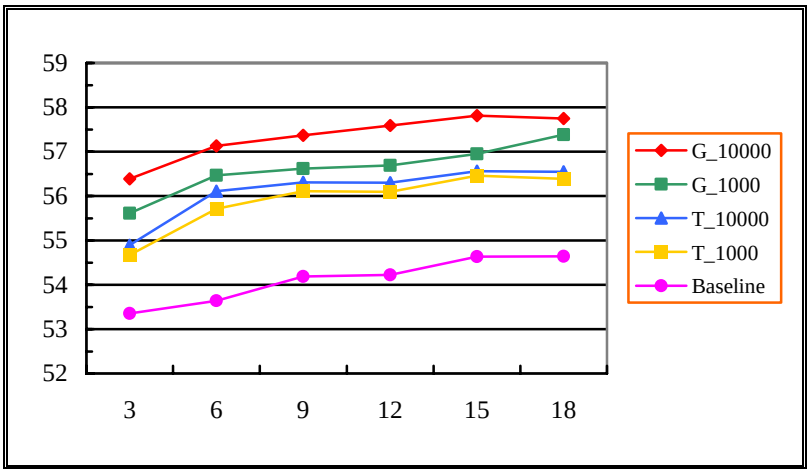

圖 4.1.2.1 不同參考分佈和不同表格大小的交叉比較實驗結果

橫軸代表的是 HMM 模型訓練的次數，縱軸則是音節辨識率。G_10000 和 G_1000 代表的是採用標準高斯分佈作為參考分佈，並且分別採樣 10000 點和 1000 點建立表格。而 T_10000 和 T_1000 代表的是採用所有訓練語料的統計為參考分佈，並且分別採樣 10000 點和 1000 點建立表格。Baseline 則是原始辨識率。

圖 4.1.2.1 中我們可以清楚的觀察到，當參考模型為標準高斯分佈，且採樣點為 10000 點時，查表式統計圖等化法可以發揮最大的效果。

4.1.3 作用在能量維與否的比較

一般都認為，傳統語音特徵參數的第 13 維度，也就是能量維，在不同的音素間扮演了相當重要的角色。因此吾人認為若查表式統計圖等化法，也能在這個維度上進行等化，勢必也會對辨識率有所提升。而在前兩個小節的實驗，吾人也是基於這樣的假設來探討查表式統計圖等化法在其他方面的特性。而本小節吾人想驗證的，便是對能量維進行等化法，是否會產生更好的辨識率。本實驗的參考模型採用的是標準高斯分佈，取樣點均為 10000 點。實驗結果如表 4.1.3.1 所示：

表 4.1.3.1	only for 13th dim	for 12 dims	for 13 dims
3	54.57 / 2.29	54.01 / 1.23	56.39 / 5.70
6	55.04 / 2.61	54.54 / 1.68	57.13 / 6.51
9	55.42 / 2.27	55.02 / 1.53	57.37 / 5.87
12	55.47 / 2.31	54.94 / 1.33	57.59 / 6.21
15	55.47 / 1.54	55.29 / 1.20	57.81 / 5.82
18	55.40 / 1.39	55.36 / 1.32	57.75 / 5.69
AVG	2.07	1.38	5.97

表 4.1.3.1 對能量維進行等化法與否的實驗結果

第一行代表的是 HMM 模型訓練的次數。第二行代表的是對前 12 維進行等化法的實驗結果。第三行代表的是對前 13 維進行等化法的實驗結果。第四行代表的是只對第 13 維進行等化法的實驗結果。而第八列代表的是三種方法相對於 baseline 在不同訓練次數的模型下的平均改善程度。表格中每一格的數據所代表的意義是，(音節辨識率/相對於 baseline 改善程度的百分比)。

實驗結果證明，對能量維進行等化法會帶來更好的辨識率的假設是成立的，並且這樣的結果，並不會隨著模型訓練次數的多寡而有所改變。

4.1.4 作用在頻譜和倒頻譜上的比較

前三個小節所探討的內容，均是探討統計圖等化法作用在倒頻譜上的結果。而本小節所要探討的是該方法作用在頻譜上的效果，也就是對對數梅爾濾波器組（Log Mel-scale Filter Banks）的輸出做等化法，進而比較作用在此二不同處上的差別。本實驗和先前的小節相同，採用的也是華語廣播新聞的資料庫，採樣點為 10000 點，參考分佈為先前實驗提到的兩類。表 4.1.4.1 是實驗結果：

表 4.1.4.1	FB-G	FB-T	CEP-G	CEP-T
3	49.23 -7.72	55.21 / 3.49	56.39 / 5.70	54.67 / 2.47
6	50.83 -5.24	56.22 / 4.81	57.13 / 6.51	55.71 / 3.86
9	51.56 -4.85	55.98 / 3.30	57.37 / 5.87	56.11 / 3.54
12	51.74 -4.57	56.01 / 3.30	57.59 / 6.21	56.09 / 3.45
15	52.07 -4.69	56.34 / 3.13	57.81 / 5.82	56.46 / 3.35
18	52.33 -4.23	56.44 / 3.29	57.75 / 5.69	56.39 / 3.20
AVG	-5.22	3.55	5.97	3.31

表 4.1.4.1 不同參考分佈的統計圖等化法作用在頻譜和倒頻譜的實驗結果

第一行代表的是 HMM 模型訓練的次數。FB-G 代表的是統計圖等化法參考分佈為標準高斯分佈，且作用在頻譜上。FB-T 代表的是統計圖等化法參考分佈為所有訓練語料，且作用在頻譜上。CEP-G 代表的是統計圖等化法參考分佈為標準高斯分佈，且作用在倒頻譜上。CEP-T 代表的是統計圖等化法參考分佈為所有訓練語料，且作用在倒頻譜上。而第八列代表的是四種方法相對於 baseline 在不同訓練次數的模型下的平均改善程度。表格中每一格的數據所代表的意義是，(音節辨識率/相對於 baseline 改善程度的百分比)。

由表 4.1.4.1 可以觀察到，當參考分佈為所有訓練語料時，查表式統計圖等化法無論是作在頻譜上或是倒頻譜上都有不錯的效果。而當參考分佈為標準高斯分佈時，只有作在倒頻譜上才有較好的效果。圖 4.1.4.1 是對前四個小節所做的綜合比較：

由圖 4.1.4.2 我們可以對 4.1 節歸納出三個結論。大致上而言，無論參考分佈為何，採樣的點數越多（但是也不宜過多）對於辨識率的提升越有幫助。而若採用標準高斯分佈為參考分佈，效果也都比採用所有訓練語料為參考分佈來的好。最後，在採用標準高斯分佈為參考分佈，且採樣點為 10000 點的前提之下，查表式統計圖等化法作用在倒頻譜上的結果是最佳的。而接下來的實驗也是基於這個前提，來探討查表式統計圖等化法和其他語音強健技術的優劣，以及其本身的改良。

