第三章　實驗語料庫介紹與設定

本章節主要是介紹本論文所使用的語料庫，以及相關的實驗設定。在第一小節吾人將說明本實驗所用的特徵參數抽取的步驟，第二小節將介紹聲學模型，第三小節則是介紹本論文所採用之兩套語料庫，第四小節則是介紹基礎實驗結果。

3.1　語音特徵參數的抽取

在語音辨識的領域裡，如何從語音訊號裡抽取出合適的特徵參數是頗重要的課題。而目前廣為大家所接受且使用的特徵參數大至上如：梅爾倒頻譜係數（Mel-Scale Frequency Cepstral Coefficients, MFCC）[Davis *et al.* 1980]，感知線性預測（Perceptual Linear Prediction, PLP）[Hermansky 1990]，線性預測係數（Linear Prediction Coefficients, LPC） [Gray *et al.* 1973]等等。本論文所使用的特徵參數是梅爾倒頻譜係數，此特徵參數之所以廣為大家所採用的主要因素在於其考慮到人耳對於不同頻率的感受程度，因此適合於語音辨識。以下吾人將簡要的介紹抽取梅爾倒頻譜係數時主要的步驟：

（1）預強調 (Pre-Emphasis)：將語音訊號通過一個高通濾波器

這個動作的目的在於，消除在發聲的過程中聲帶和嘴唇相互間所造成的效應，以補償語音訊號受到發音系統所壓抑的高頻部份。換句話說，是要突顯在高頻的能量。我們用式 (3.1.1) 來表示：

$$\hat{s}(n) = s(n) - a \bullet s(n-1) \tag{3.1.1}$$

式 (3.1.1) 中，$\hat{s}(n)$ 是預強調之後的訊號。a 是預強調的參數，一般是介於 0.9 和 1 之間。本論文所採用的參數值為 0.975。

（2）音框化 (Framing)：將某特定數量的採樣點集中成一個單位即所謂的音框一般而言，該特定數量 a 的值是 256 或 512，涵蓋的時間約為 20 至 30 ms 左右。為了避免相鄰兩音框的變化過大，我們會讓兩相鄰音框之間有一段重疊區域，此重疊區域包含了 b 個取樣點，且 b 的值約是 q 的 1/2 或 1/3。通常語音辨識所用的音訊的取樣頻率為 8KHz 或 16KHz，以 8KHz 來說，若音框長度為 256 個取樣點，則對應的時間長度是(256 / 8000) * 1000 = 32 ms。本論文在華語廣播新聞資料庫上所使用的取樣頻率為 16KHz，取樣點為 320，每單位音框的涵蓋時間為 20 ms。而在 Aurora 2.0 資料庫（稍後將做介紹）上的取樣頻率為 8KHz，取樣點為 250，每單位音框的涵蓋時間為 31.25 ms。

（3）漢明窗 (Hamming Window)：將每一個音框乘上漢明窗

此動作目的在於增加音框左端和右端的連續性。我們用式 (3.1.2) 來表示音框化後乘上漢明窗的過程：

$$w(n)=\begin{cases}(1-\alpha)-\alpha\cos(\dfrac{2\pi n}{N-1}) & n=0,1,...,N-1\\ 0 & otherwise\end{cases} \tag{3.1.2}$$

隨著式 (3.1.2) 中 α 的不同，會產出不同的漢明窗。本論文所採用的 α 為 0.46。

（4）快速傅立葉轉換 (FFT)：將訊號由時域轉換為頻域

由於訊號在時域上的變化通常很難看出訊號的特性，所以通常將它轉換成頻域上的能量分佈來觀察其特性。不同頻率的能量分佈，可以代表不同語音的特性，所以在乘上漢明窗後，每個音框還必需再經過快速傅立葉轉換，以得到其在頻譜上的能量分佈。

