第二章 實驗架構及語料庫環境

本章節主要是介紹論文中與實驗相關的設定，以及所使用的語料庫特性。第一小節將介紹實驗語料庫，第二小節則說明本實驗所採用的特徵參數擷取方法與步驟，最後是介紹聲學模型的建立及辨識效能的評估。

2.1 實驗語料庫

論文中使用的語料庫為歐洲電信標準協會(European Telecommunications Standards Institute, ETSI)[ETSI 2000] 所發行的語料：Aurora-2.0 為主。語料內容是藉由以人工的方式特別錄製的連續英文數字語料，語者為成年男女各半數，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 和 Clean 等；通道效應則包含由國際電信聯合會所訂立的二個標準 G.712 和 MIRS。根據測試語料中加入之通道噪音以及加成性噪音之種類不同，Aurora-2.0 分為三組測試群組 Set A、Set B 和 Set C，Set A 所呈現的噪音是屬於穩定性(Stationary)噪音，Set B 則是非穩定性(Nonstationary)噪音，Set C 除了穩定性與非穩定性噪音外，額外使用與訓練語料不同的通道效應(Channel Effects)。詳細情形如表 2.2.1 所示。

表 2.2.1 中的兩種通道效應，分別為 G.712 與 MIRS，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 MobileCommunications)的頻道特性。其中訊噪比(Signal-to-noise ratio, SNR)的單位為分貝(Decibel, dB)。

<table>
<tr><td colspan="4">AURORA 2.0</td></tr>
<tr><td>取樣頻率</td><td colspan="3">8KHz</td></tr>
<tr><td>編碼格式</td><td colspan="3">16 位元 PCM，無檔頭</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>語音長度</td><td colspan="3">語料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</td><td>複合情境訓練</td></tr>
<tr><td></td><td colspan="2">音段數：
8440 句
通道效應：
G.712 的通道特性
加成性噪音：
無</td><td>音段數：
8440 句
通道效應：
G.712 的通道特性
加成性噪音：
地下鐵、人聲、汽車
與展覽會館
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>測試組 A</td><td>測試組 B</td><td>測試組 C</td></tr>
<tr><td>對於右側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><td>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>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>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>
</table>

表 2.1.1 關於 AURORA 2.0 訓練語料與測試語料以及噪音介紹

2.2 特徵參數擷取

在本論文中所使用的語音特徵參數為梅爾倒頻譜參數(Mel-frequency Cepstral Coefficients, MFCC) [Davis et al. 1980]，主要目的在於模擬人耳聽覺感知特性[Hermansky 1998]作為初步處理，藉此達到降維、增強語音訊號的效果。梅爾倒頻譜參數(MFCC)的語音特徵擷取架構圖如圖 2.1.1。梅爾倒頻譜參數的計算從取框(Framing)開始，經過預強(Pre-emphasis)、漢明窗(Hamming Window)處理直到離散傅立葉轉換(Discrete Fourier Transform, DFT)將時域信號轉換成頻域成份，其後將功率頻譜(Power Spectrum)經由在梅爾頻率(Mel Frequency)平均分佈的三角濾波器組處理，最後對各個濾波器的輸出所形成的向量進行離散餘弦轉換。圖 2.2.1 的各個部分將作為簡要介紹特徵參數擷取時的主要步驟流程：

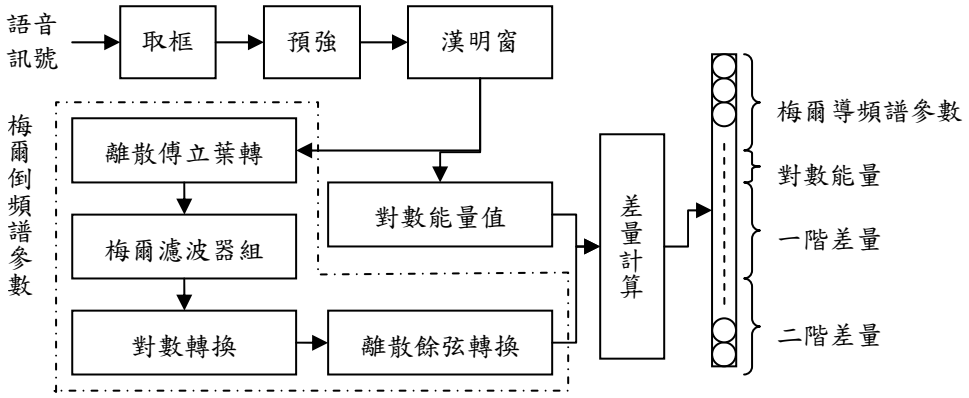

圖 2.2.1 特徵參數擷取流程圖

(1) 取框(Framing)

