第五章 尺度重刻法於語音端點偵測之應用

5.1 常見之語音端點偵測技術

語音訊號之端點偵測(Voice Activity Detection)在自動語音辨識處理過程中被視為一重要部分。事實上倘若能精確判斷語音訊號之正確端點位置，則自然的在非語音訊號部份就不會有辨識錯誤的發生，在此所謂辨識錯誤就是指原本為非語音的訊號部份被辨識出不該出現的錯誤詞。因此本章節將討論目前常見的語音端點偵測技術並應用吾人所提出的尺度重刻法於音框對數能量端點偵測技術。

5.1.1 音框對數能量偵測法(Log Energy, LE)

音框對數能量偵測法[ETSI 2000]主要是針對語音音框之對數能量值來判斷語音訊號的端點所在處，其原理基於非語音部份之音框的對數能量值相對會存在較小的值域範圍，然而在有語音之音框部份會有較高的能量值，藉此一現象，對於每一語句找出門檻值(Threshold)，利用此門檻值判斷該語句中音框的對數能量值，小於此門檻值即判定此音框為非語音音框，若大於此門檻值則判定此音框為語音音框。具體作法類似 3.1.4 節中靜音音框對數能量正規化法 I 中提到的方法。如下式(5.1.1)和(5.1.2)所表示+：

$$\tau = 1.0 \times \frac{1}{T}\sum_{j=1}^{T} \log E[j] \tag{5.1.1}$$

$$\log E[i] \begin{cases} Speech & \text{, if} \quad \log E[i] \geq \tau \\ Noies & \text{, otherwise} \end{cases} \tag{5.1.2}$$

其中 T 為一段語音的音框數，τ 為該語句的門檻值， $\log E[j]$ 為音框 j 對數能量。從圖 5.1.1(語音內容為數字：O3O)中可以觀察出音框對數能量的變化情形，圖(a)表示為時間軸上的取樣點大小，圖(b)表示為音框對數能量並表示門檻值關係。

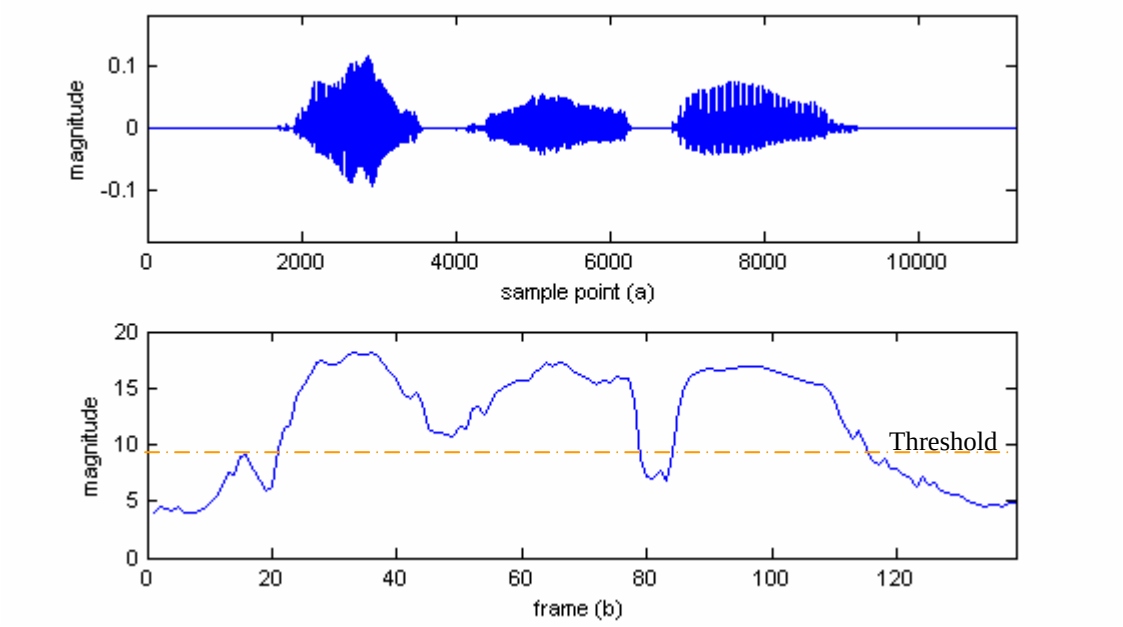

圖 5.1.1 音框能量偵測法圖示

5.1.2 頻譜熵值偵測法(Spectral Entropy, SE)