（5）三角濾波器 (Triangular Bandpass Filters)：將頻譜能量乘一組三角濾波器

為了模擬人類耳朵處理語音訊號的特性，我們會將語音訊號送入一組模擬人類耳朵的濾波器去做處理，而這組濾波器在梅爾頻率上（Mel Frequency）是重疊且平

均分佈的。我們用式 (3.1.3) 來表示之：

$$Mel(f) = \beta \bullet \log_{10}(1 + \frac{f}{700}) \tag{3.1.3}$$

式 (3.1.3) 中，參數β在本論文中是使用 1127。而除了模擬人類耳朵的功能外，三角濾波器還有兩個重要的特性：第一是降低資料量，第二是對頻譜進行平滑化並消除諧波的作用，凸顯原語音的共振峰。此外，通常會對三角濾波器輸出的值再取對數，那是因為要模擬人類耳朵對於太大或太小的聲音有著自動調整敏感度範圍的功能。因此在取過對數後，特徵參數對於語音能量的變異便不會那麼敏感。本論文濾波器的數量為 18。

（6）離散餘弦轉換 (DCT)：將訊號由頻域轉回時域

將先前取過對數的三角濾波器輸出值，帶入離散餘弦轉換，求出 D 階的梅爾刻度倒頻譜係數，這裡 D 通常取 12。離散餘弦轉換公式如式 (3.1.4)：

$$C_i(n) = \sum_{k=1}^{N} \log \left| X_i(e^{j2\pi kn/N}) \right| \cos(n(k - \frac{1}{2})\frac{\pi}{k}) \quad n = 0,1,2,... \tag{3.1.4}$$

由於先前做了快速傅立葉轉換，所以在此離散餘弦轉換的目的，是希望能將訊號轉換到倒頻譜上。

（7）差量倒頻譜參數及能量(Delta Cepstrum & Energy)

雖然已經求出 12 個特徵參數，然而在實際應用於語音辨識時，我們通常會再加上差量倒頻譜參數，以顯示倒頻譜參數對時間的變化。它的意義為倒頻譜參數相對於時間的斜率，也就是代表倒頻譜參數在時間上的動態變化，我們用式 (3.1.5) 表示之：

$$c'_i(n) = \frac{\sum_{r=1}^{R} r(c_{i+r}(n) - c_{i-r}(n))}{2\sum_{r=1}^{R} r^2} \tag{3.1.5}$$

而能量在不同的音素間，也扮演了相當關鍵的角色，因此我們也會將其作為特徵參數的一部分。我們用式 (3.1.6) 表示之，N 代表濾波器的個數：

$$e = \sum_{i=1}^{N} \log b_i^2 \tag{3.1.6}$$

其中 b_i^2 代表語音訊號通過第 i 個濾波器後的能量，e 代表語音訊號通過所有濾波器後對數能量值的總合。由先前的 12 維梅爾倒頻譜係數，加上 1 維的能量係數共 13 維，對其求另外 13 維的差量倒頻譜參數（Delta Cepstrum Coefficients），再對該 13 維的差量倒頻譜參數求取差量倒頻譜參數（Delta Delta Cepstrum Coefficients），最後總共有 39 維，而這 39 維的特徵參數，便是本論文主要採用的特徵參數。表 3.1 是以華語廣播新聞資料庫為主，將上述諸步驟以及過程中所使用的參數做個整理：

預強調	0.975
音框化	取樣頻率為 16KHz 取樣點為 320 點 涵蓋時間為 20 ms
漢明窗	0.46
三角濾波器組	1,127，18 個濾波器
離散餘弦轉換	取 12 維
能量及差量倒頻譜	能量維 1 維 1st 及 2nd 差量倒頻譜各 13 維，總共 39 維

表 3.1　抽取參數的步驟以及相對應的參數值

3.2 聲學模型的介紹及辨識效能的評估

本論文所採用的語料庫有兩種，分別為 AURORA 2.0 以及華語廣播新聞。以下針對該兩種語料庫作個別的介紹。首先在 AURORA 2.0 部分，聲學模型是採用傳統的連續密度隱藏式馬可夫模型（Continuous Density Hidden Markov Model, CDHMM），模型內狀態的轉移情形只有兩種，一種是停留在原狀態，一種是由左至右跳到下一個相鄰的狀態。此外，除了阿拉伯數字 0 有 zero 和 oh 兩種聲學模型外，數字 1 至 9 分別都只有一個相對應的聲學模型。而每個模型都有 18 個狀態（States），包含了首尾兩個模型間連接用的空狀態，且每個狀態均包含 3 個高斯混合分佈（Gaussian Mixture Distributions）。除了數字的聲學模型外，另外還有兩個模型分別為靜音模型（Silence）和短暫停模型（Short Pause）。靜音模型包含 3 個狀態，每個狀態有 6 個高斯混合分佈。而短暫停模型包含 1 個狀態，此狀態與靜音模型最中間的狀態是共用的。

