國立臺灣師範大學

資訊工程研究所碩士論文

指導教授：陳柏琳 博士

以能量為基礎之語音正規化方法研究及其於語音端點偵測之應用

On the Study of Energy-Based Speech Feature Normalization and Application to Voice Activity Detection

研究生：陳鴻彬 撰

中華民國 九十六 年 七 月

研究摘要

本論文主要探討強健(Robust)性語音辨識技術在不同噪音環境下的情況，並且於時間軸上研究雜訊語音(Noisy Speech)在對數能量上重建出乾淨語音(Clean Speech)對數能量的方法。基於每一語句對數能量特徵值的分佈特性，我們期望發展出一個有效的方法可以重刻雜訊語音對數能量的尺度，以減緩噪音環境所造成不匹配的情形，並達到更好的辨識率效果。

根據時間軸上的語音訊號觀察顯示，低能量的語音音框比高能量的語音音框更容易受到加成性噪音(Additive Noise)的影響，並且當出現嚴重的加成性噪音影響的時候，對數能量特徵強度在語句中幾乎會整個被提高，因此我們提出一個簡單但是有效的方法，稱之為對數能量尺度重刻正規化技術(Log Energy Rescaling Normalization, LERN)，適當的重刻雜訊語音的對數能量特徵值使成為接近乾淨語音的環境狀況。

語音辨識實驗採用的是包含多種噪音環境的語料，該語料是由歐洲電信標準協會(European Telecommunications Standards Institute, ETSI)所發行的 Aurora-2.0 語料庫，語料庫內容為英語發音的連續數字字串的小詞彙。提供有八種噪音來源和七種訊噪比(Signal-to-Noise Ratio, SNR)的情況。實驗方面，結果顯示對數能量尺度重刻正規化方法(LERN)的效果比其他的能量或對數能量上的正規化方法好。此外，另一組實驗則採用中文廣播新聞語料庫(Mandarin broadcast news corpus, MATBN)在大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)上的測試，並證明對數能量尺度重刻正規化方法(LERN)依然可以有效提升辨識率。

Abstract

This thesis considered robust speech recognition in various noise environments, with a special focus on investigating the ways to reconstruct the clean time-domain log-energy features from the noise-contaminated ones. Based on the distribution characteristics of the log-energy features of each speech utterance, we aimed to develop an efficient approach to rescale the log-energy features of the noisy speech utterance so as to alleviate the mismatch caused by environmental noises for better speech recognition performance.

As the time-domain phenomena of the speech signals reveal that lower-energy speech frames are more vulnerable to additive noises than higher-energy ones, and that the magnitudes of the log-energy features of the speech utterance tend to be lifted up when they are seriously interfered with additive noise, we therefore proposed a simple but effective approach, named log-energy rescaling normalization (LERN), to appropriately rescale the log-energy features of noisy speech to that of the desirable clean one.

The speech recognition experiments were conducted under various noise conditions using the European Telecommunications Standards Institute (ETSI) Aurora-2.0 database. The database contains a set of connected digit utterances spoken in English. It offers eight noise sources and seven different signal-to-noise ratios (SNRs). The experiment results showed that the performance of the proposed LERN approach was considerably better than the other conventional energy or log-energy feature normalization methods. Another set of experiments conducted on the large vocabulary continuous speech recognition (LVCSR) of Mandarin broadcast news also evidenced the effectiveness of LERN.

誌謝

時光飛逝，在師大資工所的日子，可以說是我人生中一個重大的里程碑。兩年的助理研究員再加上兩年的研究生涯，轉眼間即將道別，最不捨的就是這個逐漸壯大的資工所，這是一個讓我成長智慧與豐富知識的好地方。

這一切從數位典藏計畫開始到研究所畢業，一路上學習過多媒體系統的撰寫、PDA 手持系統的開發、網路架構、室內定位、網站建構、資料特徵擷取技術、語音辨識技術、資訊檢索技術等...由淺入深、由簡至精，最後才以「能量為基礎之語音正規化方法和語音端點偵測」成為我的研究題目。

對此論文的完成，最感謝的是我的指導老師陳柏琳教授以及林順喜教授，因為陳老師的循循善誘還有林老師的鼓勵，才得以使我完成在師大後兩年的研究生涯。每當我研究上遇見瓶頸，總是讓陳老師在一旁耐心的幫忙尋找問題所在，並給予精闢的解釋，將每一個問題的細節作最佳的解決。這才讓我們知道在做學問上紮實的重要性。而每當我遇見不如意的時候便又想起林老師對我鼓勵的一番話，才得以從拾信心。還記得是助理研究員時林老師常邀我到辦公室促膝長談，為我開導人生的未來路並且教導做學問時的各種細節，最後再給予最大的鼓勵，這是讓我最不能忘記的。

其次要感謝伴我成長的數位典藏計畫的成員們，志豪、世傑、仲良，還有在語音實驗室一起研究的所有人，士弘、士翔、炫盛、怡婷、芳輝、庭瑋、家豪、斯涵、和鴻欣，我會記得在這研究所生涯中的美好時光，因為有跟你們互相的討論才能讓我的視野更開闊，研究上也才能有較佳的結果。

生活上，我則要感謝婉如老婆大人的支持，值得高興的是從結婚到現在兒子已經周歲又三個月了，而悲傷的是母親從病危到逝世。再次謝謝老婆大人在師大資工所的這四年，幫助我順利通過人生過程中的各種不如意的時光。

最後感謝口試委員們的建議讓本論文更臻完整，並將這篇論文的成果獻給幫助過我的每一個人，沒有你的存在就沒有今天的成果。

謝謝大家 鴻彬謹誌

章節目錄

圖目錄

表目錄

第一章 序論

1.1 研究動機

科技的發展始終來自於人類的需要，期望的就是希望建立良好的生活品質與幫助人們的互動關係，然而人與人之間的「說話」或人與電腦之間的「交談」就成為溝通全人類最重要的中介與橋樑。近年來，由於電腦科學方面的技術發展快速，在硬體上許多更輕薄短小的智慧型電子設備不斷地被發展出來，而軟體上的人機互動也逐漸以語音控制或語音輸入資訊的方式取代傳統以鍵盤為主的互動方式。此外，人們對於隨時隨地取得資訊的需求日益強烈，語音即將扮演更重要的角色，擔任起人們與各種不同智慧型電子設備間最主要的人機介面。因此，自動語音辨識(Automatic Speech Recognition, ASR)技術的應用在未來勢必成為日常生活中不可或缺的一個環節。

在現實生活中，語音本來就是最簡易的互動方式，以目前的電腦語音辨識率當作參考，雖然並沒有達到百分之百正確的地步，不過使用語音來操控一些簡單安全的動作卻是沒有問題的，比如電視選台、冷氣機操作等等的機械控制，另外利用語音辨識技術更直接的應用，就是將語音這龐大的儲存方式轉存成純文字，如此一來便可以得到更多的硬體空間來節省資源的浪費，並且改變了語音資料上循序檢索的方式，在純文字的資料上，我們可以更充分地應用這份語音內容。而拜積體電路之賜，現在數以百萬計的電子元件都可以整合到一個小小的晶片上，要利用晶片來做即時小詞彙的語音辨識已經是可以做得到的。相信不久的將來，家電用品都會內建具有運算功能的晶片，所有家電也都具有上網功能，從而接受遠端語音的控制。目前語音辨識雖然已有初步的應用在通訊領域上，諸如客服中心的語音系統或是手機語音撥號功能等等，但語音辨識效率方面，若要達到快速並且完全正確的境界，仍然有一段很長的距離要走。主要影響的因素如環境的噪

音、語者間差異，以及通道效應等在真實環境才會遇到的問題，即使是目前最好的語音辨識系統，在上述的干擾下其效能依舊會大大地降低。因此對抗噪音的問題正是本論文主要探討的部份。

1.2 研究目的

自動語音辨識在真實環境中會遇到受環境噪音及通道效應影響等問題，正是語音強健性技術(Speech Robustness)長久以來一直被視為重要研究課題的原因。語音強健技術解決的辦法不外乎增強語音訊號、壓抑非語音訊號或同時進行，但本論文主要的研究目的是希望藉由訊號本身的能量大小在語音特徵參數上做適當的處理與刻度的調整，以減緩噪音干擾的影響、降低訓練環境與測試環境不匹配的情形、提升語音能量訊號及語音特徵參數本身的強健性。根據語音訊號的能量觀察，乾淨語音受到噪音干擾的結果顯示，在對數能量的表現上當受到噪音影響時將會使得對數能量產生非線性的失真，在對數能量較高的音框僅有輕微的影響；相反地，在對數能量較低的音框則會有較為嚴重的影響，整個對數能量值被提高，因此相對地壓縮語句對數能量的值域，使得語音段落(通常是對數能量值高的段落)與非語音段落(通常是對數能量值低的段落)若以對數能量來做區隔愈顯不易。本論文主要針對乾淨語音受到噪音干擾時的情況做分析與探討，進一步的參考前人所發表的方法加以改進與創新，期許達到減少噪音能量對語音能量干擾做復原的目標，使受噪音干擾的測試語料能夠與乾淨的訓練語料的能量特徵能夠相匹配，進而提高系統辨識效能。

1.3 研究內容

從語音訊號處理的文獻中，我們可以歸納環境中干擾語音訊號的噪音可概略分為二種類型：(1)加成性噪音(Additive Noise)和(2)摺積性噪音(Convolutional Noise)。

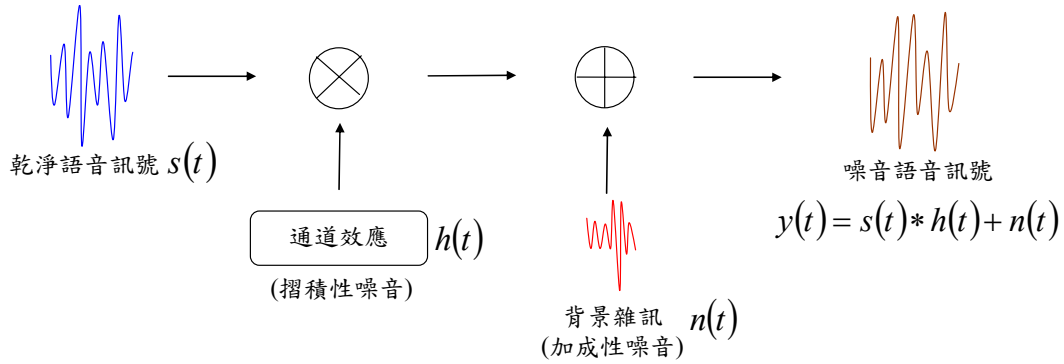

圖 1.3.1 噪音干擾示意圖

加成性噪音為收錄語音時，人員的語音與背景環境噪音以線性加成(Linearly Additive)的關係同時被錄製進去，例如周遭人走動的聲音(非穩定性噪音)或是冷氣設備所發出的噪音(穩定性噪音)等；摺積性噪音通常是指語音訊號在經由不同傳輸通道時所產生的通道效應，例如電話線路通道效應、麥克風通道效應等。加成性噪音與摺積性噪音對於語音訊號的干擾過程可以用圖 1.3.1 來表示。