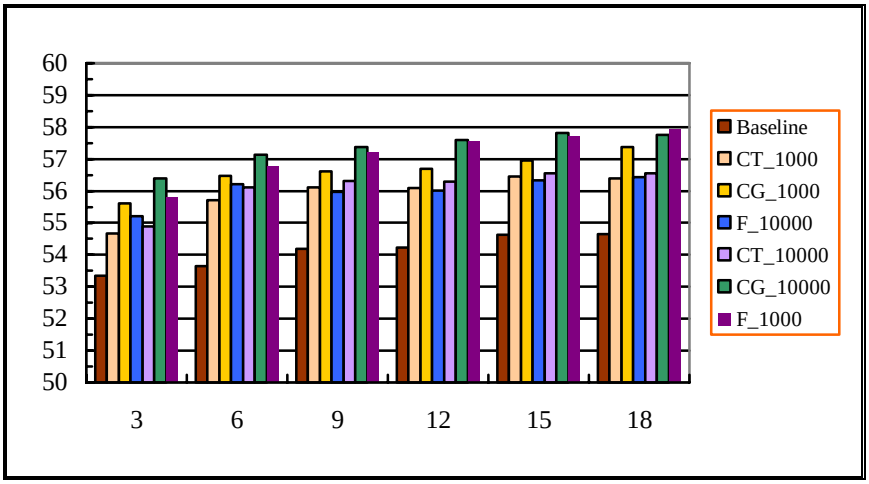

圖 4.1.4.2 各種條件下的交叉比較結果

本圖中橫軸代表的是 HMM 模型訓練的次數，縱軸則是音節辨識率。F_10000 和 F_1000 代表的是查表式統計圖等化法作用在頻譜上的實驗結果，參考分佈為所有訓練語料，並且分別於參考分佈中採樣 10000 點和 1000 點建立表格。而 CG_10000 和 CG_1000 代表的是查表式統計圖等化法作用在倒頻譜上的實驗結果，參考分佈為標準高斯分佈，並且分別於參考分佈中採樣 10000 點和 1000 點建立表格。而 CT_10000 和 CT_1000 代表的是查表式統計圖等化法作用在倒頻譜上的實驗結果，參考分佈為所有訓練語料，並且分別於參考分佈中採樣 10000 點和 1000 點建立表格。Baseline 則是原始辨識率。

4.2 兩種統計圖等化法的比較

第一小節主要是以查表式統計圖等化法做為探討對象，事實上分位差統計圖等化法，也被認為是對於提升辨識率非常有用的技術之一。不同於查表式統計圖等化法需要從參考模型的分佈裡，取出大量的採樣點建立表格，分位差統計圖等化法只需要從參考模型取出極少量的採樣點，並加上非線性轉換方程式的轉換，使得存在於訓練語料和測試語料間的不匹配現象可以降低。因此本小節的實驗主軸，便是比較此兩種統計圖等化法的優劣。

4.2.1 作用在倒頻譜上的比較

上個小節的實驗有個初步的結論，那就是查表式統計圖作用在倒頻譜上和頻譜上的效果，幾乎是在伯仲之間。而這個小節吾人先將此二種統計圖等化法先作用在倒頻譜上來比較其優劣。我們用 THEQ 來代表查表式統計圖等化法，而用 QHEQ 來代表分位差統計圖等化法（參照本論文 2.2 節）。本實驗 QHEQ 取 4 個分位差的值。表 4.2.1.1 是實驗結果：

表 4.2.1.1	QHEQ	THEQ
3	50.48 / -5.38	56.39 / 5.70
6	51.75 / -3.52	57.13 / 6.51
9	52.29 / -3.50	57.37 / 5.87
12	52.32 / -3.50	57.59 / 6.21
15	52.55 / -3.81	57.81 / 5.82
18	52.46 / -3.99	57.75 / 5.69
AVG	-3.95	5.97

表 4.2.1.1 兩種統計圖等化法作用在倒頻譜上的比較

第一行代表的是 HMM 模型訓練的次數。第二行代表的是分位差統計圖等化法作在倒頻譜上的實驗結果。第三行是查表式統計圖等化法作在倒頻譜上的實驗結果。。而第八列代表的是兩種方法相對於 baseline 在不同訓練次數的模型下的平均改善程度。表格中除每一格的數據所代表的意義是，(音節辨識率/相對於 baseline 改善程度的百分比)。

表 4.2.1.1 的實驗結果顯示，分位差統計圖等化法並不如查表式統計圖等化法，在倒頻譜上能夠有不錯的表現，甚至辨識結果比原始辨識率還差。

4.2.2 作用在頻譜上的比較

有鑑於上個小節的實驗結果，在本小節吾人試圖將此二種統計圖等化法作用在頻譜上，來比較其優劣。本實驗 QHEQ 取 4 個分位差的值。表 4.2.2.1 是實驗結果：

表 4.2.2.1	QHEQ	THEQ
3	55.69 / 4.39	55.81 / 4.61
6	56.37 / 5.09	56.80 / 5.89
9	56.20 / 3.71	57.23 / 5.61
12	56.24 / 3.73	57.56 / 6.16
15	56.41 / 3.26	57.72 / 5.66
18	56.31 / 3.06	57.93 / 6.02
AVG	3.87	5.16

表 4.2.2.1 兩種統計圖等化法作用在頻譜上的比較

第一行代表的是 HMM 模型訓練的次數。第二行代表的是分位差統計圖等化法作在頻譜上的實驗結果。第三行是查表式統計圖等化法作在頻譜上的實驗結果。而第八列代表的是兩種方法，相對於 baseline 在不同訓練次數的模型下的平均改善程度。表格中每一格的數據所代表的意義是，(音節辨識率/相對於 baseline 改善程度的百分比)。

表 4.2.2.1 的實驗結果說明了 QHEQ 不同於 THEQ 作用在倒頻譜上或頻譜上都能有不錯的效果，QHEQ 只有作用於頻譜上才能提升辨識率；然而 QHEQ 提升辨識率的程度仍然比 THEQ 差平均而言低了 1.3 個百分點。圖 4.2.2.1 是吾人對這兩個小節的實驗所做的綜合比較：

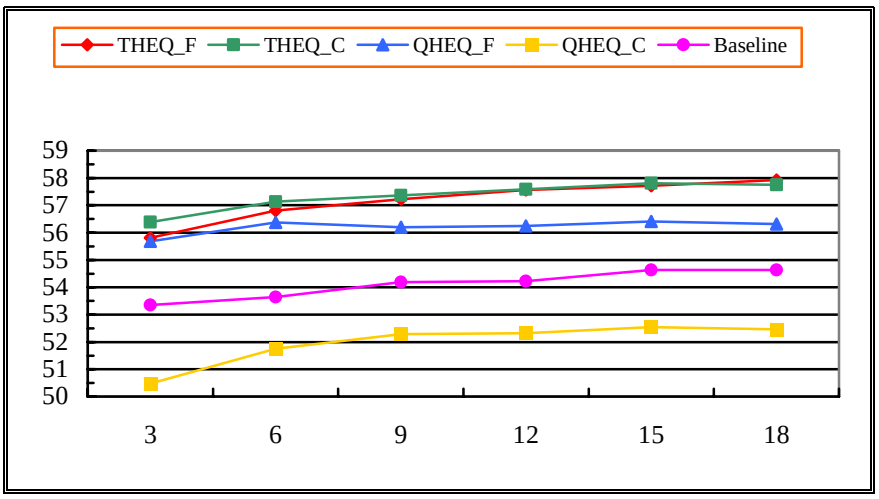

圖 4.2.2.1 兩種統計圖等化法的綜合比較

本圖中橫軸代表的是 HMM 模型訓練的次數，縱軸則是音節辨識率。THEQ_F 和 THEQ _C 代表查表式統計圖等化法分別作用在頻譜上和倒頻譜上的實驗結果。而 QHEQ_F 和 QHEQ_C 代表分位差統計圖等化法作用在頻譜上和倒頻譜上的實驗結果。Baseline 則是原始辨識率。

圖 4.2.2.1 中我們可以觀察到，無論是作用在頻譜上或是倒頻譜上，查表式統計圖等化法對於抗噪音的能力均表現的比分位差式統計圖等化法來的好，也因此辨識率相較之下也都比較高。