在華語廣播新聞部份，採用的也是連續密度隱藏式馬可夫模型。模型內狀態的轉移情形，亦是只有如上述的兩種狀況。模型的總數量有 151 個，其中包含了 1 個靜音模型（Silence），112 個聲母模型（Initials），以及 38 個韻母模型（Finals）。每個模型的狀態數分別為 3 至 6 個不等，每個狀態均有 16 個高斯混合分佈。此外，聲母和韻母共有 403 種不同的音節組合。

關於辨識率的部分，我們使用美國標準與科技組織所訂立的評估標準(US NIST F.O.M metric) [NIST] 來進行正確轉譯文句字串與辨識字串的比較，而 AURORA 2.0 是以詞正確率（Word Accuracy）評估各種語音強健技術的效用。其中考慮了詞消去率（Deletion），詞替換率（Substitution），以及詞插入率（Insertion）。定義如式 (3.2.1)：

$$\text{Word Accuracy} = 1 - \text{Deletion} - \text{Substitution} - \text{Insertion} \qquad (3.2.1)$$

另外華語廣播新聞則是以音節正確率（Syllable Accuracy）來評估各種語音強健技術的效用，評估的方式和式 (3.2.1) 是相同的，不過單位是音節而非字。

3.3 實驗語料庫介紹

3.3.1 AURORA 2.0

本語料庫是由歐洲電信標準協會(European Telecommunications Standards Institute, ETSI)[ETSI Website : http://www.etsi.org/] 所發行的語料：AURORA 2.0。其是藉由以人工的方式，加上八種來源不同的加成性噪音，分別是機場，人聲，汽車，展覽會館，餐廳，地下鐵，街道，火車站等，以及不同程度的訊噪比，分別是 -5dB，0dB，5dB，10dB，15dB，20dB，clean 等，來觀察噪音對訊號所造成的影響。其本身是一套含噪音的英語連續數字語料，參與錄音計劃的語者，皆是美國成年人;而 AURORA 2.0 是由 TIDigits 這套不含噪音的英語連續數字語料的部分內容，加上噪音而成，每個乾淨不含噪音的音段，會先通過特定的通道效應，再依照各種訊噪比，加上八種不同的加成性噪音，並以測試語料通過的通道效應，以及加成性噪音的種類，分成 A，B，C，三種測試組合。詳細如表 3.3.1 所示：