強健性語音技術的主要目的就是為了消除不同環境下的差異性以及減輕噪音對語音訊號的影響，過去已有許多方法成功地被提出，依據方法的本質可概分為以下三種方向[Gong 1995]：

1. 語音強化技術(Speech Enhancement)

目的在於提升語音訊號本身的品質，通常是假設語音訊號與噪音訊號二者在統計上是不相關(Uncorrelated)，希望能由觀察到的噪音語音(Noisy Speech)重建出乾淨語音(Clean Speech)訊號。常見的技術有頻譜消去法(Spectral Subtraction, SS)[Boll 1979]、維爾濾波器(Wiener Filter, WF)[Huang 2001]等。

2. 強健性語音特徵(Robust Speech Feature)

從語音訊號中擷取出較不易受到環境變化干擾而失真的強健性語音特徵參數。常見的技術有倒頻譜平均消去法(Cepstral Mean Subtraction, CMS)[Furui 1981]、倒頻譜正規化法(Cepstral Mean and Variance Normalization, CMVN)[Viikki and Laurila 1998]等。

3. 聲學模型調適(Acoustic Model Adaptation)

藉由少量的調適語料(Adaptation Data)對由乾淨語音所訓練而成的聲學模型做調整，主要調整聲學模型中機率分布的參數，如平均值向量(Mean Vector)或共變異矩陣(Covariance Matrix)。期望調適後的模型可以適用於新的環境，用以降低環境不匹配的現象。常見的技術有最大事後機率法則(Maximum a Posteriori, MAP)[Gauian and Lee 1994]、最大相似度線性回歸法(Maximum Likelihood Linear Regression, MLLR)[Leggetter and Woodland 1995]等。

1.4 研究貢獻

本論文依據參考文獻，主要研究語音對數能量特徵的技術，並提出對數能量尺度重刻法技術。在研究方面將探討對數能量在不同環境狀況下，接受各種噪音的影響，要點在於觀察時間軸上的對數能量變化，吾人發現在一段乾淨語句中有語音出現的段落其對數能量特徵值會較高；反之若無語音出現的段落其對數能量特徵值則會接近於一穩定的低能量值。此外，當一段語句受到嚴重的噪音干擾前與噪音干擾後，語句中的能量特徵可以明顯發現原本對數能量較低的部分會提高許多，最後根據對數能量的特性，吾人提出兩種對數能量尺度重刻的作法，並討論其強健性的效果。實驗方面，環境設定主要採用 Aurora-2.0 [ETSI]，內容為英語發音的連續數字字串的小詞彙語料庫，提供有八種噪音來源和七種訊噪比(Signal-to-Noise Ratio, SNR)的測試情況，進一步討論參考文獻中的各種技術與吾人所提出的方法在語音辨識器上的辨識結果。其中對數能量尺度重刻法使用最佳設定時，當各種噪音搭配不同訊噪比(SNR)環境下的音節辨識結果，其中，詞平均正確率與基礎實驗結果做比較，我們發現對數能量尺度重刻法在乾淨語料訓練模式下的相對進步率高出 34.68%。最後吾人將對數能量尺度重刻法實作於中文大詞彙連續語音辨識系統，其使用的語料庫為 MATBN 電視新聞語料[Wang et al. 2005]，語音辨識詞典部分則包含有七萬二千個字詞，從辨識結果上證實，在大詞彙的辨識系統中仍然可以得到有效提升。

1.5 章節大綱

本論文章節概要如下：

第二章：首要介紹實驗語料庫與相關實驗環境設定。論文中我們主要採用的是在語音辨識的學問上廣被國際學者所使用的 Aurora-2.0，資料內容是由英語發音的連續數字字串。第二小節將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。

第三章：回顧參考文獻，主要討論近年在國際學術上被廣泛及討論的語音強健技術，包含音框對數能量消去法(Frame Log Energy Subtraction, FLES)、對數能量動態範圍正規化法(Log-Energy Dynamic Range Normalization, LEDRN)和靜音音框對數能量正規化法(Silence Log-Energy Normalization, SLEN)。

第四章：對數能量尺度重刻法為本研究論文所提出的音框對數能量正規化方法，主要在於對數能量的強健性技術改進，並探討不同的參數設定狀況的於實驗在各種環境下的結果。最後則討論音框對數能量正規化於倒頻譜正規化法之加成性的實驗。

第五章：利用論文所提出的尺度重刻法應用於語音能量端點偵測，章節中將簡介近年來常見的端點偵測技術，進一步則討論語音能量端點偵測經過尺度重刻法處理的前後效果。

第六章：主要討論對數能量為基礎之語音正規化於中文大詞彙連續語音辨識系統(Large Vocabulary Continuous Speech Recognition, LVCSR)的辨識效果，並比較音框對數能量正規化法在英語數字小詞彙與中文大詞彙語料庫的差異，從最後結果得知，本論文所提出的尺度重刻法在不同語料庫的辨識率都可以有效提升。

第七章：結論與未來展望

最後為參考文獻。

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><th colspan="4">AURORA 2.0</th></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：

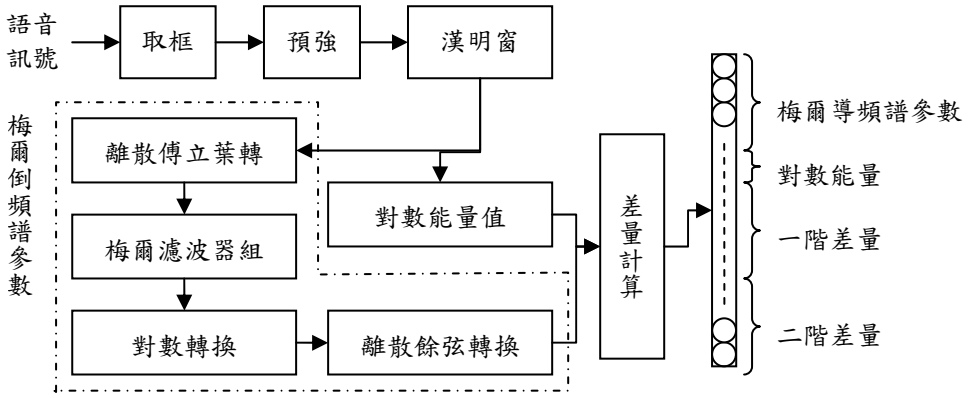

圖 2.2.1 特徵參數擷取流程圖

(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。

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

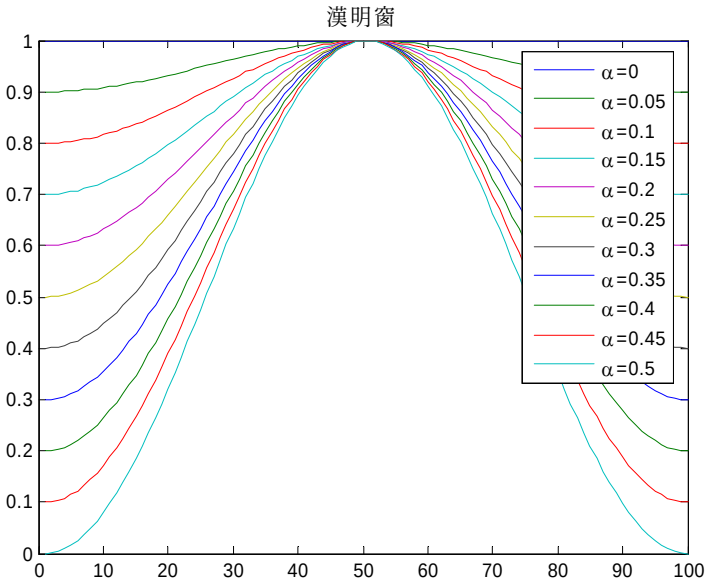

圖 2.2.2 不同 α 參數下的漢明窗示意圖

(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。

第三章 文獻回顧

3.1 對數能量特徵值之強建式技術

當語音訊號受到環境噪音干擾時，在語音資料上的現象可由對數能量參數在時間軸上的變化明顯觀察出，尤其是在非語音訊號的段落，原本的語音對數能量參數值應該偏低，但受噪音的干擾後對數能量參數值卻因此而增加。然而參考文獻中[Bocchieri and Wilpon 1992]學者也特別研究能量特徵值對於自動語音辨識的影響。本章節所回顧的對數能量特徵值的技術，目前從作法上大致可分為兩類：第一類作法是只對訓練語料的對數能量特徵值做預先處理，方法為音框能量消去(Frame Energy Subtraction, FES)和對數能量動態範圍正規化法(Log-Energy Dynamic Range Normalizaton, LEDRN)。第二類作法則是同時對訓練語料和測試語料用同方法做處理，方法為靜音音框對數能量正規化法(Silence Log-Energy Normalization, SLEN)。以下小節將詳細介紹各能量特徵值強建式技術。

3.1.1 音框能量消去法(Frame Energy Subtraction, FES)

音框能量消去法[Gomez et al. 2004]主要假設噪音語音為語音訊號與噪音訊號加成的結果，與頻譜消去法(Spectral Subtraction, SS)的假設與作法相類似，希望藉由預先估測噪音的能量值，用此噪音能量值來降低噪音對語音訊號在能量上的影響，進而減少存在訓練語料和測試語料之間環境不匹配的現象。基本上必須假設：語音訊號與噪音訊號二者在統計上是不相關(Uncorrelated)，因此噪音訊號對語音訊號的影響只是加成性(Additive)的改變。因此若希望能由觀察到的雜訊語音(Noisy Speech)重建出乾淨語音(Clean. Speech)訊號，只需要將含有噪音的語音能量值扣掉噪音能量值，特別是針對測試語料每一語句的任何音框來做處理。根據上述假設與處理方法，我們可以令 $E_y[i]$ 、 $E_x[i]$ 與 $E_n[i]$ 分別為音框 i 的噪音(Noise Speech)語音能量、乾淨語音(Clean Speech)能量以及噪音(Noise)本身的能