熵值法(Entropy)[Shannon 1948；Misra et al. 2004]在資訊理論裡扮演了相當重要的角色。其用途相當的廣，除了可以用作為資訊量程度的量測外，同時也可以看成是資訊的混淆度，藉用這樣的特性恰好可以來觀察頻譜上亂度的情況，進而判斷語音端點位置。熵值法其定義：首先假設一隨機變數 X ，其機率分佈為 $P(X = x_i) = P_i, i = 1,2,\ldots,n$ ，則其機率分佈的熵值如式(5.1.3)

$$H = -\sum_{i=1}^{n} P_i \log P_i \tag{5.1.3}$$

在這裡我們利用熵值法來計算頻譜上的熵值，但事前需先將頻譜轉換為機率質量函數(Probability Mass Function, PMF)，以方便計算熵值，作法主要針對每一音框之各頻譜帶強度(Magnitude)取其相對於全頻帶強度和的機率值，如式(5.1.4)所表示：

$$x_i = \frac{M_i}{\sum_{i=1}^{N} M_i} \qquad i = 1,2,\ldots,N \tag{5.1.4}$$

式中 M_i 代表頻譜上第 i 個頻譜帶上的強度大小，而 N 是頻帶個數，x_i 則表示該頻帶強度在此音框中所佔的比重。而對於每個音框的熵值計算如式(5.1.5)

$$H = -\sum_{i=1}^{N} x_i \log x_i \tag{5.1.5}$$

由圖 5.1.2 觀察計算出的熵值，當語音存在於某音框時，其熵值會明顯較低。而當音框不存在語音時，則熵值會偏高，因此我們可以設定一門檻值，便可被利用來做為判對語音端點的依據。