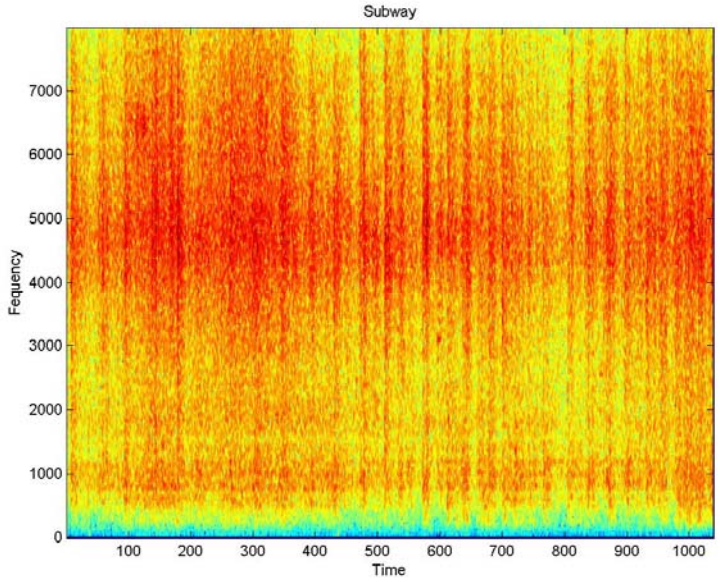
<table>
<tr><th colspan="4">AURORA 2.0</th></tr>
<tr><td>取樣頻率</td><td colspan="3">8KHz</td></tr>
<tr><td>語音內容</td><td colspan="3">英文數字：one，two，three，four，five，six，seven，eight，nine，zero，oh，共計 11 種發音。</td></tr>
<tr><td>音段長度</td><td colspan="3">包含一至七個連續數字</td></tr>
<tr><td>訓練模式</td><td colspan="2">乾淨語音訓練
音段數：
8440 句
通道效應：
G.712 的通道特性
加成性噪音：
無</td><td>複合情境訓練
音段數：
8440 句
通道效應：
G.712 的通道特性
加成性噪音：
@種類：地下鐵，人聲，汽車，展覽會館。
@訊噪比：20db，15dB，10dB，5dB 以及完全乾淨
@四種噪音以及五種訊噪比共 20 種情境</td></tr>
<tr><td>測試組合</td><td>A
音段數：28,028 句
通道效應：
G.712 的通道特性
加成性噪音：
－地下鐵
－人聲
－汽車
－展覽會館</td><td>B
音段數：28,028 句
通道效應：
G.712 的通道特性
加成性噪音：
－餐廳
－街道
－機場
－火車站</td><td>C
音段數：14,014 句
通道效應：
MIRS 的通道特性
加成性噪音：
－地下鐵
－街道</td></tr>
<tr><td colspan="4">對於上述每種加成性噪音訊噪比都控制在 20dB，15dB，10dB，5dB，0dB，-5dB，以及完全乾淨等七個程度，並且對於每種噪音的每一個訊噪比都計算一組辨識結果。</td></tr>
</table>

表 3.3.1 關於 AURORA 2.0 訓練語料與測試語料以及噪音介紹

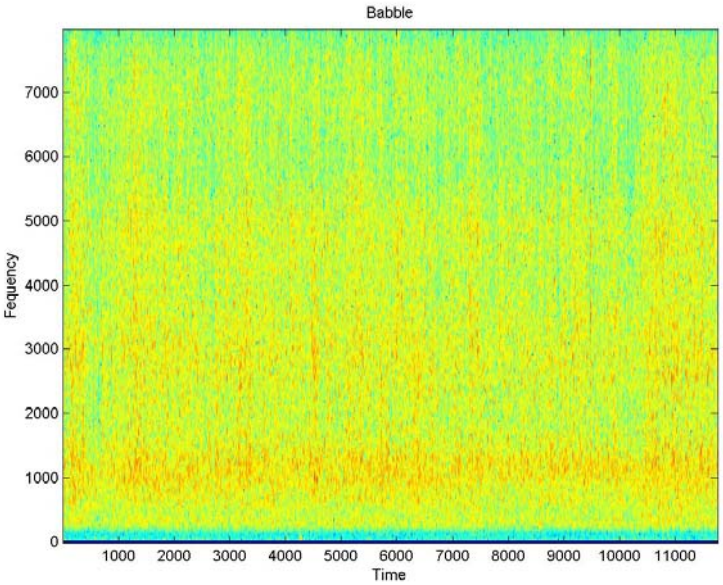
表 3.3.1 中的兩種通道效應，均是由國際電信聯合會所訂立的標準。其中 G.712 描述的是傳統電話線所使用之脈碼調變（Pulse Code Modulation, PCM）的頻道特性，而 MIRS 描述的則是類似手機 GSM（Global System of Mobile

Communications）的頻道特性。圖 3.3.1 是上述八種不同的噪音的頻譜-時間圖（Spectrogram），其中橫軸為時間，縱軸為頻率。顏色越亮的代表能量越強，越暗的代表顏色越弱：