量，而三者間有式(3.1.1)的關係：

$$E_y[i] \cong E_x[i] + E_n[i] \tag{3.1.1}$$

然而，$E_n[i]$事先並未能知道，必需透過特定方式估測而得。我們可進一步假設每一語句的前K音框為噪音音框，以取得估計值$\hat{E}_n$來取代$E_n[i]$，如式(3.1.2)所示：

$$\hat{E}_n = \frac{1}{K}\sum_{i=1}^{K} E_y[i] \tag{3.1.2}$$

在得到$\log \hat{E}_n$估計值後，採用噪音遮蔽法以防止$E_n[i] - \hat{E}_n$出現負值來取得近似的乾淨語音訊號能量，如下(3.1.3)：

$$E_x[i] = \begin{cases} Ey[i] - \alpha \cdot \hat{E}n & , Ey[i] > \hat{E}n \quad ,0 \le \alpha \le 1 \\ \beta \cdot Ey[i] & , Ey[i] \le \hat{E}n \quad ,0 \le \beta \le 1 \end{cases} \tag{3.1.3}$$

其中α為過度估測因子，β為底限因子。

3.1.2 對數能量動態範圍正規化法(Log-Energy Dynamic Range Normalizaton, LEDRN)

對數能量動態範圍正規化法[Weizhong and Douglas 2005]的目標是針對於乾淨的語音訊號做預處理(Preprossing)，讓乾淨語音訊號的對數能量可以逼近噪音語音訊號的對數能量，使得經由對數能量動態範圍正規化法(LEDRN)處理過的訓練語料所練出的聲學模型會與測試的噪音語音訊號兩者間互相匹配。

對數能量動態範圍正規化法(LEDRN)的考慮是語音訊號能量曲線波峰(Peaks)段落，不容易受到噪音影響，而在波谷(Valley)段落則會受到噪音嚴重干擾，以至於乾淨語音與噪音干擾之語音在能量特徵上有不匹配的現象。因此可以利用一個線性處理或一個非線性的處理方法，使得能量特徵波峰值維持不變，而波谷的值相對上升以達到訓練與測試的訊號能有匹配的效果。

具體作法是先對乾淨環境訓練模式的每一則訓練語句中所有T個音框之對數能量中找出最大對數能量值$LE_\max$以及最小對數能量值$LE_\min$：

$$\begin{aligned} LE\_\max &= \underset{1\le i\le T}{Max}\, LogE[i] \\ LE\_\min &= \underset{1\le i\le T}{Min}\, LogE[i] \end{aligned} \tag{3.1.4}$$

當得到最大和最小對數能量值後，根據我們事前預期測試語音的噪音干擾大小決定一個動態能量的範圍比值$D.R.$，在此可以依照不同的噪音環境定義出測試語音檔要調整的最小音框能量值：

$$D.R. = 10 \times \frac{LE\_\max}{T\_\min} \tag{3.1.5}$$

緊接著針對每句訓練語句做偵測，當所有音框中的最小對數能量值小於最小目標對數能量值$T_\min$的時候，則對語句中每個音框的對數能量作更新，更新的方法主要分為線性(Linear)調整與非線性(Non-Linear)調整，如式(3.1.6)和式(3.1.7)：

$$\begin{aligned} &if \quad LE\_\min < T\_\min \\ &then \\ &\hat{LogE}[i] = LogE[i] + \frac{T\_\min - LE\_\min}{LE\_\max - LE\_\min} \times \left(LE\_\max - LogE[i]\right) \end{aligned} \tag{3.1.6}$$

$$\begin{aligned} &if \quad LE\_\min < T\_\min \\ &then \\ &\hat{LogE}[i] = LogE[i] + \frac{T\_\min - LE\_\min}{\log(LE\_\max) - \log(LE\_\min)} \\ &\qquad \times \left(\log(LE\_\max) - \log(LogE[i])\right) \end{aligned} \tag{3.1.7}$$

其中$\hat{LogE}[i]$為更新後之對數能量。由式(3.1.6)和(3.1.7)可以發現對於語音能量較大的音框在透過對數能量動態範圍正規化法處理後只被作些許地作修正，而能量較小的音框則被作大幅地作修正，如圖 3.1.1 為所顯示，圖中橫軸為原對數能量，縱軸為調整後的對數能量對應值，藍色實線為原對數能量排序後曲線，長虛線為非線性方式調整，點虛線為線性調整後的對數能量。

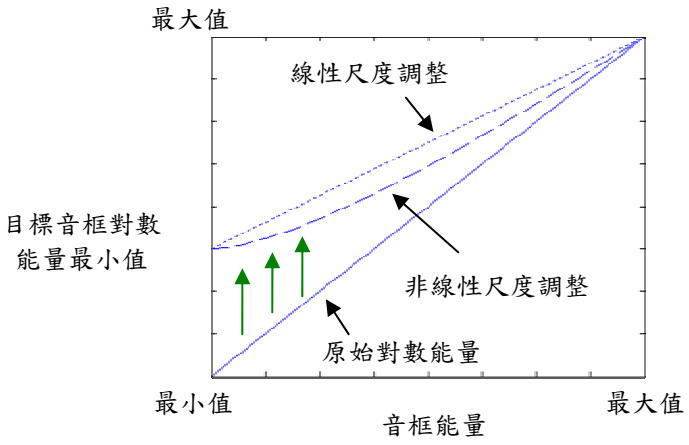

圖 3.1.1 對數能量動態範圍正規化法作用前後示意圖

3.1.3 靜音音框對數能量正規化法 I (Silence Log-Energy Normalization, SLEN I)

靜音音框對數能量正規化法 I [戴仲甫 2006]是一個快速而有效的方法，主要利用語音端點偵測(Voice Activity Detector, VAD)的方法找出非語音區間作進階處理。靜音音框對數能量正規化法的原理類似能量正規法，由於噪音語音受干擾較為嚴重的部份為波谷，而波峰則比較不受影響。實驗顯示若僅保留波峰部分，經辨識後依然可以得到不錯的正確率。從實驗觀察得知，對能量特徵而言最重要的是語音能量曲線而不是非語音能量失真的部分，也就是說當有語音的段落若具有完整的能量曲線，會比音框能量值整體的降低或昇高所造成的失真程度還重要。因此假設，具有完整的語音能量曲線，就可以得到好的辨識率。依照上述的假設，我們找出波型中非語音的部份，並且把它正規化為一個常數值，由於非語音的部份已經正規化為一個常數值，而語音部分又比較不受噪音的干擾，因此正規化處理過後的語料，不論在乾淨環境下語音段落的對數能量曲線或噪音環境下的語音對數能量的曲線會相似。

具體作法是利用噪音偵測法來找出非語音音框，以便輔助靜音音框對數能量正規化法 I。靜音音框對數能量正規化法 I 主要利用能量來判斷語音與非語音的門檻值，並將非語音音框正規化為一個常數。如下式(3.1.8)與(3.1.9)：

$$\tau = \frac{1}{T}\sum_{i=1}^{T} LogE[i] \tag{3.1.8}$$

$$Log\hat{E}[i]_{i=1\ldots T} = \begin{cases} LogE[i] & ,\text{if } LogE[i] \geq \tau \\ \Phi & ,\text{otherwise} \end{cases} \tag{3.1.9}$$

其中T為一段語音的音框數，Log$E[i]$與$Log\hat{E}[i]$分別為音框對數能量以及修正後之音框對數能量，τ則是門檻值而Φ為一個常數。

3.1.4 靜音音框對數能量正規化法 II (Silence Log-Energy Normalization, SLEN II)

靜音音框對數能量正規化法 II [Tai and Hung 2006]，同靜音音框對數能量正規化法 I 的假設，但是判斷語音與非語音的門檻值方法捨棄靜音音框對數能量正規化法 I 的作法改用變動音框位移率(Variable Frame Rate, VFR)取代。變動音框位移選擇噪音音框方法如式(3.1.10)：

$$y[i] = \frac{1}{2}\left(LogE\left[i+1\right] - y\left[i-1\right]\right) \tag{3.1.10}$$

其中$y[i]$為$LogE[i]$通過高通濾波器(High-Pass Filter)的輸出值;$LogE[i]$則為每個音框的對數能量。音框判斷法如式(3.1.12)，若高通濾波器輸出值的$y[i]$值小於門檻值τ則認定該音框為非語音音框，而τ的計算值與整個演算法如下式(3.1.11)與(3.1.12)：

$$\tau = \frac{1}{T}\sum_{i=1}^{T} \log y[i] \tag{3.1.11}$$

$$\log \hat{E}[i]_{i=1\ldots T} = \begin{cases} \log E[i] & ,\text{if } \log y[i] \geq \tau \\ \Phi & ,\text{otherwise} \end{cases} \tag{3.1.12}$$

最後如 3.1.3 節的假設，依照靜音音框對數能量正規化後，我們找出波型中非語音的部份，並且把它正規化為一個常數值，因此正規化處理過後，受噪音影響的語音能量的曲線會與乾淨的語音訊號能量波形將相似。

3.2 對數能量特徵強建技術實驗

根據 Aurora-2.0 的實驗設定如表 2.1.1，章節中的語音訊號特徵參數為梅爾倒頻譜參數。在此將介紹 3.1 節所的語音強健技術實做於對數能量特徵值上。實驗數據採分別表示乾淨環境訓練模式與複合情境訓練模式的測試結果，最後由實驗結果顯示兩種環境下的訓練語料和測試語料於音框能量上的不匹配情況，皆可以在語音強健技術處理後能夠有效降低噪音的干擾，並在詞正確率上得到明顯的提升。實驗結果比較三組測試語料(測試組 A、測試組 B 和測試組 C)，不同的噪音設定包含有乾淨(Clean)環境結果與不同訊噪比(SNR) -5dB~20dB 干擾下的結果，表格中列平均為各組別在相同下訊噪比干擾下的平均值，其次表中最後一行的 0~20dB 平均為相同噪音於訊噪比 0~20dB 干擾下的平均值結果。

3.2.1 音框對數能量消去法(FES)

音框能量消去法假設噪音語音為語音能量與噪音能量加成的結果，式(3.1.3)中α過度估測因子設定為 0.95，β底限因子為 0.05。實驗結果如表 3.2.1 與 3.2.2。