4.3 強健性語音參數技術的合併

近年來有許多研究是偏向於擷取出強健性的語音參數，也就是較具有抗噪音能力的特徵參數，來提升辨識率。本小節的實驗方向，就是對於其他強健性語音參數的技術做研究，並與上個小節的查表式統計圖等化法合併，目的是能擷取出更具有抗噪音能力的特徵參數。

4.3.1 高階倒頻譜正規化法

傳統的倒頻譜平均消去法或是倒頻譜正規化法，分別對特徵參數的第一動差（First Order Moment）和第二動差（Second Order Moment）進行正規化，其目的都是希望可以降低存在於模型和語料間不匹配的現象。最近有相關研究提出，若能對更高階的動差進行正規化，能讓不匹配的現象降到更低，因此本小節的實驗主軸在於比較對不同階的動差進行等化法後的現象。圖 4.3.1.1 是實驗結果：

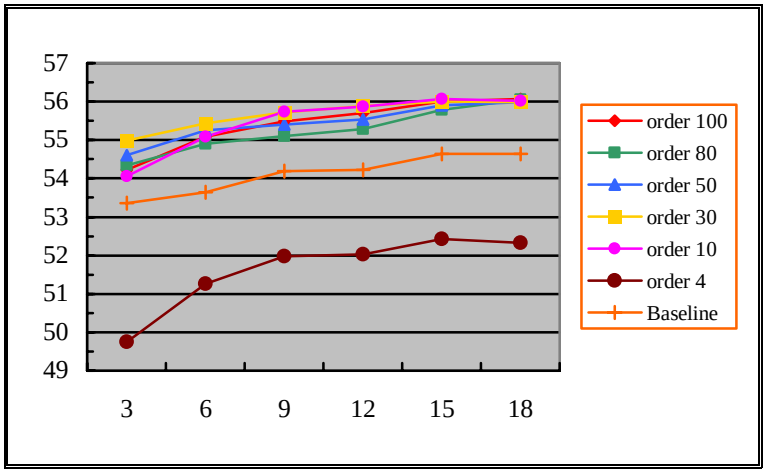

圖 4.3.1.1 對不同階動差做正規化的比較結果

本圖中橫軸代表的是 HMM 模型訓練的次數，縱軸則是音節辨識率。order 後的數字是代表對不同階的動差進行正規化。Baseline 則是原始辨識率。

由圖 4.3.1.1 中我們可以觀察到，除了對第四階的動差作正規化得到的效果不理想外，對其他階的動差進行正規化後辨識率都比 Baseline 來的好。同時我們也發現，辨識率並沒有隨著對更高階的動差進行正規化而更加的提升，反而有收斂的現象。此外，隨著模型訓練次數的增加，對不同階的動差進行正規化後的差異也越來越小。接下來吾人針對階數在 10 階以上的動差進行更細微的比較，同時本次實驗也加入倒頻譜正規化法（隊第二階動差做正規化），並人工加上 AURORA 2.0 所提供的八種噪音，伴隨著不同的訊噪比作比較。本實驗 HMM 模型採用訓練次數為 18。實驗結果如表 4.3.1.1 所示：

	CN (order=2)	order 10	order 30	order 50	order 80	order 100
Clean	57.12	56.02	55.98	55.99	56.05	56.07
20dB	55.88	55.15	55.31	55.27	55.44	55.49
15dB	51.19	50.92	50.94	51.10	51.37	51.20
10dB	40.80	41.32	41.33	41.62	42.34	41.27
5dB	24.04	24.82	23.84	24.28	26.04	24.18
0dB	1.93	5.63	3.59	3.70	6.20	3.79
-5dB	-9.43	-2.98	-5.06	-6.10	-3.78	-5.97
AVG	34.77	35.57	35.00	35.19	36.28	35.19

表 4.3.1.1 (a) 在噪音環境下對不同階動差做正規化的實驗結果

第一行代表的是不同程度的訊噪比。第二行代表的是經過倒頻譜正規化法的結果。第三至第七行分別是對第 10、30、50、80、100 階動差做正規化法的實驗結果。而第九列代表的是 0dB 至 20dB 的平均辨識率。

	CN (order=2)	order 10	order 30	order 50	order 80	order 100
Clean	4.54	2.53	2.45	2.47	2.58	2.62
20dB	11.32	9.79	10.12	10.04	10.42	10.53
15dB	21.33	20.71	20.71	21.12	21.84	21.42
10dB	47.04	48.92	48.93	50.03	52.63	48.74
5dB	260.96	272.67	257.96	264.57	290.99	263.06
0dB	118.56	154.14	134.52	135.58	159.62	136.44
-5dB	32.88	78.79	63.99	56.58	73.10	57.51
AVG	70.95	83.94	76.95	77.20	87.31	77.19

表 4.3.1.1 (b) 在噪音環境下對不同階動差做正規化的相對改善辨識率

第一行代表的是不同程度的訊噪比。第二行代表的是經過倒頻譜正規化法的結果，相對於 baseline 的改善程度。第三至第七行分別是對第 10、30、50、80、100 階動差做正規化法的實驗結果，相對於 baseline 的改善程度。而第九列代表的是 0dB 至 20dB 的平均改善程度。

由表 4.3.1.1 (a)、(b) 我們可以觀察到幾個現象：第一，隨著訊噪比的下降，語音受到噪音的干擾也越嚴重，而在這種干擾嚴重的情況下，高階倒頻譜正規化的表現會比倒頻譜正規化法來的好。第二，在訊噪比較高的部分，傳統的倒頻譜正規化法的效果則是稍微好一些。第三，無論是在普通訊噪比或是低訊噪比時，其中以對第 80 階作正規化的效果是最好的。第四，在普通訊噪比的情況下，特別是在 5dB 時，各方法可以發揮最好的抗噪音能力。我們也可以觀察到一個現象：隨著訊噪比的下降，各條件下的辨識率和 Baseline 的辨識率之間的差異會越來越大，但是當噪音擾過度嚴重時，高階倒頻譜正規化法和倒頻譜正規化法已無法適度的對不匹配的現象做彌補，因而導致各條件下的辨識率和 Baseline 的辨識率反而拉近了些。圖 4.3.1.2 是上述實驗的比較圖：

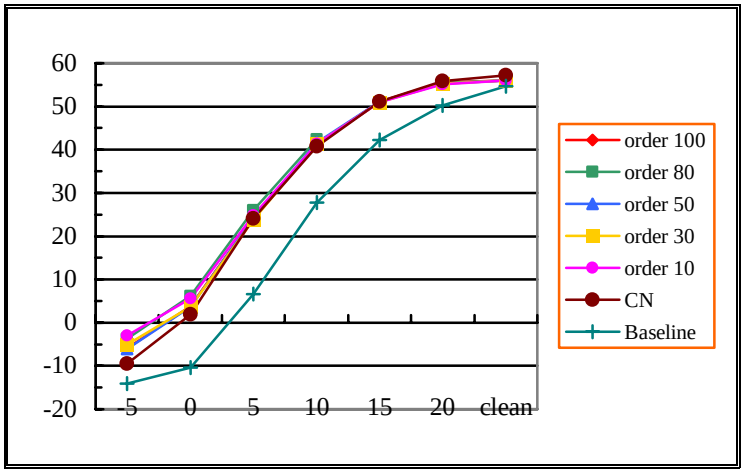

圖 4.3.1.2 在不同訊噪比下對不同階動差做正規化的比較圖

本圖中橫軸代表的是 SNR 比值，縱軸則是不同噪音下的平均音節辨識率。100、80、50、30、10 分別代表的是高階倒頻譜正規化法所作用的階數。CN 是倒頻譜正規化法。Baseline 則是原始辨識率。