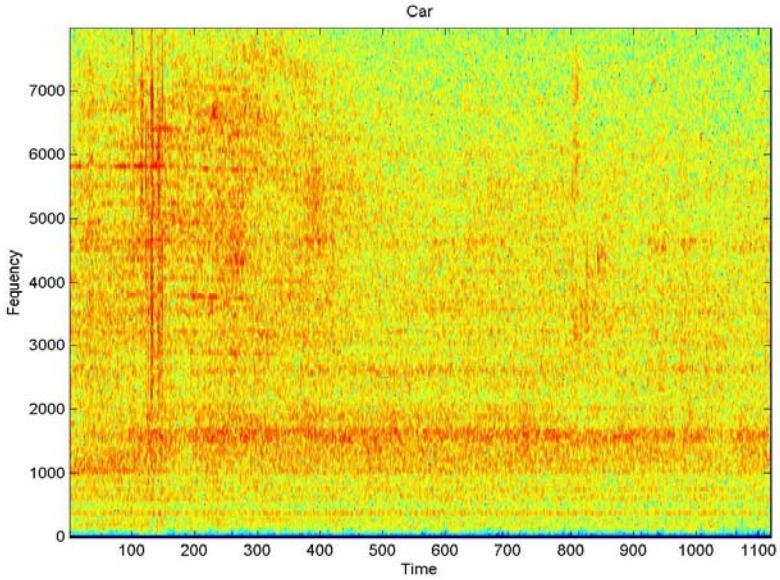
(a) 地下鐵噪音（Subway）



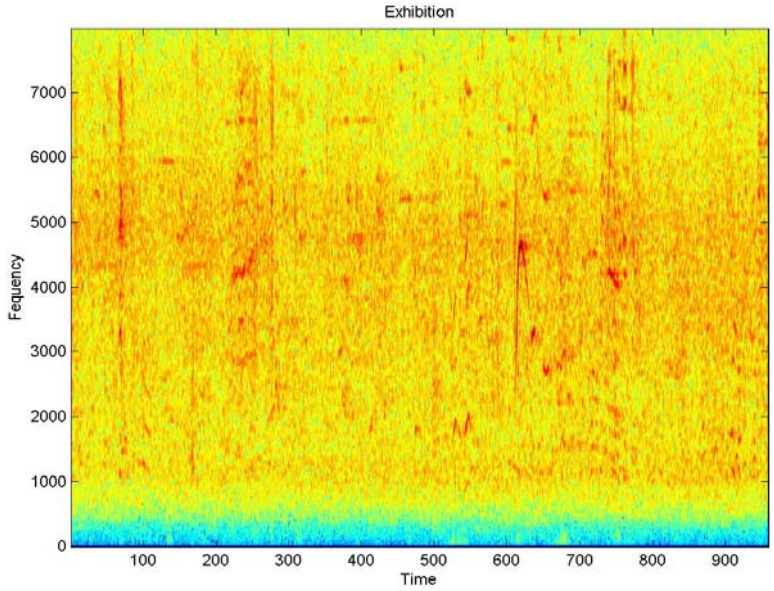
(b) 人聲噪音（Babble）



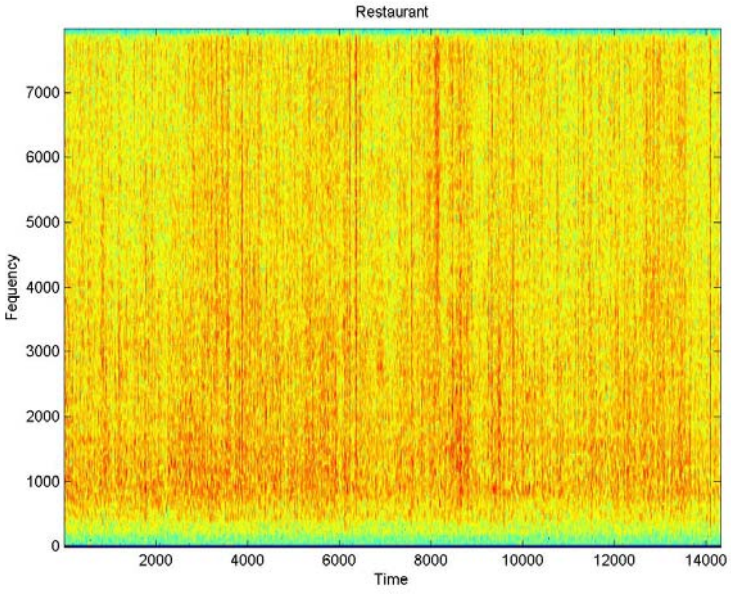
(c) 汽車噪音（Car）



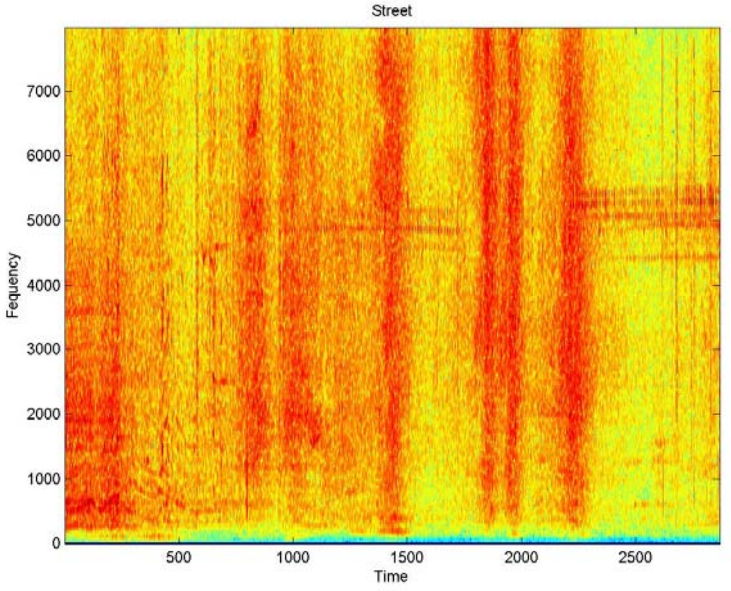
(d) 展覽會館噪音（Exhibition）



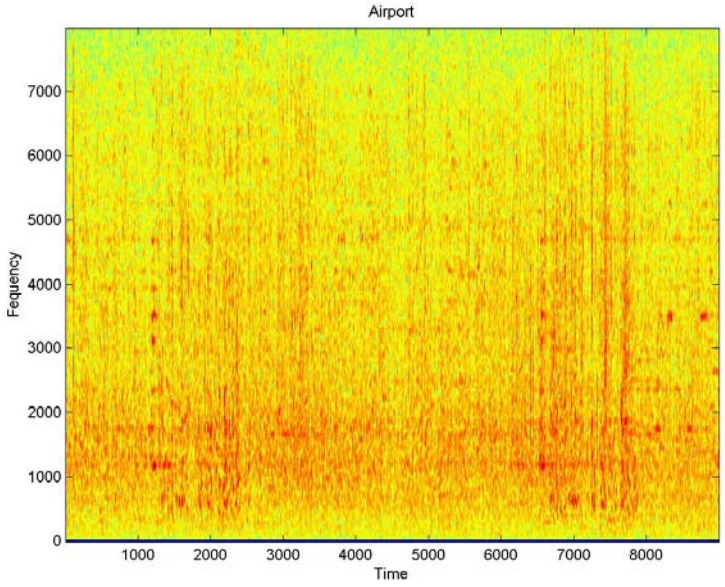
(e) 餐廳噪音（Restaurant）



(f) 街道噪音（Street）



(g) 機場噪音（Airport）



(h) 火車站噪音（Train Station）

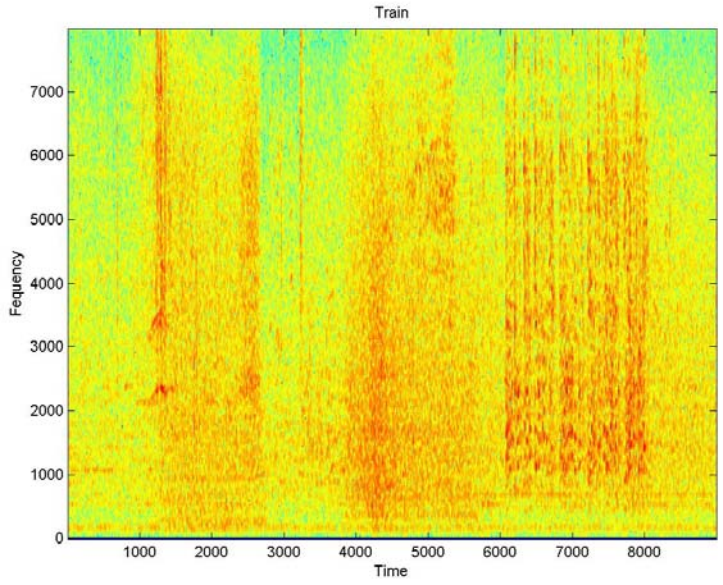


圖 3.3.1 各種噪音的頻譜時間圖

圖 3.3.1 中 (a) 至 (h) 分別是八種不同噪音的頻譜–時間圖。表 3.3.2 則是針對圖 3.3.1 中諸噪音的頻譜–時間圖所做的觀察：

(a) 地下鐵噪音	能量大都分布在 4~7KHz，次多分布在 1KHz。
(b) 人聲噪音	能量集中在 1KHz，且在 3，5，7KHz 時能量亦較高。
(c) 汽車噪音	前半段 1.5KHz 及中高頻能量集中，後半段隨著頻率遞減。
(d) 展覽館噪音	能量集中在 4~6KHz，整體分布平均。
(e) 餐廳噪音	能量大都集中在低頻，能量隨時間有增有減。