大部分的語音信號可被視為短時域穩定(Short-Term Stationary)，或稱為半穩定(Quasi-Stationary)的訊號(在語音學中的半穩定就相當於數位訊號處理中的非時變訊號)。由於語音訊號是時變的信號，任何欲測量的語音特徵是隨著時間而改變，這使得我們無法以線性非時變的方法來分析長時域(Long-Term)的語音信號

特徵。假若我們將訊號劃分為數個連續的音框(Frames)，則將有助於我們用非線性時變的模型來分析短時域語音訊號的特性。所以在語音辨識的前處理，會假設語音訊號為短時間穩定，因此利用取框的概念，來取得獨立的音框。主要目的是在限制輸入資料的長度，使得此音框的頻域特性在其長度之中是合理地穩定(Reasonably Stationary)。從頻域上觀察，可以發現在短時間(20ms~40ms)的情況下頻譜的變化是具有週期規則性。然而，為了讓音框與音框之間能夠保持前後的關連性，強調目前音框與下一音框的相互影響，在音框與音框之間會重複一小段時間，此完整動作稱為取框(Framing)。本論文中在 Aurora 2.0 語料庫上的實驗設定上，取樣頻率為 8KHz，每音框取樣點數為 200 個樣本點，其單位音框的涵蓋時間為 25 ms。

(2) 預強(Pre-emphasis)

預強功用是將語音訊號通過一個高通濾波器(High-Pass Filter) ，主要是加強聲波高頻的部份。由於人嘴唇所發出的聲音，受到傳播時輻射效應的影響，使得收聽到的語音其頻譜具有隨著頻率增加而強度降低的特性。但人類耳朵的外聽道約 2.5 至 3 公分長，其共振作用可以提高 2000~5000Hz 聲音的強度，剛好可以彌補高頻能量的損失，故能自動補償此效應。

$$H(z) = 1 - \alpha \cdot z^{-1} \tag{2.2.1}$$

式(2.2.1)可以用來表達預強，其中 $H(z)$ 為高通濾波器在 Z 轉換(Z-Transform)的表示。實作上可以在時域上處理如式(2.2.2)，其中 $s(n)$ 為第 n 個採樣點，$\hat{s}(n)$ 為第 n 個採樣點經預強後的值。α 為預強的參數，本論文設定為 0.975。

$$\hat{s}(n) = s(n) - \alpha \cdot s(n-1) \tag{2.2.2}$$

(3) 漢明窗(Hamming Window)

由於每個音框都會經由離散傅立葉轉換成頻域的訊號，但每個音框是設定在有限的固定時間點，所以音框左端和右端的邊緣會造成訊號不連續現象，會使得頻域(Frequency Domain)上產生摺積的效果，所以在離散傅立葉轉換前會乘上一個漢明窗，特性在於主辦葉(Main Lobe)較寬，邊葉(Side Lobe)較窄，因此能有效的壓抑訊號的二端，聚集中央部份的特徵。漢明窗的公式如下，其中 α 為控制漢明窗的參數，本論文設定為 0.46。

$$w(n)=\begin{cases}(1-\alpha)-\alpha\cos\left(\dfrac{2\pi n}{N-1}\right) & n=0,1,\ldots,N-1\\ 0 & otherwise\end{cases} \tag{2.2.3}$$