由圖 4.3.1.2 中我們可以觀察到，高階倒頻譜正規化法無論作用在何種階數，在比較不受噪音干擾的情況下，其效果大致上是大同小異的，並且和傳統的倒頻譜正規化法的比較也是在伯仲之間。唯獨在受到噪音干擾的程度比較嚴重時，作用在不同階數的效果才有明顯差異。並且我們可以很清楚的觀察到，在訊噪比偏低

的情況下，高階倒頻譜正規化法的效用均比傳統的正規統計圖等化法來的更好。圖 4.3.1.3 是上述諸條件下，對於不同噪音間抗噪音能力的比較：

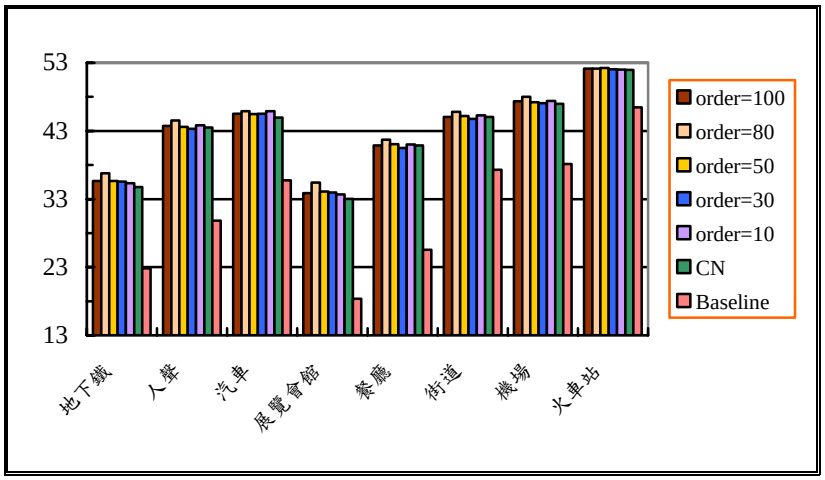

圖 4.3.1.3 在不同噪音下對不同階動差做正規化的比較圖

本圖中橫軸代表的是各種不同噪音，縱軸則是不同訊噪比下的平均音節辨識率，包含了 20dB、15dB、10dB、5dB 等訊噪比的實驗數據。100、80、50、30、10 分別代表的是高階倒頻譜正規化法所作用的階數。CN 是倒頻譜正規化法。Baseline 則是原始辨識率。

由圖 4.3.1.3 中我們可以觀察到兩個現象。第一，在各種不同噪音干擾的情況下，高階倒頻譜的確表現的比倒頻譜正規化法好。第二，在高階倒頻譜正規化法對第 80 階作用的條件下，無論是在較穩定的噪音下或是不穩定的噪音下，效果也是最佳的。

4.3.2 查表式統計圖等化法與高階倒頻譜正規化法的合併

上個小節吾人所探討的是高階倒頻譜正規化法和倒頻譜正規化法的優劣。而本小結吾人想探討的是，查表式統計圖等化法和上述兩種正規化法合併之後，是否會有加成性的效果。實驗步驟是先對語料進行查表式統計圖等化法後（作用在倒頻譜上），再進行高階倒頻譜正規化法（對第 80 階的動差進行正規化），或是倒頻譜正規化法。實驗結果如表 4.3.2.1 所示：

表 4.3.2.1	HEQ + HOCMN	HEQ + CN
3	55.35 / 3.75	56.13 / 5.21
6	56.16 / 4.70	57.05 / 6.36
9	56.44 / 4.15	57.17 / 5.50
12	56.54 / 4.28	57.13 / 5.37
15	56.75 / 3.88	56.85 / 4.06
18	57.01 / 4.34	56.72 / 3.81
AVG	4.18	5.05

表 4.3.2.1 查表式統計圖等化法與高階倒頻譜正規化法的合併後，作用在倒頻譜上的結果
第一行代表的是 HMM 模型訓練的次數。第二行代表的是查表式統計圖等化法與高階倒頻譜正規化法合併的實驗結果。第三行是查表式統計圖等化法與倒頻譜正規化法合併的實驗結果。而第八列代表的是兩種方法相對於 baseline 在不同訓練次數的模型下的平均改善程度。表格中每一格的數據所代表的意義是，(音節辨識率/相對於 baseline 改善程度的百分比)。

由表 4.3.2.1 的實驗數據我們可以觀察到，查表式統計圖等化法與倒頻譜正規化法合併之後的效用，要來的比與高階倒頻譜正規化法合併後的效用來的高。不過該實驗所用的兩種方法都是作用在倒頻譜上，因此接下來吾人嘗試將查表式統計圖等化法先作用在頻譜上，接著再進行高階倒頻譜正規化法或是倒頻譜正規化法。實驗結果如表 4.3.2.2 所示：

表 4.3.2.2	HEQ + HOCMN	HEQ + CN
3	55.51 / 4.05	55.52 / 4.08
6	55.94 / 4.29	56.05 / 4.49
9	55.84 / 3.04	56.44 / 4.15
12	55.98 / 3.19	56.45 / 4.11
15	55.99 / 2.49	56.59 / 3.59
18	56.15 / 2.76	56.66 / 3.70
AVG	3.30	4.02

表 4.3.2.2 查表式統計圖等化法作用在頻譜上後，再與高階倒頻譜正規化法的合併的結果第一行代表的是 HMM 模型訓練的次數。第二行代表的是查表式統計圖等化法與高階倒頻譜正規化法合併的實驗結果。第三行是查表式統計圖等化法與倒頻譜正規化法合併的實驗結果。第八列代表的是兩種方法相對於 baseline 在不同訓練次數的模型下的平均改善程度。表格中每一格的數據所代表的意義是，(音節辨識率/相對於 baseline 改善程度的百分比)。

由表 4.3.2.2 我們可以觀察到，無論是高階倒頻譜正規化法或是倒頻譜正規化法與作用在頻譜上的查表式統計圖等化法合併效用，都比作用在倒頻譜上的差。圖 4.3.2.1 是對上述諸方法的綜合比較：

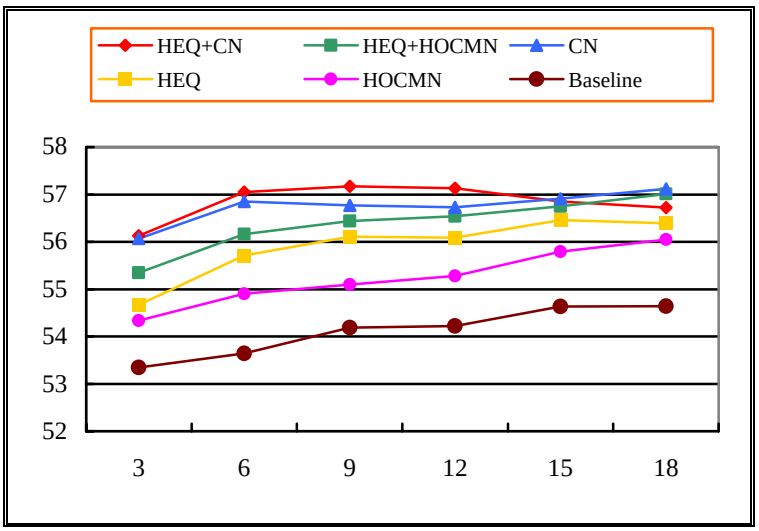

圖 4.3.2.1 查表式統計圖等化法與高階倒頻譜正規化法的綜合比較結果

本圖中橫軸代表的是 HMM 模型訓練的次數，縱軸則是音節辨識率。HEQ+CN 代表的是查表式統計圖等化法與倒頻譜正規化法的合併。HEQ+HOCMN 代表的是查表式統計圖等化法與高階倒頻譜正規化法的合併。Baseline 則是原始辨識率。