		乾淨環境訓練模式							
	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	0~20dB 平均
測試組 A	地下鐵	98.99	96.99	92.54	79.86	52.99	24.65	9.98	69.41
	人聲	99.03	97.04	91.32	76.54	49.12	19.50	5.74	66.70
	汽車	98.78	98.21	96.30	88.13	61.94	26.78	10.08	74.27
	展覽會館	99.01	96.76	92.75	80.19	59.09	31.38	12.59	72.03
	平均	98.95	97.25	93.23	81.18	55.79	25.58	9.60	70.60
測試組 B	餐廳	98.99	96.84	91.50	78.20	52.23	23.67	6.85	68.49
	街道	99.03	97.28	93.86	81.53	55.89	28.69	12.12	71.45
	機場	98.78	96.99	94.90	82.97	57.89	27.11	10.17	71.97
	火車站	99.01	97.75	94.66	85.99	59.80	26.26	9.60	72.89
	平均	98.95	97.22	93.73	82.17	56.45	26.43	9.69	71.20
測試組 C	地下鐵	99.23	91.93	81.36	61.59	35.52	15.57	9.43	57.19
	街道	98.94	94.44	86.76	69.89	46.58	25.39	15.15	64.61
	平均	99.09	93.19	84.06	65.74	41.05	20.48	12.29	60.90

表 3.2.1 音框能量消去法於乾淨環境訓練模式實驗結果

		複合情境訓練模式							
	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	0~20dB 平均
測試組A	地下鐵	98.22	93.40	89.35	82.44	70.46	45.10	18.08	76.15
	人聲	98.34	96.16	92.26	86.03	75.57	50.70	18.95	80.14
	汽車	98.45	97.49	96.81	94.15	86.79	65.05	30.78	88.06
	展覽會館	98.40	96.14	94.79	90.28	80.65	54.92	23.94	83.36
	平均	98.35	95.80	93.30	88.23	78.37	53.94	22.94	81.93
測試組B	餐廳	98.22	96.01	91.74	83.42	73.35	49.25	17.44	78.75
	街道	98.34	96.64	94.17	90.11	78.72	52.48	23.46	82.42
	機場	98.45	97.55	96.09	90.90	80.97	60.93	27.05	85.29
	火車站	98.40	96.98	94.91	91.18	79.91	59.52	28.23	84.50
	平均	98.35	96.80	94.23	88.90	78.24	55.55	24.05	82.74
測試組C	地下鐵	98.28	94.29	90.57	82.81	64.42	33.25	13.05	73.07
	街道	98.37	96.22	93.86	86.06	68.65	41.11	19.17	77.18
	平均	98.33	95.26	92.22	84.44	66.54	37.18	16.11	75.12

表 3.2.2 音框能量消去法於複合情境訓練模式實驗結果

其中表 3.2.2 能量消去法於複合情境訓練模式需同時對訓練和測試語料作處理。

3.2.2 對數能量動態範圍正規化法(LEDRN)

對數能量動態範圍正規化針對乾淨環境訓練模式的語音對數能量做預處理，讓語料在乾淨環境下的對數能量模擬出該語料在噪音環境下的語音對數能量大小，因此在小節中的實驗以乾淨環境訓練模式為主。實驗中我們測試非線性動態範圍正規化在不同動態範圍值(*D.R.*)的正確率：如表 3.2.3，實驗結果發現當動態範圍值(*D.R.*)設定為 12dB 時，針對 Aurora-2.0 的所以測試語料會有最佳的正確率，詳細各項數據如表 3.2.4：

(D.R.)	測試組 A	測試組 B	測試組 C	總平均
10dB	68.73	69.57	57.02	66.72
11dB	73.92	76.39	62.98	72.72
12dB	74.18	76.45	62.52	72.75
13dB	71.88	74.72	61.36	70.91
14dB	70.23	73.30	61.04	69.62
15dB	68.19	71.14	60.31	67.80
16dB	68.29	71.24	60.88	67.99
17dB	66.84	69.56	60.46	66.65
18dB	66.30	68.76	60.35	66.09

表 3.2.3 動態範圍值於動態範圍非線性正規化平均實驗結果

	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	0~20dB 平均
測試組 A	地下鐵	87.04	96.41	92.48	80.99	58.46	28.55	11.88	71.38
	人聲	87.15	97.31	95.50	88.54	67.59	30.65	6.08	75.92
	汽車	86.01	97.41	95.71	87.29	66.39	34.75	11.36	76.31
	展覽會館	86.15	96.48	93.24	82.57	60.69	32.52	10.71	73.10
	平均	86.59	96.90	94.23	84.85	63.28	31.62	10.01	74.18
測試組 B	餐廳	87.04	96.38	94.14	86.21	66.41	30.73	6.26	74.77
	街道	87.15	97.34	94.56	86.15	65.18	36.79	13.78	76.00
	機場	86.01	97.38	96.06	91.50	74.23	41.07	13.36	80.05
	火車站	86.15	96.30	93.00	85.90	64.95	34.62	11.29	74.95
	平均	86.59	96.85	94.44	87.44	67.69	35.80	11.17	76.45
測試組 C	地下鐵	86.67	91.40	81.58	63.13	37.52	16.58	9.43	58.04
	街道	87.18	94.44	88.78	72.43	50.24	29.05	15.15	66.99
	平均	86.93	92.92	85.18	67.78	43.88	22.82	12.29	62.52

表 3.2.4 最佳動態範圍值(12dB)於動態範圍非線性正規化實驗結果

緊接著，對數能量動態範圍正規化實驗中我們測試線性動態範圍正規化在不同動態範圍值(*D.R.*)的正確率：如表 3.2.5，實驗結果發現當動態範圍值(*D.R.*)設定為 16dB 時，針對 Aurora-2.0 的所以測試語料會有最佳的正確率，詳細各項數據如表 3.2.6：

(D.R.)	測試組 A	測試組 B	測試組 C	總平均
10dB	15.43	14.67	13.68	14.78
11dB	41.36	39.47	32.46	38.83
12dB	61.66	61.06	47.25	58.54
13dB	69.29	70.55	54.17	66.77
14dB	72.10	74.11	57.07	69.90
15dB	73.11	75.47	58.65	71.16
16dB	73.08	75.83	59.85	71.53
17dB	72.56	75.32	60.14	71.18
18dB	71.48	74.49	60.07	70.40

表 3.2.5 動態範圍值於動態範圍線性正規化平均實驗結果

	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	0~20dB 平均
測試組A	地下鐵	97.24	96.13	93.09	83.21	61.59	29.26	11.48	72.66
	人聲	97.49	96.07	93.74	86.19	64.12	32.29	9.85	74.48
	汽車	96.90	95.94	93.14	83.51	61.80	32.78	11.12	73.43
	展覽會館	96.88	96.05	92.41	80.78	58.41	31.60	11.97	71.85
	平均	97.13	96.05	93.10	83.42	61.48	31.48	11.11	73.11
測試組B	餐廳	97.24	96.10	94.14	85.97	64.78	33.34	11.64	74.87
	街道	97.49	96.49	92.35	83.98	63.39	37.00	15.93	74.64
	機場	96.90	96.06	94.33	88.55	70.59	40.20	14.41	77.95
	火車站	96.88	95.34	93.06	84.67	63.56	35.54	12.65	74.43
	平均	97.13	96.00	93.47	85.79	65.58	36.52	13.66	75.47
測試組C	地下鐵	96.07	91.93	81.55	59.78	31.01	13.72	9.73	55.60
	街道	95.98	93.11	84.79	65.87	41.96	22.82	14.48	61.71
	平均	96.03	92.52	83.17	62.83	36.49	18.27	12.11	58.65

表 3.2.6 最佳動態範圍值(16dB)於動態範圍線性正規化實驗結果

3.2.3 靜音音框對數能量正規化法 I (SLEN I)

靜音音框對數能量正規化法 I 主要利用能量來判斷語音與非語音的門檻值，並將非語音音框正規化為一個常數Φ，在此設定Φ常數為 1。結果如表 3.2.7 與 3.2.8。

		乾淨環境訓練模式							
	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	0~20dB 平均
測試組A	地下鐵	98.86	92.57	82.81	63.65	38.66	19.25	10.38	59.39
	人聲	98.61	96.77	92.90	82.01	61.25	31.47	11.76	72.88
	汽車	98.66	95.05	90.58	79.93	59.38	30.69	14.88	71.13
	展覽會館	98.95	92.93	84.20	68.44	43.47	20.46	11.69	61.90
	平均	98.77	94.33	87.62	73.51	50.69	25.47	12.18	66.32
測試組B	餐廳	98.86	96.07	91.65	81.85	60.15	31.62	12.99	72.27
	街道	98.61	95.07	87.00	72.52	49.21	26.66	13.85	66.09
	機場	98.66	96.15	92.34	85.03	64.45	35.28	15.27	74.65
	火車站	98.95	95.71	90.81	79.88	58.99	31.29	13.21	71.34
	平均	98.77	95.75	90.45	79.82	58.20	31.21	13.83	71.09
測試組C	地下鐵	98.83	84.77	68.28	47.31	26.71	12.99	7.77	48.01
	街道	98.64	89.12	79.96	58.65	34.95	20.47	10.73	56.63
	平均	98.74	86.95	74.12	52.98	30.83	16.73	9.25	52.32

表 3.2.7 靜音音框對數能量正規化法 I 於乾淨環境訓練模式實驗結果

		複合情境訓練模式							
	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	0~20dB 平均
測試組A	地下鐵	98.43	97.24	95.79	91.19	75.65	43.78	14.58	80.73
	人聲	98.52	97.70	96.86	93.68	83.25	55.20	18.56	85.34
	汽車	98.12	97.41	96.42	93.44	85.18	61.11	23.89	86.71
	展覽會館	98.52	97.25	95.74	90.93	79.33	49.71	17.40	82.59
	平均	98.40	97.40	96.20	92.31	80.85	52.45	18.61	83.84
測試組B	餐廳	98.43	97.67	96.62	93.37	81.67	54.47	21.22	84.76
	街道	98.52	97.46	95.92	92.08	79.63	52.75	20.74	83.57
	機場	98.12	97.55	96.90	93.89	84.37	62.00	28.78	86.94
	火車站	98.52	97.78	96.14	91.51	81.46	59.36	27.21	85.25
	平均	98.40	97.62	96.40	92.71	81.78	57.15	24.49	85.13
測試組C	地下鐵	98.65	95.73	92.54	84.19	61.01	25.24	10.04	71.74
	街道	98.19	96.19	94.62	87.70	70.31	39.63	17.93	77.69
	平均	98.42	95.96	93.58	85.95	65.66	32.44	13.99	74.72