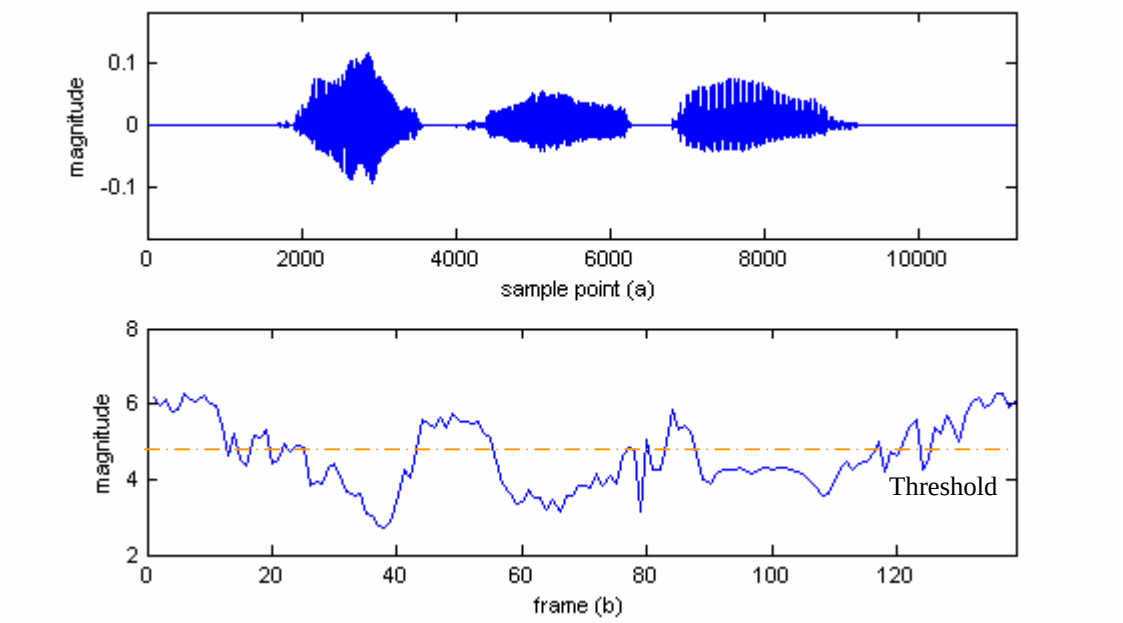

圖 5.1.2 音框熵值法圖示

5.1.3 長時期頻譜差異法(Long-Term spectral divergence, LTSD)

長時期頻譜差異法(LTSD)[Ramírez et al. 2004][Górriz et al. 2006] 主要目的在頻譜值上找出語音和非語音的片段，此方法假設在一段長時間的頻譜中可以找出較具有語音特徵的最大頻譜值，做法是藉由預測一段長期頻譜封包(Long-Term Spectral Envelope, LTSE)，將其最大頻譜值當作語音訊號的成分。作法可以分為兩部份，首先在頻譜封包(LTSE)中求取語音特徵的最大頻譜值，針對每一音框作

傅利葉轉換求取其各功率頻譜，假設 $X(k,l)$ 為此訊號在第 l 個音框上的第 k 個頻率之頻譜強度(Magnitude Spectrum)，接著利用長期封包設定取其該音框之前後 N 個音框範圍中的最大值。定義如式(5.1.6)：

$$LTSE_N(k,l) = \max\{X(k,l+j)\}_{j=-N}^{j=+N} \tag{5.1.6}$$

則此長期頻譜封包(LTSE)即為(前後2N+1)個相鄰音框中之各頻率的最大頻譜值。第二部分則利用長期頻譜封包取得每一個音框的長期頻譜差異值(LTSD)，定義如下式(5.1.7)：

$$LTSD_N(l) = 10\log_{10}\left(\frac{1}{NFFT}\sum_{K=0}^{NFFT-1}\frac{LTSE^2(k,l)}{N^2(k)}\right) \tag{5.1.7}$$

其中 $N(k)$ 為噪音之頻譜強度， $NFFT$ 則為離散傅利葉轉換的點數。最後計算出的長期頻譜差異值，當語音存在於某音框時，其長期頻譜差異值會明顯較高。而當音框不存在語音時，則長期頻譜差異值則偏低，因此我們可以設定一門檻值，便可被利用此長期頻譜差異值作為判對語音端點的依據。圖 5.1.3 表示為長期頻譜差異值變化情形。