圖 4.3.2.1 中的查表式統計圖等化法是作用在倒頻譜上。由該圖我們可以觀察到兩個現象：第一，合併之後的效用大致上都比單獨使用該技術來的好。第二，查表式統計圖等化法與倒頻譜正規化法合併的效果，除了在模型訓練次數越來越高的情況下會稍微差一點，大致上而言要比與高階倒頻譜正規化法來的好。

4.3.3 作用在 AURORA 2.0 的結果

前幾個小節吾人所探討的是將諸語音強健技術結合後作用在華語廣播新聞上。實驗結果證實，結合後的技術的確是能更加的提升語音的強健性。而本小節吾人欲將上述技術作用在由歐洲標準電信協會所提供的語料庫 AURORA 2.0。目的是再次的驗證語音強健技術結合後的效用。實驗結果如表 4.3.3.1 所示，第九行的 AVG 代表-5dB 至 Clean 的平均值：

乾淨語音訓練模式								
	-5dB	0dB	5dB	10dB	15dB	20dB	Clean	AVG
Baseline	8.45	14.99	40.99	72.83	89.71	94.94	97.15	59.87
CN	14.54	33.42	62.39	82.33	91.39	94.86	97.21	68.02
HOCMN	14.87	31.80	60.92	82.20	91.57	94.95	97.18	67.64
HEQ	14.44	29.79	58.67	81.83	91.58	95.03	97.12	66.92
HEQ+CN	14.46	30.29	59.61	82.19	91.66	95.03	97.12	67.19
HEQ+HOCMN	14.40	29.69	58.45	81.76	91.60	95.05	97.14	66.87

複合情境訓練模式								
	-5dB	0dB	5dB	10dB	15dB	20dB	Clean	AVG
Baseline	24.89	58.26	82.78	91.27	94.11	95.22	95.94	77.50
CN	33.10	67.39	86.82	92.16	94.31	95.24	95.47	80.64
HOCMN	31.91	66.31	85.60	92.08	94.24	95.32	95.73	80.17
HEQ	30.60	65.19	85.65	92.07	94.24	95.16	95.32	79.75
HEQ+CN	30.78	65.11	85.60	92.10	94.20	95.14	95.41	79.76
HEQ+HOCMN	30.54	64.96	85.64	92.12	94.23	95.15	95.37	79.72

表 4.3.3.1 查表式統計圖等化法與高階倒頻譜正規化法的合併且作用在 AURORA 2.0 的結果此表是用來比較此五種語音強健技術的抗噪音能力。分別在不同訊噪比以及不同訓練模式下的實驗結果。

由表 4.3.3.1 我們可以觀察到幾個現象。第一，無論在何種訓練模式下，隨著訊噪比越來越低，查表式統計圖等化法加上其它正規化法的表現都不比單獨使用來的好，這點和作用在華語廣播新聞的結果是相異的；唯獨在高訊噪比的情況下，表現才比單獨使用稍微好。第二，在技術合併為前提之下，對於複合情境訓練模式，整體來說兩種技術的表現是在伯仲之間，然而在低訊噪比的情況下，查表式統計圖等化法加上倒頻譜正規化法的表現稍好些。而普通訊噪比的情況下，查表式統計圖等化法加上高階倒頻譜正規化法的表現則是稍好些，高訊噪比的情況下兩者互有優劣，不過差異都非常的小。表 4.3.3.2 是吾人將上述實驗依照不同測試組別以及不同訓練模式所整理的綜合比較的實驗結果，第四行的 AVG 代表 -5dB 至 Clean 的平均值：

原始辨識率			
	乾淨語音訓練	複合情境訓練	AVG
SET A	58.60	76.51	67.56
SET B	61.13	78.47	69.80
SET C	59.08	73.84	66.46
AVG	59.60	76.27	67.94

查表式統計圖等化法			
	乾淨語音訓練	複合情境訓練	AVG
SET A	65.15	79.16	72.16
SET B	68.69	80.33	74.51
SET C	62.55	78.05	70.30
AVG	65.46	79.18	72.32

查表式統計圖等化法 + 倒頻譜正規化法			
	乾淨語音訓練	複合情境訓練	AVG
SET A	65.44	79.16	72.30
SET B	68.95	80.37	74.66
SET C	62.76	78.12	70.44
AVG	65.72	79.22	72.47

查表式統計圖等化法 + 高階倒頻譜正規化法			
	乾淨語音訓練	複合情境訓練	AVG
SET A	65.10	79.13	72.12
SET B	68.64	80.30	74.47
SET C	62.45	78.07	70.26
AVG	65.40	79.17	72.29

表 4.3.3.2 查表式統計圖等化法和其他正規化法合併後，在不同測試組別的實驗結果

此表是用來比較此三種語音強健技術的抗噪音能力。分別在不同測試組別以及不同訓練模式下的實驗結果。

由表 4.3.3.2 我們可以觀察到，無論是在乾淨語音訓練模式或者複合情境訓練模式下，查表式統計圖等化法加上倒頻譜平均消去法，在各個測試組別的表現均來的比其他兩者稍好。而查表式統計圖等化法加上高階倒頻譜平均消去法，和單純只用查表式統計圖等化法的效用，則是相去不遠。因此吾人的結論是，當查表式統計圖等化法和其他語音強健技術合併時，在不同的條件下或許會有加成性的效果，但是大部分的情況下加成性的效果都是有限的，並不如作用在華語廣播新聞資料庫上那樣的明顯，可能的原因是因為華語廣播新聞的模型數量（151 個）遠大於 AURORA 2.0 的模型數量（13 個）的關係，因此模型間在語言學上的複雜性也相對的比較高。圖 4.3.3.1 和圖 4.3.3.2 是上述語音強健技術，在不同訊噪比和不同噪音下的比較結果：

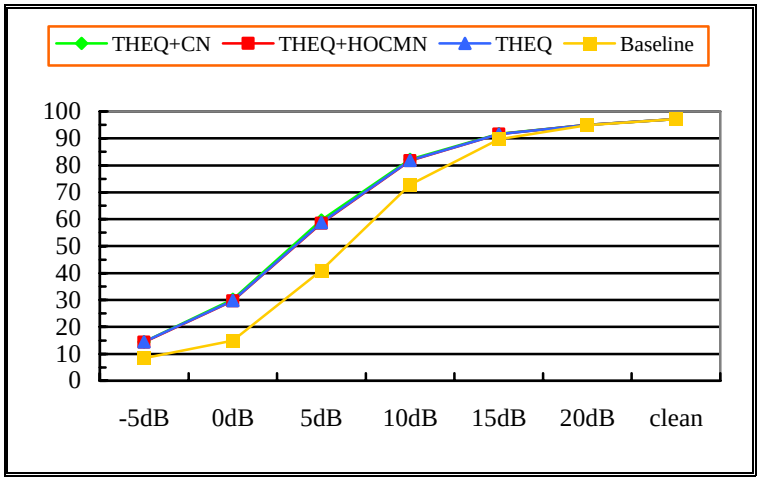

圖 4.3.3.1 (a) 查表式統計圖等化法和正規化法合併後，乾淨語音訓練模式下不同訊噪比的結果

本圖是乾淨訓練模式下的實驗結果。其中橫軸代表的是 SNR 比值，縱軸則是不同噪音下的平均辨識率。THEQ+CN 代表的是查表式統計圖等化法與倒頻譜正規化法的合併。THEQ +HOCMN 代表的是查表式統計圖等化法與高階倒頻譜正規化法的合併。THEQ 是查表式統計圖等化法。Baseline 則是原始辨識率。