表 3.2.8 靜音音框對數能量正規化法 I 於複合情境訓練模式實驗結果

3.2.4 靜音音框對數能量正規化法 II (SLEN II)

在靜音音框對數能量正規化法 II 中的設定與法 I 相同，將非語音音框正規化為一個常數Φ，並設定Φ常數為 1。實驗結果如表 3.2.9 與 3.2.10。

		乾淨環境訓練模式							
	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	0~20dB 平均
測試組A	地下鐵	98.99	94.96	89.07	71.97	44.00	19.77	10.59	63.95
	人聲	98.73	96.86	93.77	84.89	64.57	34.07	9.73	74.83
	汽車	98.90	95.35	92.10	82.82	64.84	35.73	13.81	74.17
	展覽會館	98.92	94.08	89.02	75.38	50.85	24.50	10.80	66.77
	平均	98.89	95.31	90.99	78.77	56.07	28.52	11.23	69.93
測試組B	餐廳	98.99	96.90	93.40	85.69	64.97	36.69	12.34	75.53
	街道	98.73	96.22	91.23	78.81	57.22	31.26	12.94	70.95
	機場	98.90	96.63	93.68	87.21	68.77	41.16	15.06	77.49
	火車站	98.92	96.11	92.72	83.28	64.12	35.82	14.93	74.41
	平均	98.89	96.47	92.76	83.75	63.77	36.23	13.82	74.59
測試組C	地下鐵	99.08	88.58	74.52	51.92	27.76	11.82	7.98	50.92
	街道	98.76	91.78	84.04	65.05	41.11	21.89	11.88	60.77
	平均	98.92	90.18	79.28	58.49	34.44	16.86	9.93	55.85

表 3.2.9 靜音音框對數能量正規化法 II 於乾淨環境訓練模式實驗結果

		複合情境訓練模式							
	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	0~20dB 平均
測試組A	地下鐵	97.33	96.44	94.23	88.95	75.62	47.07	16.33	80.46
	人聲	97.37	97.01	95.74	91.38	79.53	50.48	16.20	82.83
	汽車	96.99	97.05	95.82	92.99	83.95	56.58	15.96	85.28
	展覽會館	97.44	96.14	94.88	90.65	80.13	54.34	21.01	83.23
	平均	97.28	96.66	95.17	90.99	79.81	52.12	17.38	82.95
測試組B	餐廳	97.33	96.65	94.96	90.97	78.63	54.16	19.04	83.07
	街道	97.37	96.67	95.62	90.36	78.69	53.33	21.83	82.93
	機場	96.99	96.75	95.82	92.78	84.31	63.91	28.54	86.71
	火車站	97.44	96.95	95.16	90.71	80.47	59.98	25.15	84.65
	平均	97.28	96.76	95.39	91.21	80.53	57.85	23.64	84.34
測試組C	地下鐵	97.21	94.93	91.83	84.03	64.51	30.15	10.53	73.09
	街道	97.37	95.74	93.92	86.64	71.37	42.17	18.95	77.97
	平均	97.29	95.34	92.88	85.34	67.94	36.16	14.74	75.53

表 3.2.10 靜音音框對數能量正規化法 II 於複合情境訓練模式實驗結果

3.2.5 結論

根據以上實驗結果，將三組噪音環境整理成表 3.2.11 與表 3.2.12，表格中測試組 A、測試組 B 和測試組 C 所表示的正確率為各組實驗之 0dB 至 20dB 的算數平均數結果，平均值(Average)則為測試組 A 四份加上測試組 B 四份和測試組 C 兩份的平均值結果，如下式：

$$\frac{\text{測試組 A} \times 4 + \text{測試組 B} \times 4 + \text{測試組 C} \times 2}{10} \tag{3.2.1}$$

其中第二章之基礎實驗(Baseline)結果將納入表格中一起比較，最後一行為進步率(relative improvement, R.I.)算法如下式 3.2.2。實驗結果如表 3.2.11。

$$R.I.(\%) = \frac{NewScore - Baseline}{100 - Baseline} \times 100\% \tag{3.2.2}$$

乾淨環境訓練模式					
0~20dB 平均	測試組 A	測試組 B	測試組 C	平均	進步率
Baseline	58.94	58.48	59.97	58.96	non
FES	70.60	71.20	60.90	68.90	24.23
LEDRN Linear	73.11	75.47	58.65	71.16	29.73
LEDRN Non-Linear	74.18	76.45	62.52	72.75	33.61
SLEN I	66.32	71.34	52.32	65.53	16.00
SLEN II	69.93	74.59	55.85	68.98	24.41

表 3.2.11 對數能量之強建式技術比較於乾淨環境訓練模式實驗結果

複合情境訓練模式					
0~20dB 平均	測試組 A	測試組 B	測試組 C	平均	進步率
Baseline	85.22	83.99	80.67	83.82	non
FES	81.93	82.74	75.12	80.89	-18.10
SLEN I	83.94	85.13	74.72	82.57	-7.71
SLEN II	82.95	84.34	75.53	82.02	-11.11

表 3.2.12 對數能量之強建式技術比較於複合情境訓練模式實驗結果

實驗探討：

綜合對數能量特徵值之強建式技術的實驗結果，從表 3.2.11 中可以得知以上的方法與基礎實驗比較都有明顯的進步。但是從各組別來觀察，發現結果當中的測試組 C 的效果表現有好有壞，其中則以非線性的對數能量動態範圍正規化法(LEDRN)進步最多。因此結果顯示上述強建式的技術只能對測試組 A 和 B 兩種情境下的噪音干擾有提高辨識率的效果。然而，在複合情境訓練模式實驗結果卻不理想，綜觀兩種訓練模式的實驗結果，其中以非線性的對數能量動態範圍正規化法(LEDRN)的強建式技術最佳。而測試組 C 的 MIRS 通道加成性噪音干擾，明顯是所有方法都無法有效對抗。

第四章 音框對數能量正規化

針對語音能量參數受到環境噪音干擾的改變現象，首先會在 4.1 節針對音框對數能量特徵做噪音干擾前後的觀察與進一步的探討，並且根據觀察結果提出本論文的音框對數能量正規化作法，其次討論正規化之實驗結果，最後則試驗音框對數能量正規化和倒頻譜正規化法之加成性。

4.1 音框對數能量特徵

根據觀察對數能量特徵值發現自動語音辨識結果會明顯的受到噪音干擾而辨識率變差，因此我們特別觀察語音對數能量特徵在不同噪音環境下的變化情形。在時間軸上，觀察對數能量由非語音到語音段落的改變過程，發現通常在一段乾淨語句中有語音出現的段落其對數能量特徵值會較高；反之若無語音出現的段落其對數能量特徵值則會接近於一穩定的低能量值。此外，當一段語句受到嚴重的噪音干擾前與噪音干擾後，語句中的能量特徵可以明顯看出在原本能量較低的部分會提高許多。

噪音環境干擾對於語句中對數能量特徵影響的變化程度可用圖 4.1.1 來做說明(圖示(a)為 20dB 噪音干擾環境對應乾淨環境情形，而圖示(b)為 10dB 噪音干擾環境對應乾淨環境情形)，語料來源是從 Aurora-2.0 訓練語料庫中乾淨訓練模式對應於的複合情境訓練模式的部分語料，圖示(a)中黑色點是以乾淨語句的每音框(Frame)對數能量同時以橫軸與縱軸座標值所繪出的參考點；而紅色叉則是以噪音干擾的語音對數能量為縱軸座標值對應乾淨語句的橫軸座標值所繪出的參考點；圖示(b)的參考點則與圖示(a)設定相同。由圖 4.1.1 可得知，當受到噪音影響時將會使得對數能量產生非線性的失真：在對數能量較高的音框僅有輕微的影響；但是相反地，在對數能量較低的音框上則會有嚴重的影響，故噪音干擾會讓

大部分的對數能量值被提高甚多，反而整體縮小對數能量的值域範圍，使得語音段落與非語音段落在對數能量的區隔上愈顯不易。

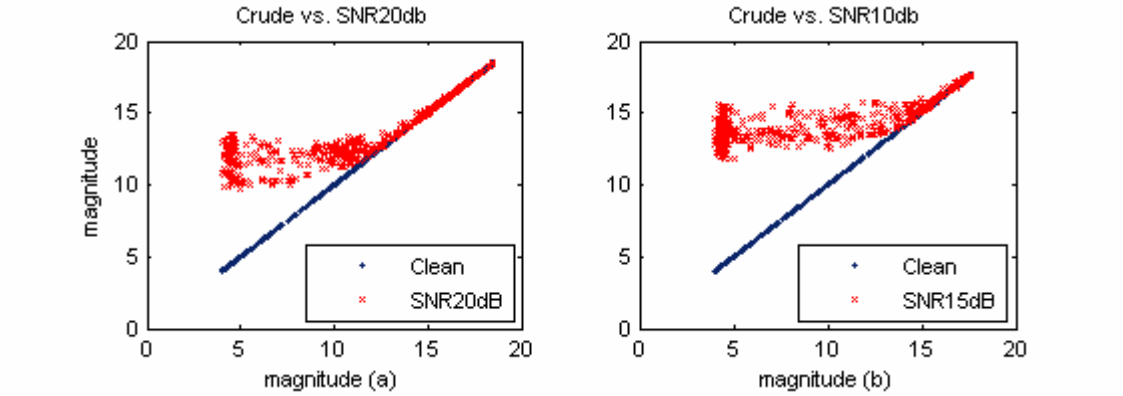

圖 4.1.1 加成性噪音影響語音對數能量特徵示意圖

在時間軸上對數能量的表現亦可用圖 4.1.2 來說明，語料來源是從Aurora-2.0 語料庫中的測試組A內挑選(語音內容為數字：139O)，環境狀況設定在乾淨環境、訊噪比(SNR)20dB噪音干擾和訊噪比(SNR)10dB噪音干擾。噪音環境對於語句對數能量特徵的影響表示：圖中橫座標表示連續的語音音框；縱軸座標代表在每一音框的對數能量；黑色實線代表原始乾淨語音的對數能量；藍色點線與紅色虛線分別代表訊噪比為 20dB與 10dB的噪音語音的對數能量。由圖亦可看出對數能量在乾淨環境下較低的音框，受噪音的影響程度較為嚴重。從圖 4.1.2 中對數能量在 5 附近的值域區間，原本是屬於非語音部分，但接受到噪音的干擾後，使得其對數能量相對提升。吾人認為上述情形是主要造成乾淨環境和噪音環境語音二者間在對數能量表現不匹配(Mismatch)的主要原因。