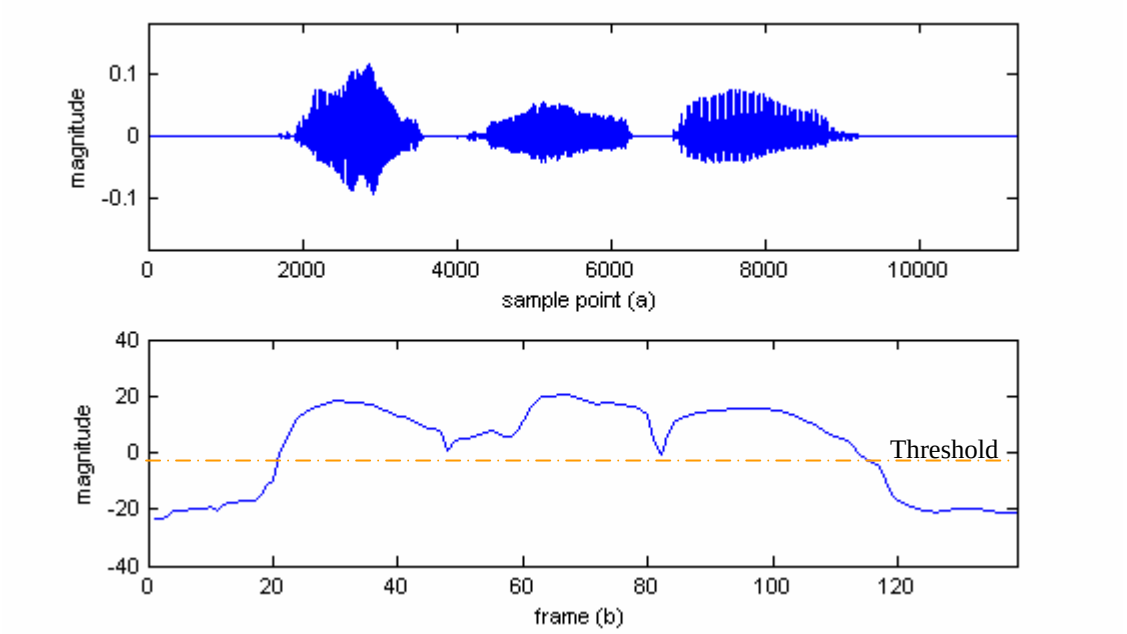

圖 5.1.3 長期頻譜差異值圖示

5.2 語音端點偵測實驗

此節將討論常見之語音端點偵測技術，包括能量偵測法、頻譜熵值偵測法與長時期頻譜差異法。在實驗環境設定上，由於在 Aurora-2.0 實驗語料庫標準設定中並沒有提供語音與非語音之音框標示，在此我們特別以人工目測法對 Aurora-2.0 中測試組 A 與測試組 B 之語料做標示，語料數共為 56056 句，實驗中並以此人工標示為基準，做為最後測試不同之語音端點偵測技術的正確答案。此外在各端點技術的門檻值設定，於實驗中我們皆假設測試語料的前五個音框為靜音音框，因此門檻值採用前五個音框值之平均。

正確率計算：

我們定義 HR0(Non-Speech Hit-Rate)與 HR1(Speech Hit-Rate)[Ramírez et al. 2004]，如下式：

$$\mathrm{HR0} = \frac{N_{0,0}}{N_0^{ref}} \tag{5.2.1}$$

$$\mathrm{HR1} = \frac{N_{1,1}}{N_1^{ref}} \tag{5.2.2}$$

其中 N_0^{ref} 與 N_1^{ref} 表示為非語音音框和語音音框在目測法下之標準個數，而 $N_{0,0}$ 表示為非語音音框被正確判斷個數，而 $N_{1,1}$ 則表示語音音框被正確判斷的個數。端點偵測數據將如下小節所示。

對數能量偵測法(Log Energy)：

對數能量偵測法針對語音音框之對數能量值，利用門檻值判斷該語句中之音框屬於語音訊號或非語音訊號，實驗結果如表 5.2.1 與表 5.2.2。

HR0	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	平均
測試組A	地下鐵	24.26	37.85	41.16	41.63	43.29	41.85	41.94	38.85
	人聲	23.31	31.81	35.75	37.52	39.55	39.66	40.11	35.39
	汽車	23.02	29.31	29.27	31.74	33.23	31.47	31.02	29.87
	展覽會館	23.05	31.70	33.54	36.11	36.35	36.96	36.25	33.42
	平均	23.41	32.67	34.93	36.75	38.10	37.48	37.33	34.38
測試組B	餐廳	24.26	34.18	37.06	39.99	41.18	41.74	40.70	37.02
	街道	23.31	35.32	40.61	42.13	40.66	41.20	39.32	37.51
	機場	23.02	32.55	36.15	37.04	39.86	40.21	38.31	35.31
	火車站	23.05	28.91	30.63	32.31	33.89	32.65	32.63	30.58
	平均	23.41	32.74	36.11	37.87	38.90	38.95	37.74	35.10