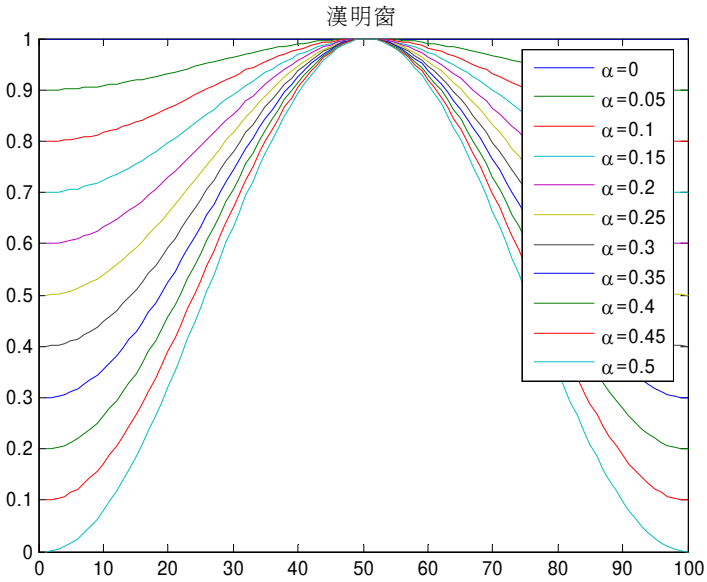

圖 2.2.2 不同 α 參數下的漢明窗示意圖

(4) 離散傅立葉轉換(Discrete Fourier Transform)

語音訊號在時域上變化迅速且會隨著時間不斷的改變並且不容易觀察出週期性

的變化，使得在時域上沒有辦法作有效的觀察。為了找出語音訊號的特性，可以轉換到頻域上做觀察，因為短時間內語音訊號在頻域上的能量分佈是有規律性的，所以一般會可以經由離散傅立葉轉換轉換(DFT)，換式如下：

$$X_i[n] = \sum_{n=0}^{N-1} x_i[n] e^{-\frac{j2\pi nk}{N}} \quad , \ 0 \le k < N \tag{2.2.4}$$

式中 x_i 是第 i 個音框向量，$x_i[n]$ 為第 i 個音框向量中的第 n 個值，N 為頻域上取樣點數。實作上則常會使用快速傅立葉轉換(Fast Fourier Transform, FFT) 取代離散傅立葉轉換以增加計算速度。但是用快速傅立葉轉換在音框的取樣點數上必須限制在 2 的倍數，不足 2 的倍數部分必須要補上零值，此補零方式會造成傅立葉轉換轉換的誤差。

(5) 梅爾頻率濾波器組(Mel-frequency Filter Bank)

人耳聽覺系統中不同頻率是由不同位置的耳蝸(Cochlea)神經反應來接受不同的頻率成分。梅爾頻率濾波器組藉著模擬耳蝸內部基底膜(Basiler Membrane)傳遞刺激到聽覺神經的方式，以達到語音資訊的擷取。梅爾頻率濾波器組作法上可以分為梅爾頻率(Mel Frequency)及三角濾波器(Triangular Bandpass Filters)兩個部分[Davis 1980]，而除了模擬人類耳朵的功能外，三角濾波器還有另外兩個功能，第一個功能是降低資料量，第二個功能是對頻譜進行平滑化並消除諧波(Harmonic)的作用，保留原本語音的共振峰(Formant)。

人耳對於頻率的變化在高頻與低頻時的敏感度不同，人耳感受在低頻部分是比較敏銳的，而在高頻部分人耳的感受就會越來越粗糙。在相對低頻時，對於頻率變化的感受是呈線性的；而當頻率大於 1KHz 時，人耳對於頻率的感受是呈對數變化的。所以梅爾頻率便將此對數變換模擬化，式子如下，其中 β 參數通常設定為 1127。

$$Mel(f) = \beta \log_{10}(1 + \frac{f}{700}) \tag{2.2.5}$$

三角濾波器部份，研究發現人耳聽覺神經不只接受單一特定頻率刺激，會被此受到一定範圍內的頻率影響，而距離某特定頻率越遠影響會越小。所以在經過梅爾頻率的對數轉換後，必須再通過 M 個三角帶通濾波器處理，使梅爾頻率上平均分佈來模擬人耳聽覺特性，三角濾波器的公式如下：

$$H_m[k]=\begin{cases}0 & ,k<f[m-1]\\ \dfrac{k-f[m-1]}{f[m]-f[m-1]} & ,f[m-1]\le k\le f[m]\\ \dfrac{f[m+1]-k}{f[m+1]-f[m]} & ,f[m]\le k\le f[m+1]\\ 0 & ,k>f[m+1]\end{cases} \tag{2.2.6}$$

其中 $f[m]$ 為第 m 個三角濾波器的中心點，$H_m[k]$ 為 k 頻率在第 m 個三角濾波器的權重(Weight)，N 為音框大小。$f[m]$ 可進一步表示成：

$$f[m]=(\frac{N}{F_s})Mel^{-1}\left(Mel(f_L)+m\cdot\frac{Mel(f_K)-Mel(f_L)}{M+1}\right) \tag{2.2.7}$$