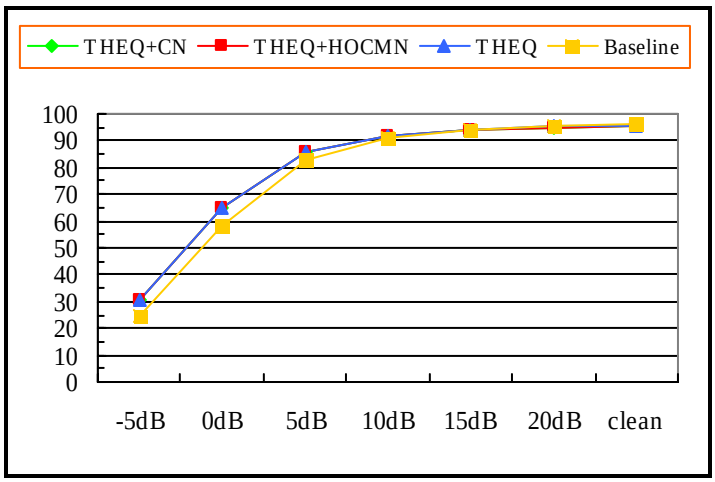

圖 4.3.3.1 (b) 查表式統計圖等化法和正規化法合併後，複合情境訓練模式下不同訊噪比的結果

本圖是複合情境訓練模式下的實驗結果。其中橫軸代表的是 SNR 比值，縱軸則是不同噪音下的平均辨識率。THEQ +CN 代表的是查表式統計圖等化法與倒頻譜正規化法的合併。THEQ +HOCMN 代表的是查表式統計圖等化法與高階倒頻譜正規化法的合併。THEQ 是查表式統計圖等化法。Baseline 則是原始辨識率。

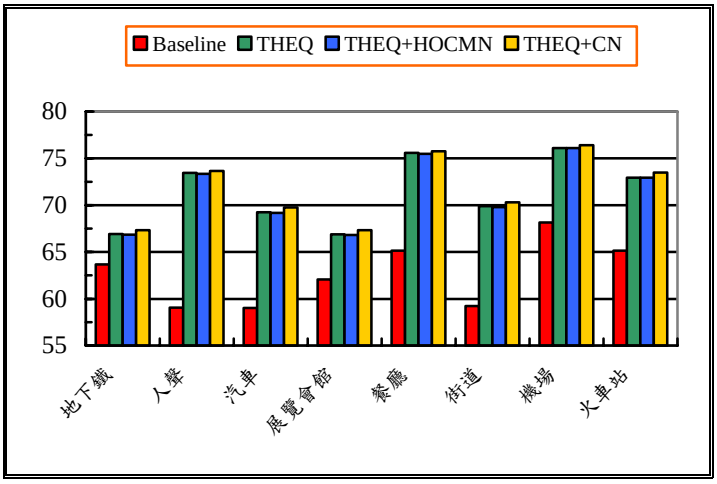

圖 4.3.3.2 (a) 查表式統計圖等化法和正規化法合併後，乾淨語音訓練模式下不同噪音的結果

本圖是乾淨訓練模式下的實驗結果。其中橫軸代表的是各種不同噪音，縱軸則是不同訊噪比下的平均音節辨識率，包含了 20dB、15dB、10dB、5dB、0dB 等訊噪比的實驗數據。Baseline 則是原始辨識率。THEQ 是查表式統計圖等化法。THEQ +HOCMN 代表的是查表式統計圖等化法與高階倒頻譜正規化法的合併。THEQ +CN 代表的是查表式統計圖等化法與倒頻譜正規化法的合併。

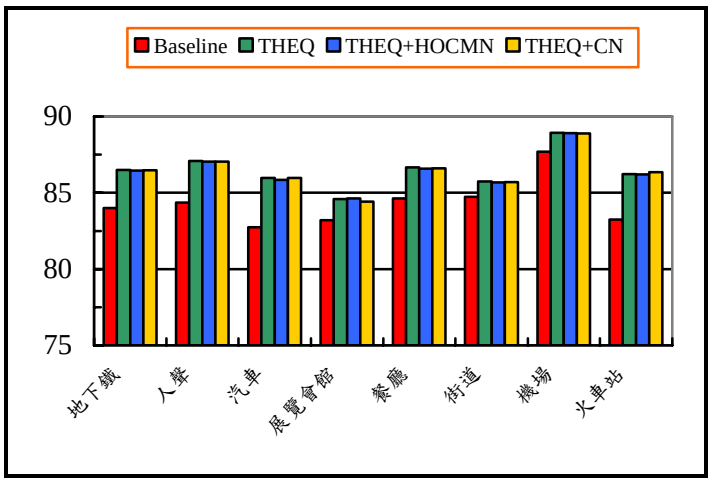

圖 4.3.3.2 (b) 查表式統計圖等化法和正規化法合併後，複合情境訓練模式下不同噪音的結果

本圖是複合情境訓練模式下的實驗結果。其中橫軸代表的是各種不同噪音，縱軸則是不同訊噪比下的平均音節辨識率，包含了 20dB、15dB、10dB、5dB、0dB 等訊噪比的實驗數據。Baseline 則是原始辨識率。THEQ 是查表式統計圖等化法。THEQ +HOCMN 代表的是查表式統計圖等化法與高階倒頻譜正規化法的合併。THEQ +CN 代表的是查表式統計圖等化法與倒頻譜正規化法的合併。

由圖 4.3.3.1 和 4.3.3.2 我們可以觀察到，在乾淨語音訓練模式下，對於不同的噪音，查表式統計圖等化法加上倒頻譜正規化法的表現相較之下是稍微好的。然而在複合情境訓練模式下，三者的差異性則沒有那麼明顯，並且和 Baseline 的差距也比較小了，這是因為存在於模型和語料之間的不匹配現象比較小的關係，使得諸技術的效果，就不如在乾淨語音訓練模式下明顯。

4.4 頻譜熵值特徵

4.4.1 原始頻譜熵值作為特徵參數

近三年來，有學者試圖將自語音訊號頻譜上所獲得的熵值（Spectral Entropy）來作為特徵參數。在資訊理論裡，熵值是用來描述資訊分布的情形（混亂的程度），因此運用在頻譜上，我們可以藉由計算頻譜上的熵值來判斷某時間點語音訊號的特性。舉例來說，當語音訊號的某段落為非語音時（例如靜音），通常共振峰比較不明顯，頻譜曲線也比較平滑，也因此由該段落語音在頻譜上所計算出的熵值是比較高的；反之，當語音訊號的某段落為語音時，會有比較明顯的共振峰，頻譜曲線上則是起伏比較明顯的，因此也會具有比較高的熵值。換句話說，我們可以用熵值來作為判斷語音訊號為靜音與否的依據 [Hung 1998]。不過在計算熵值之前，需要先將頻譜轉換為機率質量函數（Probability Mass Function, PMF），如此該段頻譜區域所算出來的機率值總和才會為 1。