表 5.2.1 能量偵測法之非語音音框正確率結果

HR1	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	平均
測試組A	地下鐵	99.76	97.70	94.73	89.38	80.97	72.77	64.47	85.68
	人聲	99.73	99.24	98.37	96.56	92.29	85.18	76.74	92.59
	汽車	99.70	99.23	98.64	96.34	94.06	88.40	83.10	94.21
	展覽會館	99.77	98.32	96.43	92.62	85.85	78.25	71.29	88.93
	平均	99.74	98.62	97.04	93.73	88.29	81.15	73.90	90.35
測試組B	餐廳	99.76	98.91	97.73	95.19	89.97	83.58	74.56	91.39
	街道	99.73	97.98	95.07	90.69	84.91	75.75	68.37	87.50
	機場	99.70	99.09	98.27	96.20	92.01	86.63	77.94	92.83
	火車站	99.77	99.18	98.39	96.77	92.91	88.12	81.04	93.74
	平均	99.74	98.79	97.36	94.71	89.95	83.52	75.48	91.36

表 5.2.2 能量偵測法之語音音框正確率結果

頻譜熵值偵測法(Spcetral Entropy)：

熵值法如式(5.1.5)藉由觀察頻譜上亂度的情形判斷該語句中之音框屬於語音訊號或非語音訊號，實驗結果如表 5.2.3 與表 5.2.4。

HR0	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	平均
測試組A	地下鐵	9.76	18.34	19.87	18.97	18.05	18.91	18.30	17.46
	人聲	9.94	29.54	29.43	29.30	27.65	25.81	25.03	25.24
	汽車	9.48	10.47	8.99	9.33	9.11	8.11	9.20	9.24
	展覽會館	9.36	20.08	20.42	21.14	21.26	21.13	22.13	19.36
	平均	9.64	19.61	19.68	19.69	19.01	18.49	18.66	17.82
測試組B	餐廳	9.76	33.44	32.09	32.87	32.00	33.06	31.54	29.25
	街道	9.94	21.89	23.93	24.49	22.71	23.46	24.02	21.49
	機場	9.48	37.50	36.90	33.34	32.61	31.35	31.88	30.44
	火車站	9.36	18.43	17.61	17.18	16.68	16.04	16.56	15.98
	平均	9.64	27.81	27.63	26.97	26.00	25.98	26.00	24.29

表 5.2.3 頻譜熵值偵測法之非語音音框正確率結果

HR1	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	平均
測試組A	地下鐵	99.04	96.04	93.00	89.45	83.58	77.33	75.41	87.69
	人聲	99.03	96.53	94.28	92.10	88.19	82.95	78.33	90.20
	汽車	99.11	99.36	99.20	98.38	97.09	95.80	92.76	97.38
	展覽會館	99.25	94.23	92.89	89.85	84.09	79.37	75.53	87.89
	平均	99.11	96.54	94.84	92.45	88.24	83.86	80.51	90.79
測試組B	餐廳	99.04	92.88	91.49	87.79	82.97	77.05	71.27	86.07
	街道	99.03	94.25	90.83	89.05	86.01	80.96	76.76	88.13
	機場	99.11	87.82	86.35	84.81	79.59	76.60	69.13	83.34
	火車站	99.25	95.97	95.49	94.66	91.53	88.35	85.01	92.89
	平均	99.11	92.73	91.04	89.08	85.02	80.74	75.54	87.61

表 5.2.4 頻譜熵值偵測法之語音音框正確率結果

長時期頻譜差異偵測法(LTSD)：

長時期頻譜差異法用意在頻譜值上找出語音和非語音的片段，方法中之長期封包設定該音框之前後$2N+1$個音框範圍，實驗設定$N=3$，定義如式(5.1.6)。實驗結果如表 5.2.5 與表 5.2.6。

HR0	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	平均
測試組A	地下鐵	16.67	31.82	36.06	38.10	42.40	43.03	44.42	36.07
	人聲	15.77	25.09	29.50	32.51	37.00	40.04	43.05	31.85
	汽車	15.79	24.36	26.17	29.93	35.98	36.69	38.82	29.68
	展覽會館	15.77	27.47	30.26	34.62	37.19	40.45	40.52	32.32
	平均	16.00	27.18	30.50	33.79	38.14	40.05	41.70	32.48
測試組B	餐廳	16.67	27.48	31.05	35.14	38.22	40.53	41.79	32.98
	街道	15.77	29.60	35.56	38.80	40.29	42.43	42.99	35.06
	機場	15.79	24.68	28.86	31.71	35.90	38.65	39.75	30.76
	火車站	15.77	23.19	26.44	29.26	33.79	34.89	36.95	28.61
	平均	16.00	26.24	30.48	33.72	37.05	39.13	40.37	31.86