其中 F_s 為取樣頻率，f_L 為三角濾波器組中最低的頻率，f_K 為三角濾波器組中最高的頻率，M 為三角濾波器組的個數。本論文共取 23(即 M=23)個三角濾波器。

(6) 對數轉換(Logarithm)

由於在人耳的構造上，外耳與中耳所接觸的音波振動藉由三小聽骨傳遞到後方的內耳，因此當音波振動由空氣傳遞到液體的過程已經造成能量的損失，所以人耳除了對於頻率的變化會隨著高頻而敏感度遞減，此外對於頻率能量的變化現實上是不敏感的。在此，實驗設定上為達到模擬人耳的特性，所以一般會對梅爾三角濾波器輸出的值作對數運算。

(7) 離散餘弦轉換(Discrete Cosine Transform, DCT)

梅爾倒頻譜參數求取的最後一個步驟，就是把對數轉換後的三角濾波器輸出再經

由離散餘弦轉換(Discrete Cosine Transform, DCT)，目的是希望能將訊號轉換到倒頻譜上(Cepstrum)。主要用意在於減少維度間的關係，有助於隱藏式馬可夫模型在儲存共變異矩陣時資料的縮減，其次能夠再次做降低特徵維度的動作，增加辨識系統的效率。梅爾倒頻譜參數表示為式(2.2.8)：

$$c[n]=\sqrt{\frac{2}{M}}\sum_{j=1}^{M}\log\left(Mel_j\right)\cos\left(\frac{n\cdot\pi}{M}\left(j-0.5\right)\right) \quad , \quad n=0,1,\cdots,L<M \tag{2.2.8}$$

其中 $c[n]$ 表示語音特徵向量中第 n 維的特徵值，L 為語音特徵向量的總維度個數，M 三角帶通濾波器的個數，Mel_j 表示第 j 個梅爾三角濾波器輸出值。

(8) 語音對數能量計算(Log Energy)

對數能量特徵參數，在不同音素(Phoneme)之間的差異頗大，因此對數能量在語音特徵上亦扮演著重要的角色，而本論文便是著重於語音能量計算上的探討與研究。式(2.2.9)為對數能量的計算，N 為音框樣本點數，其中 x_i^2 代表語音訊號樣本點上第 n 個的能量值，$LogE_i$ 則代表第 i 個音框訊號之對數能量。

$$LogE_i=\log\sum_{n=1}^{N}x_i^2(n) \tag{2.2.9}$$

(9) 時間差量計算(Time Derivatives)

由於假設語音訊號在短時間內是穩定的，所以每隔一短時間取得一個音框，但實際上已經造成語音訊號的破壞。為了補償音框與音框間在時間軸上的連續性關係，因此在取得梅爾倒頻譜參數與對數能量的 L 維語音特徵向量外，會再加上一階差量 $\Delta C_t[n]$ (Fist-order Differential) 與二階差量 $\Delta^2 C_t[n]$ (Second-order Differential)，計算方式分別如下所示：

$$\Delta C_t[n] = \frac{\sum_{p=1}^{P} p(C_{t+p}[n] - C_{t-p}[n])}{2 \cdot \sum_{p=1}^{P} p^2} \tag{2.2.10}$$

$$\Delta^2 C_t[n] = \frac{\sum_{p=1}^{P} p(\Delta C_{t+p}[n] - \Delta C_{t-p}[n])}{2 \cdot \sum_{p=1}^{P} p^2} \tag{2.2.11}$$

其中 n 為梅爾倒頻譜參數加上能量共 13 維，$c_t[n]$ 為時間點 t 上第 n 維的梅爾倒頻譜參數，P 為音框前後的考量個數。加入兩階的差量計算，最後特徵擷取的維度為 39 維。特徵擷取的參數詳細如列表 2.2.1。

取樣頻率	8 KHz
音框點數	200 點, 25ms
音框重複	80 點, 10ms
預強	0.97
漢明窗	0.46
三角濾波器	23 組
梅爾倒頻譜係數	12 維
對數能量	1 維
差量計算	梅爾倒頻譜係數 12 維加對數能量 1 維,取一階與二階差量倒頻譜各 13 維，總共 39 維