(f) 街道噪音	能量分布極度不均衡,特色是有能量時,低頻高頻都會很強。
(g) 機場噪音	能量集中在 1KHz，且隨頻率增加而遞減。
(h) 火車站噪音	能量分布成週期性,前段分布於中高頻,後段分布於中低頻。

表 3.3.2 針對噪音頻譜-時間圖所做的觀察

由圖 3.3.1 以及表 3.3.2 我們可以觀察到,汽車噪音以及展覽館噪音的能量分部並不會隨著時間點的不同有太大的差異性，因此是屬於比較穩定的噪音；而機場，街道，以及火車站，隨著時間的不同能量分布的差異性頗大，所以是屬於比較不穩定的噪音。

3.3.2 華語廣播新聞

本論文大部分的實驗，所使用的語料都是以廣播新聞為主的語料庫，該語料庫的內容介紹如表 3.3.3：

訓練語料	期間為 1999 年 3 月 25 日至 9 月 22 日 包含 9 個主播所播報的新聞，共取 5471 句，總共約 240 分鐘。 平均句長為 2.56 秒。
測試語料	期間為 2001 年 8 月 1 日至 8 月 24 日 隨機取 17 天共 89 句，共約 45 分鐘。 平均句長為 27.78 秒。

表 3.3.3 華語廣播新聞資料簡介

3.4 AURORA2.0 基礎系統實驗結果

本小節，吾人針對本論文所使用的 AURORA 2.0 語料庫，求其個別的基礎實驗結果，且採用的是如表 3.1 所摘要出的 39 維 MFCC 特徵參數。圖 3.4.1 是 AURORA 2.0 所作的實驗結果。AVG 依照慣例取的是 20dB 至 0dB 的平均值：

測試組別 A								
	Subway		Babble		Car		Exhibition	