表 5.2.5 長時期頻譜差異偵測法之非語音音框正確率結果

HR1	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	平均
測試組A	地下鐵	99.99	98.51	96.78	92.68	85.16	76.81	67.32	88.18
	人聲	99.99	99.80	99.02	97.45	92.85	84.03	74.11	92.46
	汽車	99.96	99.41	98.70	96.14	93.26	86.82	78.53	93.26
	展覽會館	100.00	99.14	97.30	94.00	87.45	80.55	72.50	90.13
	平均	99.99	99.21	97.95	95.07	89.68	82.05	73.11	91.01
測試組B	餐廳	99.99	99.42	98.30	95.66	90.93	83.76	74.01	91.72
	街道	99.99	98.38	95.12	90.58	84.53	74.90	65.99	87.07
	機場	99.96	99.77	99.13	97.52	93.26	88.02	78.78	93.78
	火車站	100.00	99.66	98.99	97.08	93.17	87.59	79.36	93.69
	平均	99.99	99.31	97.89	95.21	90.47	83.57	74.53	91.57

表 5.2.6 長時期頻譜差異偵測法之語音音框正確率結果

5.3 對數能量尺度重刻法於對數能量端點偵測之實驗

根據第四章的對數能量觀察，提出對數能量尺度重刻法 I (LERN I)，在此我們特別針對音框對數能量偵測法作進一步處理，我們希望藉由使對數能量正規化後的對數能量值能夠提高語音片段與非語音片段的判斷效果。圖 5.3.1 表示為對數能量尺度重刻法 I 的處理前後變化情形，圖(a)為對數能量處理前曲線，圖(b)表示為尺度大小 M=100 處理後的對數能量曲線。實驗結果如表 5.3.1 與表 5.3.2。