表 2.2.1 本論文中使用之語音特徵參數設定

2.3 聲學模型

在聲學模型(Acoustic Models)的設定，每個數字模型(1~9 及 zero 和 oh)皆由一個由左到右(left-to-right)形式的連續密度隱藏式馬可夫模型(Continuous Density Hidden Markov Model, CDHMM)表示，其中每個模型包含有 16 個狀態(States)與首尾兩個模型間連接用的空狀態，共 18 個狀態來表示每一個模型。而在每個狀態內則利用 3 個高斯混合分佈(Gaussian Mixture Distributions)表示。另外靜音模型的部份採用二種型式，一種為長靜音(Silence)模型，內包含 3 個狀態，每個狀態有 6 個高斯混合分佈，主要用來表示語句開始跟結束時的靜音；另一個為間歇(Short Pause)模型，包含 1 個狀態，表示語句內字與字之間的短暫停止，上述所有聲學模型的訓練與本論文所有的實驗則是使用英國劍橋大學電機系所發展出來的隱藏式馬可夫模型(HMM)開發的 HTK 工具套件[HTK toolkit 2006]完成。

2.4 辨識效能評估

在辨識效能的評估方面，實驗數據計算方式我們參考的是美國標準與科技組織(National Institute of Standards and Technology)所訂立的評估標準(US NIST F.O.M metric) [NIST]與 HTK 工具套件的設定值，並藉由 HTK 方便快速地工具套件做辨識字串的比較。實驗數據則依 Aurora-2.0 標準設定，以詞正確率(Word Accuracy)評估各種語音強健技術的效果。然而計算詞正確率需要利用到動態程式設計(Dynamic Programming)來做詞(單字詞)對齊(Alignment)，其中考慮有輸入詞總數、詞取代個數(Substitutions)、詞插入個數(Insertions)和詞刪除個數(Deletions)。本論文中所有實驗皆以詞正確率(Word Accuracy)百分比來表示，定義如式(2.4.1)：

$$\text{詞正確率}(\%) = \frac{\text{輸入詞總數} - (\text{詞取代個數} + \text{詞插入個數} + \text{詞刪除個數})}{\text{輸入詞總數}} \times 100\% \tag{2.4.1}$$

2.5 基礎實驗結果

根據 Aurora-2.0 實驗語料庫標準設定，首先我們求其個別情況下的基礎實驗(Baseline Experiment)，每一個音框由 12 維的梅爾倒頻譜特徵值與 1 維的對數能量加上其一階與二階的時間軸導數(Time Derivatives)所形成的 39 維語音特徵向量所組成。其中 12 維的梅爾倒頻譜特徵是由 23 個梅爾頻譜上濾波器組的輸出經餘弦轉換求得，結果如下表：表中乾淨環境 (Clean-Condition) 訓練模式與複合情境 (Multi-Condition) 訓練模式，如表 2.1.1 關於 Aurora-2.0 的訓練語料、測試語料以及噪音設定，其中分別表示乾淨環境(Clean-Condition) 訓練語料和複合情境(Multi-Condition)訓練語料所訓練的聲學模組針對不同噪音干擾下的辨識結果。噪音干擾程度則分別為乾淨無干擾狀況、20dB 狀況到訊噪比-5dB 狀況。

測試組 A	乾淨環境訓練模式					複合情境訓練模式				
訊噪比	地下鐵	人聲	汽車	展覽會館	平均	地下鐵	人聲	汽車	展覽會館	平均
Clean	98.99	99.00	98.87	99.11	98.99	98.22	98.34	98.48	98.36	98.35
20dB	95.30	90.63	95.82	95.19	94.24	96.28	96.58	97.55	97.13	96.89
15dB	87.35	74.15	86.07	89.54	84.28	94.50	94.53	97.02	95.71	95.44
10dB	68.65	51.45	64.21	73.13	64.36	90.76	90.78	94.90	93.03	92.37
5dB	39.76	28.75	34.00	45.11	36.91	82.53	80.71	86.88	86.36	84.12
0dB	14.43	14.03	13.54	17.65	14.91	58.61	56.89	54.28	59.30	57.27
-5dB	7.95	7.80	7.93	9.04	8.18	22.84	23.43	18.01	20.95	21.31
平均	61.10	51.80	58.73	64.12	58.94	84.54	83.90	86.13	86.31	85.22

表 2.5.1 AURORA 2.0 測試組 A 的基礎實驗結果

表 2.5.1 為測試組 A 組噪音環境下的基礎實驗數據，表中縱軸為各噪音在訊噪比 20dB~0dB 的平均值，橫軸則為四種噪音由左至右分別是地下鐵、人聲、汽車、展覽會館在同一訊噪比情況下的平均值。表中以展覽會館噪音(Exhibition)對語音訊號的影響較小，而人聲(Babble)的干擾較為嚴重。