在這裡根據上述現象的觀察，因此我們假設若對數能量受噪音干擾的部分能重建出乾淨的語音還原成不受干擾時的情況，此時的辨識結果應該可以有效提升。所以在此特別針對對數能量維度，利用 Aurora-2.0 語料庫的語料對應特性，在相同內容的語料情形下可以利用乾淨環境的對數能量值將所有相對應語料之受噪音干擾的對數能量值作取代實驗。經過語音辨識步驟，最後可以取得一組將對數能量取代後的上限數據結果，如表 4.1.1。由結果顯示乾淨環境訓練模式平均正確率的上限可以高達 83.26%，而複合情境訓練模式可以達到 91.09%，可以

發現單獨的對數能量取代即可以有效影響辨識率結果。因此基於上述所觀察到的現象與結論，以下小節將提出對數能量尺度重刻法於強健性語音特徵處理的技術。

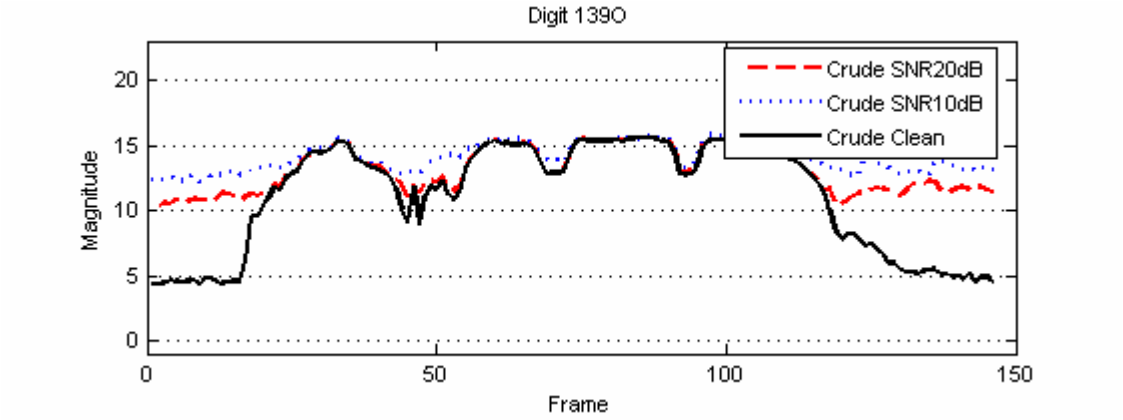

圖 4.1.2 數字語音之對數能量特徵示意圖(語音內容為：139O)

SNR 0~20dB 平均	測試組 A	測試組 B	測試組 C	總平均	基礎實驗總平均
乾淨環境訓練模式	84.05	85.8	76.58	83.26	58.96
複合情境訓練模式	91.45	93.07	86.41	91.09	83.82

表 4.1.1 無干擾之對數能量值取代干擾後的正確率上限

4.1.1 對數能量尺度重刻法 I (Log Energy Rescaling Normalization, LERN I)

在本節首先提出一個創新的語音對數能量特徵正規化技術－對數能量尺度重刻法 I，目的希望利用對數轉換函數方式來對語音對數能量參數作正規化，主要是考慮語句本身的音框對數能量特徵在不同噪音環境下的變化，試圖重建出乾淨的語音對數能量特徵。

根據上述所觀察到的現象顯示在一段乾淨語句中有語音出現的段落其對數能量特徵值會呈現較高值；反之若無語音出現的段落其對數能量特徵值則會趨近於一穩定的低能量值。另一方面，從觀察的現象可以發現受噪音干擾的語句，當

噪音能量大小沒有超過原本語音能量的情況下，該語句中受噪音干擾的音框對數能量會與相同語句在乾淨環境下的音框對數能量部分相近，因此我們認為語句中噪音能量沒有超過語音能量大小的情況下，對於大過噪音能量的音框部分在辨識系統上有相對的可靠性，而噪音所能干擾的範圍內之對數能量則容易造成錯誤的辨識結果。

因為上述的噪音干擾現象，對數能量尺度重刻的基本原理，希望將原特徵能量值乘上該特徵值所對應的等份區間之對數轉換函數值，改變特徵值尺度使處理後的特徵值能充分表現出乾淨環境下語音片段與非語音片段的對數能量差距，然而分別對每一語句劃分為均勻的相同等份，目的在於使語句和語句間互不相作用影響，僅考慮該語句中所有音框的等份數，最後利用各音框等份位置的對數轉換函數值做正規化。在這裡我們設定正規化對數函數輸出值介於 0 到 1 之間，使轉換後音框對數能量與原本對數能量的差異量自小到大有遞減的現象，因此在經過對數能量尺度重刻處理過後的音框對數能量，將可以得到原來對數能量值較低的語音段落其對數能量值越低，以及對數能量值較高的語音段落其對數能量值可以儘量維持不變，讓噪音語句在經過正規化後的對數能量可以趨近出乾淨環境下的對數能量特徵。

對數能量尺度重刻具體作法如下面步驟。首先，自每一語句(包含測試及訓練語句)的所有音框中找出最大對數能量值 $LE\_\max$ 以及最小對數能量值 $LE\_\min$ ：

$$\begin{aligned} LE\_max &= \underset{1\le i\le T}{Max}\, LogE[i], \\ LE\_min &= \underset{1\le i\le T}{Min}\, LogE[i] \end{aligned} \tag{4.1.1}$$

式中 T 為每一語句音框數，其次根據 $LE\_\max$ 及 $LE\_\min$ 決定對數能量值域範圍，並將此一範圍劃分成 M 個等份，因此每個等份區間的寬度為 L 可表示如下：

$$L = \frac{LE\_max - LE\_min}{M} \tag{4.1.2}$$

在本論文我們初步將等份個數 M 設為 100，而等份 m 所對應的對數轉換函數值如圖 4.1.3 和式(4.1.3)：

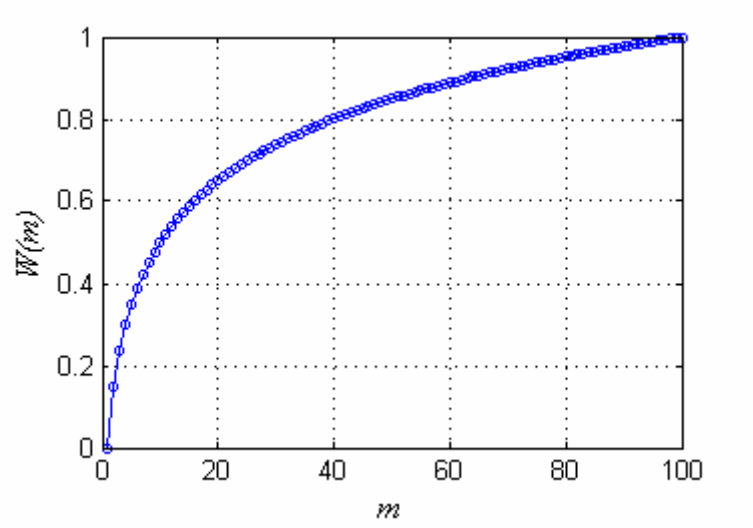

圖 4.1.3 對數轉換函數 $W(m)$ 圖示

$$W(m)=\frac{\log(m)}{\log(M)} \tag{4.1.3}$$

另一方面，每個音框對數能量所落至該等份的索引值可表示成：

$$Index_i=\left\lfloor \frac{LogE[i]-LE\_min}{L} \right\rfloor \tag{4.1.4}$$

因此音框 i 經正規化後的對數能量 $Log\hat{E}[i]$ 可以表示成：

$$Log\hat{E}[i]=LogE[i]\times W(Index_i) \tag{4.1.5}$$

從圖 4.1.2 看出，未經處理的噪音語音對數能量值在不同訊噪比狀況下(Clean, 15dB, 5dB)其音框能量值的曲線相對提昇許多，尤其以 5dB 噪音干擾的情況下，語句的對數能量的值域被嚴重壓縮。而在經過對數能量尺度重刻處理後，不同訊噪比情況下，於非語音段落的對數能量值由 10 大小降低至 5 以下，我們可以容易地由圖 4.1.4 的非語音區間察覺我們所提出對數能量尺度重刻方法能將噪音環境的對數能量曲線靠近乾淨語音對數能量曲線，但是轉換函數會讓轉換後的對數

能量過小造成對數能量曲線不連續的現象，如圖 4.1.4 橢圓區域範圍中的對數能量曲線，會有突然的波谷(Valley)出現。此外我們再一次將經過對數能量尺度重刻法處理後的音框對數能量利用圖 4.1.1 的方式畫出表示圖 4.1.5，由圖中可以看出，處理後的對數能量點雖然在低能量區間會呈現散亂的情況，並且因為對數轉換函數的關係，對數能量點在噪音環境下會比在乾淨環境降低的更多，如圖中橢圓區塊的值，但實際上大部分的能量點會接近於乾淨語句的每音框(Frame)對數能量參考點。另一方面，圖 4.1.4 與圖 4.1.5 的乾淨環境下之對數能量參考點是有經過對數能量尺度重刻法處理的結果，主要原因參考節後的實驗數據比較中得知有較佳的辨識效果，故圖中的對數能量參考點皆有採用對數能量尺度重刻法處理。

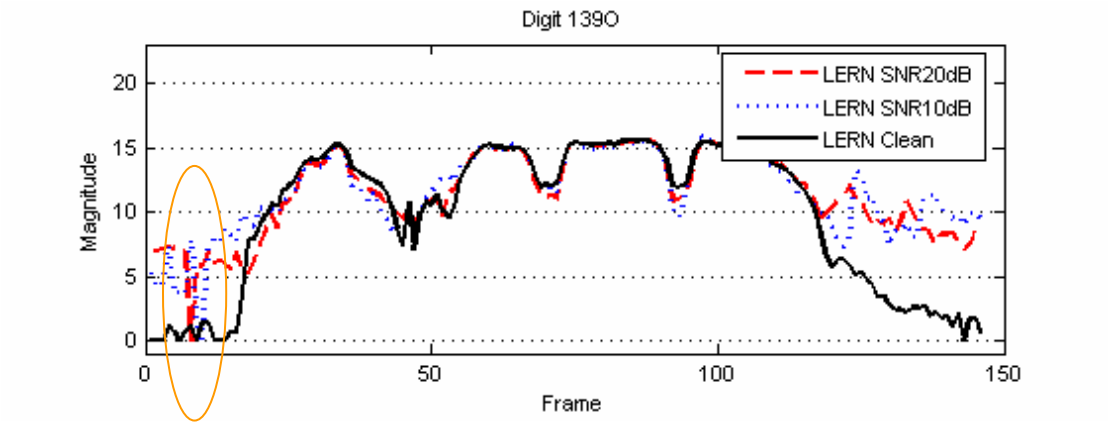

圖 4.1.4 對數能量尺度重刻法於語音對數能量特徵之作用結果圖示(1)