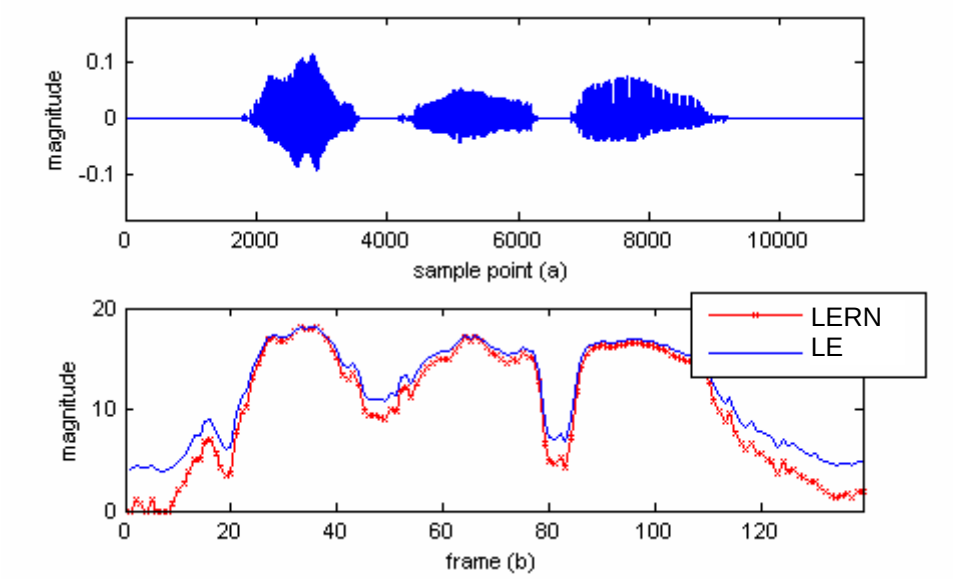

圖 5.3.1 尺度重刻法於音框能量偵測圖示

HR0	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	平均
測試組A	地下鐵	25.96	41.22	45.15	46.13	48.31	47.21	48.07	43.15
	人聲	24.97	35.22	39.83	42.67	45.00	45.48	46.73	39.98
	汽車	24.56	37.73	40.10	44.55	48.02	47.49	48.75	41.60
	展覽會館	24.86	38.21	40.89	44.67	45.78	47.15	47.33	41.27
	平均	25.09	38.10	41.49	44.50	46.78	46.83	47.72	41.50
測試組B	餐廳	25.96	37.01	40.14	44.12	45.31	46.46	45.99	40.71
	街道	24.97	39.52	45.38	47.77	46.66	47.74	46.37	42.63
	機場	24.56	36.10	40.81	42.33	46.03	47.08	45.96	40.41
	火車站	24.86	35.57	38.91	42.23	45.09	44.91	46.12	39.67
	平均	25.09	37.05	41.31	44.11	45.77	46.55	46.11	40.86

表 5.3.1 尺度重刻法於音框能量偵測之非語音音框正確率結果

HR1	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	平均
測試組A	地下鐵	99.75	97.36	94.00	87.93	78.28	69.07	59.74	83.73
	人聲	99.71	99.16	98.10	95.87	90.79	82.36	72.15	91.16
	汽車	99.68	99.03	98.11	94.96	90.98	82.78	72.89	91.20
	展覽會館	99.76	98.04	95.69	90.93	82.88	72.93	63.58	86.26
	平均	99.73	98.40	96.47	92.42	85.73	76.78	67.09	88.09
測試組B	餐廳	99.75	98.80	97.41	94.61	88.68	81.29	70.94	90.21
	街道	99.71	97.77	94.46	89.43	82.81	72.51	63.96	85.81
	機場	99.68	98.97	98.01	95.57	90.38	83.87	73.14	91.38
	火車站	99.76	99.01	98.01	95.95	90.94	84.57	74.56	91.83
	平均	99.73	98.64	96.97	93.89	88.20	80.56	70.65	89.81

表 5.3.2 尺度重刻法於音框能量偵測之語音音框正確率結果

實驗探討：

依據表 5.2.1 至 5.2.6 與表 5.3.1 和表 5.3.2 的辨識結果。我們整理如表 5.3.3，表中我們可以從總正確率(Overall Hit Rate)觀察出對數能量尺度重刻法 I 的處理前後的能量偵測正確率最高，次高則為原能量偵測(LEn)，結果顯示對數能量尺度重刻法能增加正確率效果。此外我們發現各方法的平均正確率上在非語音部分差異性較大，其中以頻譜熵值偵測法的效果較差。

偵測技術	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	平均
Entropy	HR0	9.64	23.71	23.66	23.33	22.51	22.23	22.33	21.06
	HR1	99.11	94.64	92.94	90.76	86.63	82.30	78.02	89.20
Overall Hit Rate		54.37	59.17	58.30	57.05	54.57	52.27	50.18	55.13
LTSD_N3	HR0	16.00	26.71	30.49	33.76	37.60	39.59	41.04	32.17
	HR1	99.99	99.26	97.92	95.14	90.08	82.81	73.82	91.29
Overall Hit Rate		57.99	62.98	64.20	64.45	63.84	61.20	57.43	61.73
LE	HR0	23.41	32.70	35.52	37.31	38.50	38.22	37.54	34.74
	HR1	99.74	98.71	97.20	94.22	89.12	82.34	74.69	90.86
Overall Hit Rate		61.57	65.70	66.36	65.76	63.81	60.28	56.11	62.80
LERN I	HR0	25.09	37.57	41.40	44.31	46.28	46.69	46.92	41.18
	HR1	99.73	98.52	96.72	93.16	86.97	78.67	68.87	88.95
Overall Hit Rate		62.41	68.05	69.06	68.73	66.62	62.68	57.89	65.06

表 5.3.3 端點偵測技術之正確率比較