測試組 B	乾淨環境訓練模式					複合情境訓練模式				
訊噪比	餐廳	街道	機場	火車站	平均	餐廳	街道	機場	火車站	平均
Clean	98.99	99.00	98.87	99.11	98.99	98.22	98.34	98.48	98.36	98.35
20dB	92.63	95.04	93.17	95.65	94.12	95.64	97.04	97.55	96.20	96.61
15dB	79.61	85.67	82.14	86.73	83.54	91.93	95.28	96.06	94.08	94.34
10dB	59.20	64.45	60.36	65.47	62.37	87.07	92.56	93.26	91.98	91.22
5dB	34.08	37.79	34.98	34.87	35.43	76.76	80.53	85.54	82.14	81.24
0dB	14.37	20.47	18.01	14.96	16.95	53.76	54.11	64.12	54.12	56.53
-5dB	7.40	9.89	9.04	8.08	8.60	21.86	21.07	28.99	18.14	22.52
平均	55.98	60.68	57.73	59.54	58.48	81.03	83.90	87.31	83.70	83.99

表 2.5.2 AURORA 2.0 測試組 B 的基礎實驗結果

表 2.5.2 為測試組 B 組噪音環境下的基礎實驗數據，表中以火車站噪音(Train Station)對語音訊號的影響較小，而街道(Street)的干擾較為嚴重。

測試組 C	乾淨環境訓練模式			複合情境訓練模式		
訊噪比	地下鐵	街道	平均	地下鐵	街道	平均
Clean	99.20	99.09	99.15	98.37	98.28	98.33
20dB	87.75	91.72	89.74	96.59	96.19	96.39
15dB	78.57	84.79	81.68	94.78	94.68	94.73
10dB	62.17	68.71	65.44	91.83	91.41	91.62
5dB	38.23	46.01	42.12	77.43	77.45	77.44
0dB	17.59	24.18	20.89	40.90	45.44	43.17
-5dB	10.13	13.15	11.64	14.61	17.59	16.10
平均	56.86	63.08	59.97	80.31	81.03	80.67

表 2.5.3 AURORA 2.0 測試組 C 的基礎實驗結果

表 2.5.3 測試組別 C 的為 MIRS 通道效應與手機上 GSM 的頻道特性相同，兩種噪音分別是地下鐵與街道，從表中可以發現地下鐵的噪音干擾比較嚴重。

基礎實驗探討：

依據表 2.5.1 至 2.5.3 的三類測試組別之辨識結果。首先可以發現實驗中的三類測試組在各噪音干擾環境下，隨著訊噪比值的下降，表示噪音干擾程度越強，辨識率也會同時跟著降低，證明了噪音對於語音辨識系統，確實是有嚴重的影響。其

次，複合情境訓練模式在相同訊噪比的噪音干擾環境，對照實驗表格中數據除乾淨情境(Clean)外的辨識率都比乾淨環境訓練模式來的好。然而在不加入任何噪音時的乾淨環境，由於對於乾淨環境訓練模式來說，存在於訓練語料和測試語料之間不匹配(Mismatch)的情況比較小，相對於複合情境訓練模式，存在於訓練語料和測試語料間不匹配的情況則是比較嚴重的，因此乾淨環境訓練模式的辨識率才會比較高。

至於各組別之正確率比較，我們可以得知在乾淨環境訓練模式的平均正確率相差不大，而在複合情境訓練模式的平均正確率則約各相差 2 個百分比的差距，此原因我們可以從實驗設定中合理地推論，在乾淨環境訓練模式下所訓練的聲學模組對於測試語料的噪音干擾並沒有特別的關係存在，然而在複合情境訓練模式下的加成性噪音包含有地下鐵、人聲、汽車與展覽會館，因為複合情境訓練模式所訓練的聲學模組對於測試組 A 的噪音有相同噪音環境干擾的關係，所以測試組 A 的平均正確率會較高，其次測試組 B 則是因為 G.712 的通道特性與測試組 A 相同，所以平均正確率會略差一些。最後測試組 C 變差的原因是加成性噪音為 MIRS 的通道效應與複合情境訓練模式下的 G.712 通道特性相互不匹配而造成，所以平均正確率表現最差。