在計算頻譜熵值的時候，是以每個音框為單位，對線性頻譜上的整個頻譜（Full Band）來計算熵值，然而這樣的方式只能粗略的估計頻譜上共振峰起伏的情況，無法確切的描述其所在的位置，因此通常會將全頻段等分為幾個子頻段（Sub Band），用以計算各個子頻段的熵值，並且分別針對各個子頻段轉換為機率質量函數，目的在於希望這些子頻段所計算出的熵值除了能夠用來作為判斷靜音與否的依據，同時也能描述出共振峰所在的位置。而該資訊被認為對於語音辨識是有幫助的，這也是頻譜熵值被用來作為特徵參數的主因。表 4.4.1.1 是吾人將全頻段等切為不同數量子頻段的實驗結果：

表 4.4.1.1	Baseline	Sub-3	Sub-6	Sub-9
3	53.35	52.71 / -1.20	51.97 / -2.59	51.67 / -3.15
6	53.64	53.07 / -1.07	52.81 / -1.55	52.64 / -1.86
9	54.19	53.17 / -1.88	53.12 / -1.98	53.23 / -1.77
12	54.22	52.97 / -2.31	53.35 / -1.61	53.46 / -1.40
15	54.63	52.89 / -3.19	53.21 / -2.60	53.56 / -1.96
18	54.64	52.82 / -3.33	53.10 / -2.82	53.35 / -2.36
AVG	-	-2.16	-2.19	-2.08

表 4.4.1.1 將全頻段等切為不同數量子頻段的實驗結果

第一行是 HMM 模型訓練的次數。第二行是原始辨識率。Sub-3 是將全頻段等切為三個子頻段的實驗結果。Sub-6 是將全頻段等切為六個子頻段的實驗結果。Sub-9 是將全頻段等切為九個子頻段的實驗結果。第八列代表的是三種方法相對於 baseline 在不同訓練次數的模型下的平均改善程度。表格中除了第二行外，每一格的數據所代表的意義是，(音節辨識率/相對於 baseline 改善程度的百分比)。

由表 4.4.1.1 我們可以觀察到，無論子頻段數量的多寡，對於辨識率本身並無太大的影響。上述實驗是將線性頻率上的整個頻譜等切成為數個子頻段。此外我們也可以在非線性梅爾頻譜上將整個頻譜切分為多個部分重疊的子頻段。表 4.4.1.2 是將整個頻譜切等切為十八個子頻段以及將整個頻譜切分為多個部分重疊的子頻段的實驗結果：

表 4.4.1.2	Linear-Sub-18	Mel-Sub-18
3	52.13 / -2.29	50.25 / -5.81
6	52.81 / -1.55	51.09 / -1.55
9	53.07 / -2.07	51.53 / -4.91
12	53.08 / -2.10	51.85 / -4.37
15	53.00 / -2.98	52.09 / -4.65
18	53.11 / -2.80	52.07 / -4.70
AVG	-2.30	-4.33

表 4.4.1.2 將全頻段等切或根據梅爾刻度不等切的結果

第一行是 HMM 模型訓練的次數。Linear-Sub-18 是將整個頻譜等切為十八個子頻段。Mel-Sub-18 是將整個頻譜劃分為十八個部分重疊的子頻段。第八列代表的是兩種方法相對於 baseline 在不同訓練次數的模型下的平均改善程度。表格中除了第二行外，每一格的數據所代表的意義是，(音節辨識率/相對於 baseline 改善程度的百分比)。

由表 4.4.1.2 我們可以觀察到，將全頻段等切為子頻段的效果，要比根據梅爾刻度將全頻段不等切的效果來的好。前兩個實驗的結果，並不如預期對於辨識率有所提升，與其它學者在國外的研究成果不一致 [Misra *et al.* 2004]，因此吾人嘗試對頻譜熵值進行處理後再作為特徵參數，諸如熵值正規化、計算差量頻譜熵值。我們由下面的式子來了解：

$$\tilde{Y}_t^i = \frac{Y_t^i - \bar{Y}^i}{\sigma^i} \quad where \quad \bar{Y}^i = \frac{1}{T}\sum_{t=0}^{T-1} Y_t^i \ , \ \sigma^i = \sqrt{\frac{1}{T}\sum_{t=0}^{T-1}(Y_t^i - \bar{Y}^i)^2} \tag{4.4.1}$$

$$\tilde{Y}_t^i = \frac{\sum_{r=1}^{R} r(Y_{t+r}^i - Y_{t-r}^i)}{2\sum_{r=1}^{R} r^2} \tag{4.4.2}$$

$$\tilde{Y}_t^i = \frac{(\sum_{r=1}^{R} Y_{t+r}^i + Y_{t-r}^i) + Y_t^i}{2R+1} \tag{4.4.3}$$

式(4.4.1) 是類似倒頻譜正規化法的方式，吾人稱之為頻譜熵值正規化法（以下簡稱 Spectral Entropy Normalization, SEN），其中 $\tilde{Y}_t^i$ 是時間 t 第 i 個子頻段正規化後的頻譜熵值，Y_t^i 代表的是在時間 t 第 i 個子頻段所獲得的頻譜熵值，$\overline{Y}^i$ 代表的是所有時間點上第 i 個子頻段頻譜熵值的平均值，σ^i 是所有時間點上第 i 個子頻段頻譜熵值的標準差。式(4.4.2) 是類似計算差量倒頻譜的方式，吾人稱之為差量頻譜熵值（以下簡稱 Spectral Delta Entropy, SDE），目的是希望獲得頻譜熵值在時間上變化的程度，其中 $\tilde{Y}_t^i$ 是時間 t 第 i 個子頻段的差量頻譜熵值，r 是時間間隔，Y_{t+r}^i 和 Y_{t-r}^i 則分別是時間點 $t+\gamma$ 以及時間點 $t-\gamma$ 第 i 個子頻段的頻譜熵值。式(4.4.3) 是類似式(4.4.1) 的方式（以下簡稱 Spectral Smothed Entropy, SSE），不過是對特定長度的時間區段進行正規化，而非對所有時間點進行正規化，其中 $\tilde{Y}_t^i$ 是時間 t 第 i 個子頻段正規化後的頻譜熵值，Y_t^i 是時間 t 第 i 個子頻段正規化前的頻譜熵值，r 是時間間隔，Y_{t+r}^i 和 Y_{t-r}^i 則分別是時間點 $t+\gamma$ 以及時間點 $t-\gamma$ 第 i 個子頻段的頻譜熵值。由於前兩個實驗的數據說明了辨識率並不會隨著子頻段的數量增加而有所增益，因此本實驗以等分為三個子頻段為基準。表 4.4.1.3 是實驗結果：

表 4.4.1.3	Linear-Sub-3	Linear-Sub-3 + SEN	Linear-Sub-3 + SDE	Linear-Sub-3 + SSE
3	52.71 / -1.20	50.74 / -4.89	52.10 / -2.34	50.55 / -5.25
6	53.07 / -1.07	51.23 / -4.49	52.53 / -2.07	51.14 / -4.66
9	53.17 / -1.88	52.23 / -3.62	52.50 / -3.12	51.38 / -5.19
12	52.97 / -2.31	53.18 / -1.92	52.61 / -2.97	51.47 / -5.07
15	52.89 / -3.19	54.17 / -0.84	52.41 / -4.06	51.74 / -5.29
18	52.82 / -3.33	54.43 / -0.38	52.39 / -4.12	51.83 / -5.14
AVG	-2.16	-2.69	-3.11	-5.10