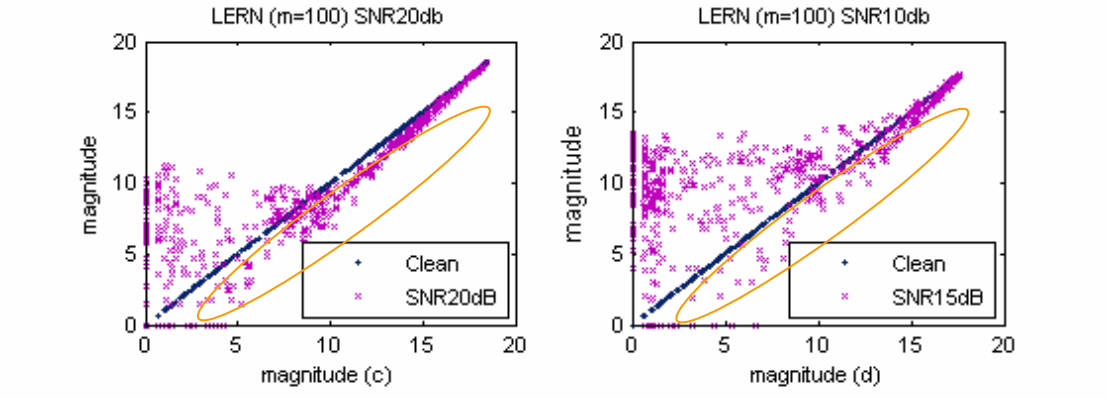

圖 4.1.5 對數能量尺度重刻法於語音對數能量特徵之作用結果圖示(2)

4.1.2 對數能量尺度重刻法 II (LERN II)

對數能量尺度重刻法 II，同方法 I 根據上述所觀察到的現象顯示在一段乾淨語句中有語音出現的段落其對數能量特徵值會呈現較高值；反之若無語音出現的段落其對數能量特徵值則會趨近於零。因此對數能量尺度重刻 II 的目標希望將原特徵能量值乘上該特徵值所對應的權重值。在此我們定義權重值函數如下式：

$$W(i)=\left(\frac{\log E[i]-\alpha\cdot LE\_\min}{LE\_\max-\alpha\cdot LE\_\min}\right)^{\beta} \tag{4.1.6}$$

式中 $LE\_\max$ 和 $LE\_\min$ 為最大對數能量值以及最小對數能量值，是由每一獨立測試語句中的所有音框 T 找出，其中 α 與 β 為控制權重曲線差異的參數，α 表示調整其值域範圍大小，而 β 指數使其成一非線性函數，權重曲線可用下圖 4.1.6 和圖 4.1.7 表示，範例圖 4.1.6 控制 α 為變數 β 為常數 1，另一圖 4.1.7 控制 β 為變數 α 為常數 1，圖中我們設定橫軸的對數能量值 $\log E[i]$ 由 0 值開始到最大值 100，縱軸則為權重值 $W(i)$ 的值域從 0 到 1。

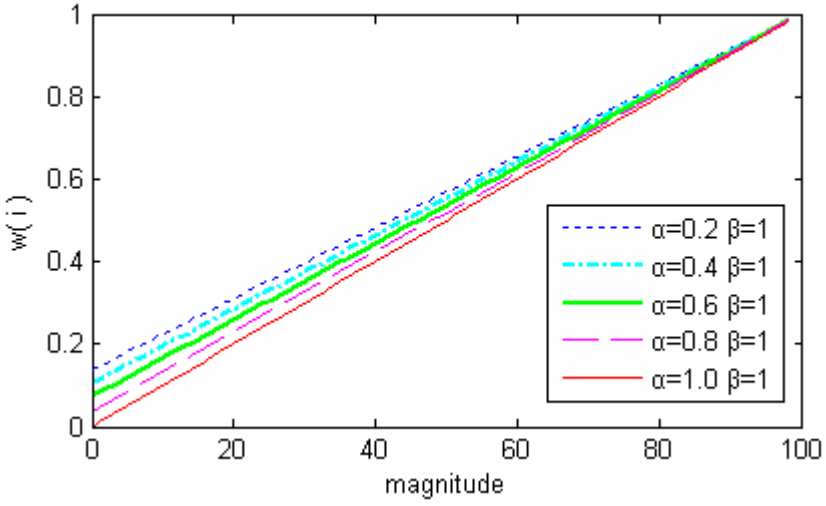

圖 4.1.6 在 β 值固定的情況下(β=1)，不同的 α 值對 $W(i)$ 的影響

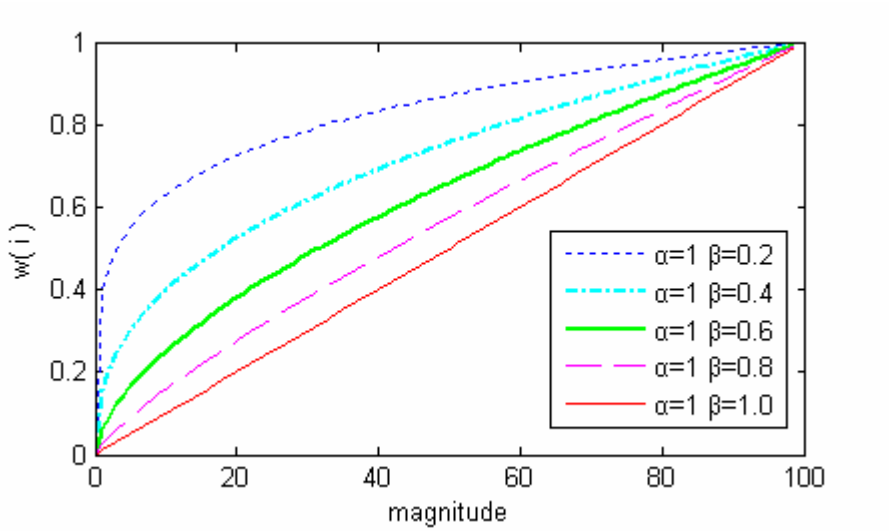

圖 4.1.7 在α值固定的情況下(α=1)，不同的β值對$W(i)$的影響

由圖 4.1.6 可以了解當α為變數時允許對能量值範圍做出微調的改變，而圖 4.1.7 則當β為變數時可以得到一非線性曲線，從曲線清楚看出在相對於對數能量較低的音框時可以有較低的權重值。因此當我們對原始對數能量乘上該音框的權重值後，最終能夠達到如觀察乾淨語句時有語音段落其對數能量特徵值較高；非語音段落其對數能量特徵值趨於零的目的，表示如下式：

$$\log \hat{E}[i] = \log E[i] \times \left(\frac{\log E[i] - \alpha \cdot LE\_\min}{LE\_\max - \alpha \cdot LE\_\min} \right)^{\beta} \tag{4.1.7}$$

至於權重函數中的α與β雙變數，在此為求取最佳值，我們進一步的利用線性迴歸的曲線擬合方法來取得α與β適當解。「曲線擬合」的概念是當給定一些資料點數(u_i, v_i)，若要以一個函數描述反應變數v_i與解釋變數u_i關係，通常可使用迴歸模型(Regression Models)來表示。換句話說，迴歸模型可用來解釋給定u_i的情況下，預測v_i的值為何。通常迴歸公式$G(u_i)$可依參數(Coefficients)組合不同表示成線性(linear)或非線性(nonlinear)型式，並且$G(u_i)$參數的選擇影響預測值$\tilde{v}_i$的準確性甚鉅，一般利用誤差平方和最小化(Minimization of the Sum of Squares Error)求得，亦即將所有u_i分別代入迴歸公式所求得的預測值$\tilde{v}_i$和實際觀測值v_i的誤差值平方合必須最小，其意謂著經由迴歸模型所預測出的值會跟實際的值較相似，

此法又可稱最小平方迴歸法(Least Squares Regression)。因此，對權重函數於誤差平方和 E^2 可以定義成式(4.1.8)：

$$E(\alpha,\beta)=\sum_{i=1}^{N}\left(v_i-\left(\log E[i]\times\left(\frac{\log E[i]-\alpha\cdot LE\_\min}{LE\_\max-\alpha\cdot LE\_\min}\right)^{\beta}\right)\right)^2 \tag{4.1.8}$$

但因為函數是非線性型式無法對參數 α、β 的導式為零求解，所以最後可利用梯度下降法(Gradient Descent)來求得適當解。

4.2 音框對數能量尺度重刻法實驗結果

本節討論音框對數能量正規化之強健技術。實驗結果將比較三組不同噪音設定下之乾淨環境結果與不同訊噪比-5dB~20dB 干擾下的結果，表格中之三組噪音環境(測試組 A、測試組 B 與測試組 C)所表示的正確率為 0dB 至 20dB 的算數平均數結果，最後總平均值(Average)計算方式為測試組 A 四份加上測試組 B 四份和測試組 C 兩份的平均值，如下：

$$\frac{\text{測試組 A}\times 4+\text{測試組 B}\times 4+\text{測試組 C}\times 2}{10} \tag{4.2.1}$$

4.2.1 對數能量尺度重刻法 I 實驗

對數能量尺度重刻法 I，在此將探討三組實驗數據。第一組實驗：在對數能量尺度重刻法上我們使用不同刻度作測試，主要針對對數表的 M 個等份各別設定為 50、70 到 500 與 1000 多種尺度做觀察。最後產生的結果如表 4.2.1。

從實驗一表格中得知，以 M=100 的尺度設定在乾淨語料訓練模式下會有最佳的正確率，而複合情境訓練模式下則是以 500 等份與 1000 等份的尺度設定會有好的效果。因此在第二組實驗：我們針對 100 的尺度設定，比較結果如表 4.2.2 與表 4.2.3。