	Clean	Multi	Clean	Multi	Clean	Multi	Clean	Multi
Base	97.24	95.76	97.16	96.07	97.11	96.00	97.10	95.93
20	94.90	95.36	94.35	95.59	95.11	95.20	94.42	95.34
15	89.62	94.38	85.19	93.86	90.99	94.48	89.79	94.23
10	72.24	91.68	64.60	90.42	72.95	91.77	73.80	90.90
5	43.26	82.16	37.06	81.86	28.21	81.87	40.33	82.14
0	18.39	56.43	14.18	60.07	7.96	50.43	11.97	53.44
-5	9.26	19.44	8.34	29.69	7.04	18.31	8.30	19.56
AVG	63.68	84.00	59.08	84.36	59.04	82.75	62.06	83.21
備註	測試組別 A 的通道效應: G.712 傳統電話線之脈碼調變 PCM (Pulse Code Modulation)							

表 3.4.1 (a) Aurora 2.0 測試組 A 的基礎實驗結果

本圖的實驗結果是測試語料 A 組中，四種不同噪音在兩種訓練情境下，對於不同程度的噪音干擾所呈現的辨識結果。Clean 是乾淨環境下訓練出來的模型。Multi 是複合情境下訓練出來的模型。四種噪音由左至右分別是地下鐵，人聲，汽車，展覽會館。通道效應為 G.712。