表 4.4.1.3 將頻譜熵值進行三種正規化後的結果

第一行是 HMM 模型訓練的次數。Linear-Sub-3 是將整個頻譜等切為三個子頻段。Linear-Sub-3+EN 是將整個頻譜等切為三個子頻段，並對計算出的熵值進行 EN。Linear-Sub-3+DE 是將整個頻譜等切為三個子頻段，並對計算出的熵值進行 DE。Linear-Sub-3+PE 是將整個頻譜等切為三個子頻段，並對計算出的熵值進行 PE。第八列代表的是四種方法相對於 baseline 在不同訓練次數的模型下的平均改善程度。表格中每一格的數據所代表的意義是，(音節辨識率/相對於 baseline 改善程度的百分比)。

由表 4.4.1.3 我們可以觀察到，對頻譜熵值進行 EN 的結果，要比進行其他正規化方式要來的好，不過在不受噪音干擾的環境下，仍然比進行 EN 前的結果來的差。

4.4.2 頻譜熵值在噪音環境下的抗噪音能力

上小節的實驗中，無論是將整個頻譜等切為數個子頻段，或在非線性梅爾頻譜上將整個頻譜切分為多個部分重疊的子頻段，在計算各個子頻段的熵值前，均會先各別將各個子頻段轉換為機率質量函數，事實上這樣的做法會面臨到一個問題，也就是子頻段彼此間的熵值是獨立的。舉例來說，假設在時間點 *t*，子頻段 *sub-1* 包含了一個峰值（Peak），相較於子頻段 *sub-1*，子頻段 *sub-2* 具有更大的峰值。然而由於分別將子頻段轉換為機率質量函數，導致兩個子頻段所計算出來的熵值

是差不多的。換句話說，子頻段 *sub-1* 原本應該具有比較高的熵值，卻因為分別將子頻段轉換為機率質量函數，反而具有比較低的熵值。因此以下的實驗，吾人先將整個頻譜轉為機率質量函數，依然保持各個頻率間峰值間的相對關係，接著再將整個頻譜等切成三個子頻段以計算熵值，並以此（Linear-Full-3）作為實驗組。表 4.4.2.1 是實驗結果：

表 4.4.2.1	Linear-Full-3	Linear-Sub-3
3	52.13 / -2.29	52.71 / -1.20
6	52.81 / -1.55	53.07 / -1.07
9	53.07 / -2.07	53.17 / -1.88
12	53.08 / -2.10	52.97 / -2.31
15	53.00 / -2.98	52.89 / -3.19
18	53.11 / -2.80	52.82 / -3.33
AVG	-2.30	-2.16

表 4.4.2.1 將整個頻譜或各別子頻段轉為機率質量函數的結果

第一行是 HMM 模型訓練的次數。Linear-Full-3 是把整個頻譜等切為三個子頻段並將整個頻譜轉為機率質量函數。Linear-Sub-3 是把整個頻譜等切為三個子頻段並分別將三個子頻段轉為機率質量函數。第八列代表的是兩種方法相對於 baseline 在不同訓練次數的模型下的平均改善程度。表格中每一格的數據所代表的意義是，(音節辨識率/相對於 baseline 改善程度的百分比)。

由表 4.4.2.1 我們可以觀察到，在不受噪音干擾的情況下，無論將整個頻譜轉為機率質量函數或是各別將子頻段轉為機率質量函數的結果是大同小異的。接下來的實驗，吾人試圖在廣播新聞資料庫上人工加入 AURORA 2.0 所提供的八種噪音，配合不同程度的訊噪比，來比較頻譜熵值在噪音環境下抗噪音能力。實驗中的特徵參數是採取 39 維傳統梅爾倒頻譜特徵參數，加上 3 維頻譜熵值共 42 維。其中頻譜熵值的計算方式是將整個頻譜等切為 3 個子頻段，並且對計算出的頻譜熵值再進行上個小節所提到的 SEN。圖 4.4.2.1 和圖 4.4.2.2 是實驗結果：

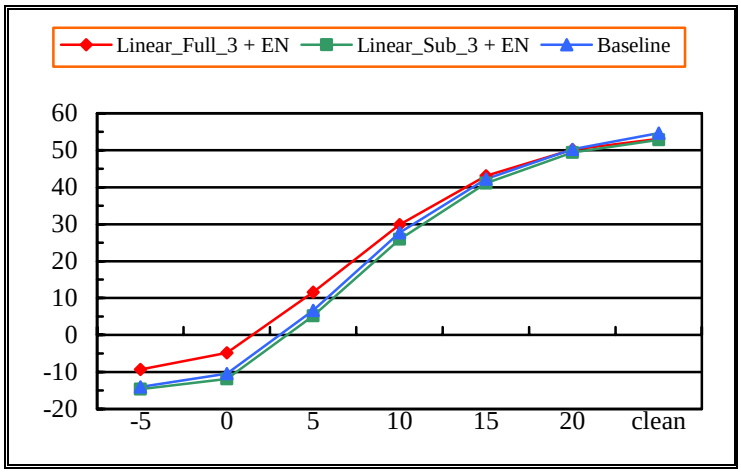

圖 4.4.2.1 將整個頻譜或各別子頻段轉為機率質量函數後，在不同訊噪比下的結果

本圖中橫軸代表的是 SNR 比值，縱軸則是不同噪音下的平均音節辨識率。Linear_Full_3 + EN 代表的是把整個頻譜等切為三個子頻段後，再將整個頻譜轉為機率質量函數後計算出熵值。Linear_Sub_3 + EN 代表的是把整個頻譜等切為三個子頻段後，再將各個子頻段轉為機率質量函數後計算出熵值。Baseline 則是原始辨識率。

由圖 4.4.2.1 我們可以觀察到，將整個頻譜轉為機率質量函數的方式，隨著訊噪比越來越低，相較於各別把子頻段轉為機率質量函數，抗噪的效果也越來越好。

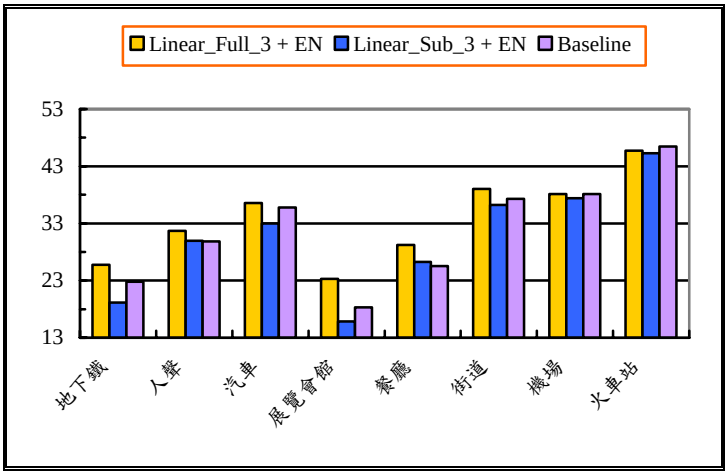

圖 4.4.2.2 將整個頻譜或各別子頻段轉為機率質量函數後，在不同噪音下的結果

本圖中橫軸代表的是各種不同噪音，縱軸則是不同訊噪比下的平均音節辨識率。Linear_Full_3 + EN 代表的是把整個頻譜等切為三個子頻段後，再將整個頻譜轉為機率質量函數後計算出熵值。Linear_Sub_3 + EN 代表的是把整個頻譜等切為三個子頻段後，再將各個子頻段轉為機率質量函數後計算出熵值。Baseline 則是原始辨識率。

由圖 4.4.2.2 我們可以觀察到，無論是在極度不穩定的噪音環境下或是稍微穩定的噪音環境下，將整個頻譜轉為機率質量函數的方式，相較於各別把子頻段轉為機率質量函數，抗噪的效果均是比較好的。