測試組別 B								
	Restaurant		Street		Airport		Train Station	
	Clean	Multi	Clean	Multi	Clean	Multi	Clean	Multi
Base	97.24	95.76	97.16	96.07	97.11	96.00	97.10	95.93
20	94.84	94.14	94.74	95.98	95.32	95.29	95.80	94.82
15	89.13	93.15	88.51	94.74	92.31	94.54	92.10	93.52
10	74.09	89.87	64.68	92.29	80.50	92.31	79.79	90.93
5	47.74	82.22	33.86	82.86	51.92	86.82	45.51	82.32
0	19.90	63.77	14.42	57.80	20.61	69.46	12.50	54.64
-5	9.12	31.56	8.16	25.27	9.33	33.37	8.05	21.88
AVG	65.14	84.63	59.24	84.73	68.13	87.68	65.14	83.25
備註	測試組別 B 的通道效應: G.712 傳統電話線之脈碼調變 PCM (Pulse Code Modulation)							

表 3.4.1 (b) Aurora 2.0 測試組 B 的基礎實驗結果

本圖的實驗結果是測試語料 B 組中，四種不同噪音在兩種訓練情境下，對於不同程度的噪音干擾所呈現的辨識結果。Clean 是乾淨環境下訓練出來的模型。Multi 是複合情境下訓練出來的模型。四種噪音由左至右分別是餐廳，街道，機場，火車站。通道效應為 G.712。

測試組別 C				
	Subway		Street	
	Clean	Multi	Clean	Multi
Base	97.21	95.86	97.10	96.01
20	93.37	95.15	93.53	95.34
15	86.18	94.01	88.00	94.65
10	71.60	89.99	68.56	90.45
5	43.05	76.85	37.18	79.02
0	18.18	42.68	14.18	48.19
-5	10.19	14.03	8.74	21.52
AVG	62.48	79.74	60.29	81.53
備註	測試組別 C 的通道效應: MIRS 類似手機上 GSM 的頻道特性			

表 3.4.1 (c) Aurora 2.0 測試組 C 的基礎實驗結果

本圖的實驗結果是測試語料 C 組中，兩種不同噪音在兩種訓練情境下，對於不同程度的噪音干擾所呈現的辨識結果。Clean 是乾淨環境下訓練出來的模型。Multi 是複合情境下訓練出來的模型。兩種噪音分別是地下鐵，街道。通道效應為 MIRS。

	乾淨語音訓練	複合情境訓練	AVG
測試組別 A	60.97	83.58	72.28
測試組別 B	64.41	85.07	74.74
測試組別 C	61.39	80.64	67.94
AVG	62.26	83.10	71.65

表 3.4.1 (d) 本圖的實驗結果是三組測試語料在兩種訓練情境下的綜合辨識結果

由表 3.4.1 中，我們可以觀察到幾個現象如下：

第一：

無論是哪個測試組別或是哪種噪音，隨著訊噪比的值下降，辨識率也一致的跟著降低，這證明了噪音對於語音本身和辨識系統，確實是有不小的影響。

第二：

大致上而言，複合情境訓練模式的辨識率普遍的都比乾淨語音訓練模式來的好，唯獨在不加入任何噪音時，這是因為對於乾淨語音訓練模式來說，由於存在於語料和模型間不匹配的情況是比較小的，而對於複合情境訓練模式，存在於語料和模型間不匹配的情況則是比較嚴重的，因此前者的辨識率才會比較高。

第三：

測試 A 組的地下鐵辨識率，無論在哪種訓練模式，都比測試 C 組的辨識率高，這是很合理的現象，那是因為測試 C 組的通道效應和訓練語料是不同的。然而測試 B 組的街道辨識率，無論是複合情境訓練模式或是乾淨語音訓練模式，理論上來說均應該要比測試 C 組的街道辨識率高，然而實驗結果卻不完全是如此。合理的解釋是，C 組的通道效應 MIRS，對於低頻的頻率響應是比較低的，而街道的能量大致上都集中在低頻，因此通道效應 MIRS 減少街道噪音對語音的影響。

第四：

測試 B 組所採用的噪音，其能量分布大都集中於較低頻處，這和人類語音的能量分布集中在頻率較高處 [Lord *et al.* 1980] 是有明顯的差異的，因此測試 B 組的辨識率，普遍來說都比測試 A 組來的高。而特性和語音類似的人聲噪音，則比較不利於語音辨識。