	乾淨環境訓練模式				複合情境訓練模式			
M 等份	測試組 A	測試組 B	測試組 C	總平均	測試組 A	測試組 B	測試組 C	總平均
基礎實驗	58.94	58.48	59.97	58.96	85.22	83.99	80.67	83.82
M = 50	74.10	76.71	63.07	72.94	86.33	86.25	81.04	85.24
M = 60	74.16	76.71	63.30	73.01	86.32	86.24	80.97	85.22
M = 70	74.32	76.79	63.46	73.13	86.34	86.28	81.05	85.26
M = 80	74.35	76.76	63.62	73.17	86.36	86.31	81.14	85.29
M = 90	74.35	76.70	63.67	73.15	86.33	86.25	81.22	85.27
M = 100	74.36	76.72	63.83	73.20	86.31	86.27	81.22	85.28
M = 110	74.34	76.65	63.83	73.17	86.37	86.28	81.25	85.31
M = 120	74.28	76.60	63.85	73.12	86.38	86.25	81.30	85.31
M = 130	74.28	76.56	63.90	73.11	86.38	86.20	81.31	85.29
M = 140	74.23	76.47	63.90	73.06	86.37	86.16	81.37	85.28
M = 150	74.23	76.43	63.90	73.05	86.39	86.19	81.38	85.31
M = 500	73.47	75.21	63.33	72.14	86.75	85.85	81.62	85.36
M = 1000	73.24	74.90	63.74	72.00	86.67	85.89	81.68	85.36

表 4.2.1 對數能量尺度重刻法 I 於不同尺度等份的實驗結果

		乾淨環境訓練模式							
	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	0~20dB 平均
測試組 A	地下鐵	99.08	97.45	94.44	84.03	58.27	25.88	10.25	72.01
	人聲	98.97	98.19	95.86	85.64	60.70	25.27	2.96	73.13
	汽車	98.93	98.00	96.21	88.49	69.25	39.61	14.67	78.31
	展覽會館	99.20	97.01	93.67	83.12	61.71	34.31	15.80	73.96
	平均	99.05	97.66	95.05	85.32	62.48	31.27	10.92	74.36
測試組 B	餐廳	99.08	98.16	95.36	86.06	62.97	29.66	7.61	74.44
	街道	98.97	97.70	94.77	85.67	65.72	38.42	15.54	76.46
	機場	98.93	98.27	96.33	89.74	70.12	38.38	12.11	78.57
	火車站	99.20	98.12	95.53	88.18	67.02	38.20	14.63	77.41
	平均	99.05	98.06	95.50	87.41	66.46	36.17	12.47	76.72
測試組 C	地下鐵	99.39	93.61	85.51	68.19	41.23	17.56	10.01	61.22
	街道	99.09	94.80	87.82	72.34	50.33	26.90	15.81	66.44
	平均	99.24	94.21	86.67	70.27	45.78	22.23	12.91	63.83

表 4.2.2 對數能量尺度重刻法 I 於乾淨環境訓練模式(100 等份)實驗結果

		複合情境訓練模式							
	訊噪比	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	0~20dB 平均
測試組 A	地下鐵	98.43	96.90	96.07	91.83	82.74	59.87	23.73	85.48
	人聲	98.31	97.31	96.04	93.11	83.77	57.83	22.04	85.61
	汽車	98.18	97.26	96.72	94.33	86.43	59.47	21.09	86.84
	展覽會館	98.46	97.38	96.73	93.30	86.45	62.70	23.11	87.31
	平均	98.35	97.21	96.39	93.14	84.85	59.97	22.49	86.31
測試組 B	餐廳	98.43	96.96	95.12	89.56	80.84	57.08	21.28	83.91
	街道	98.31	97.40	96.13	93.11	84.07	59.95	24.94	86.13
	機場	98.18	97.73	96.63	94.12	87.09	67.91	31.82	88.70
	火車站	98.46	97.35	95.77	93.24	83.77	61.52	24.81	86.33
	平均	98.35	97.36	95.91	92.51	83.94	61.62	25.71	86.27
測試組 C	地下鐵	98.65	96.78	95.21	91.13	76.91	39.64	12.77	79.93
	街道	98.16	96.80	95.89	91.72	79.56	48.52	21.16	82.50
	平均	98.41	96.79	95.55	91.43	78.24	44.08	16.97	81.22

表 4.2.3 對數能量尺度重刻法 I 於複合情境訓練模式(100 等份)實驗結果

表 4.2.2 與表 4.2.3 中 0-20dB 平均為訊噪比 0~20dB 干擾下的平均值結果。綜合乾淨環境訓練模式與複合情境訓練模式的實驗結果於各組別之正確率比較，我們得知在三類測試組的正確率都有提升的效果，但對於測試組 C 中 MIRS 通道加成性噪音效果的提升卻比較小。其次，對於測試組 B 的非穩定性(Nonstationary)噪音則有最佳的提升效果。

實驗三我們觀察對數能量尺度重刻法I若僅使用在測試語料狀況下與同時使用在訓練語料和測試語料的不同，實驗結果如表 4.2.4(使用 M =100 的等份設定)。結果中發現，若同時針對在訓練語料和測試語料的情況下使用對數能量尺度重刻法，會有較高的正確率。在此我們分析表 4.2.4 結果並參考圖 4.2.1，圖(a)與表示只對測試語料處理，圖(b)則是同時對訓練語料和測試語料處理，對於實驗數據結果認為是因為訓練語料在低能量部分並沒有靠近靜音(能量值近似於零)的情況，而是在低能量部分仍然會有小能量的噪音值產生。所以實驗設定為訓練語料和測試語料同時經過我們的方法處理後會使對數能量的值域範圍較為相似，因此辨識率可以相對提高許多。

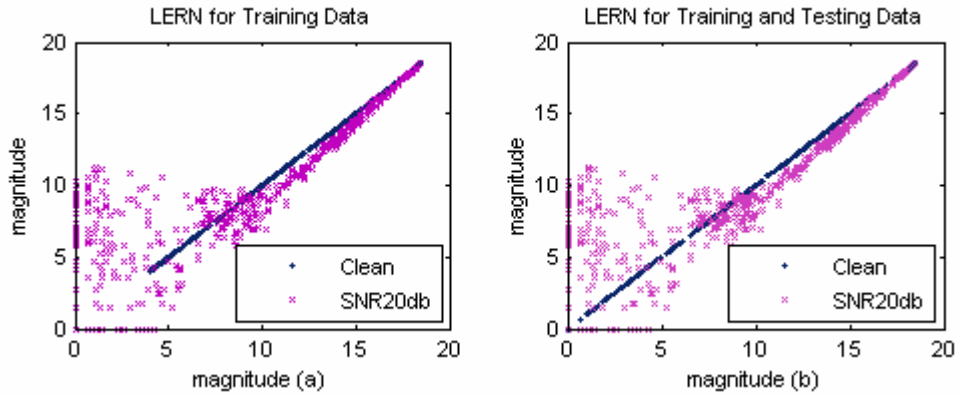

圖 4.2.1 對數能量尺度重刻法 I 於訓練語料和測試語料的差異圖示

(100 等份)	LERN 處理對象	測試組 A	測試組 B	測試組 C	總平均
乾淨環境	訓練語料＋測試語料	74.36	76.72	63.83	73.20
訓練模式	測試語料	72.20	75.45	60.58	71.18
複合情境	訓練語料＋測試語料	86.31	86.27	81.22	85.28
訓練模式	測試語料	77.76	81.98	69.86	77.87

表 4.2.4 對數能量尺度重刻法 I 於訓練語料和測試語料的差異結果

4.2.2 對數能量尺度重刻法 II 實驗

對數能量尺度重刻法 II，在此分為人工設定(Hand Set)參數與線性迴歸的曲線擬合法訓練參數的兩類實驗。實驗一：對於方法 II 我們以人工設定之α或β值為控制權重曲線差異的參數，使用不同的參數值只針對乾淨環境訓練模式的測試語料做處理，結果如表 4.2.5(實驗 1)。數據中以α為 1.0 和β為 0.2 之參數設定的組合最佳，平均正確率可達 72.02%，但是可以明顯發現在不同參數設定下的測試組 C，其結果大多呈現正確率下降的效果，並且在此發現當α參數為 1.0 和β為 1.0 的時候，對數能量尺度重刻法 II 為一線性調整，如式 4.2.1，由結果看出其效果極差。

$$Log\hat{E}[i] = LogE[i] \times \frac{LogE[i] - LE\_\min}{LE\_\max - LE\_\min} \tag{4.2.1}$$

其次，我們仍然使用人工設定之α或β值為控制權重曲線差異的參數，但同時針

對乾淨環境訓練模式和複合情境訓練模式的訓練語料與測試語料作處理，結果如表 4.2.6(實驗 2)。在乾淨環境訓練模式下的數據以α為 1.0 和β為 0.4 之參數設定的組合最佳，平均正確率可達 73%，而複合情境訓練模式下的結果則平均可以有 75%左右。最後比較表 4.2.5 與表 4.2.6，我們得知對數能量尺度重刻法 II 在使用人工設定參數值的情況下，當處理對象有同時對訓練語料和測試語料的時候會有比較好的結果。

乾淨環境訓練模式				
參數值	測試組 A	測試組 B	測試組 C	總平均
α: 0.2 β: 1.0	70.81	72.86	58.38	69.15
α: 0.4 β: 1.0	70.89	73.13	57.10	69.03
α: 0.6 β: 1.0	69.63	71.86	54.26	67.45
α: 0.8 β: 1.0	55.89	65.66	47.73	58.16
α: 1.0 β: 0.2	73.53	75.85	61.35	72.02
α: 1.0 β: 0.4	67.29	71.52	55.37	66.60
α: 1.0 β: 0.6	57.16	62.89	46.19	57.26
α: 1.0 β: 0.8	47.88	54.27	37.65	48.39
α: 1.0 β: 1.0	40.49	46.95	30.96	41.17

表 4.2.5 對數能量尺度重刻法 II 於人工設定參數值之測試(實驗 1)

	乾淨環境訓練模式				複合情境訓練模式			
參數值	測試組 A	測試組 B	測試組 C	總平均	測試組 A	測試組 B	測試組 C	總平均
α: 0.2 β: 1.0	57.66	59.49	57.29	58.32	85.66	84.60	80.47	84.20
α: 0.4 β: 1.0	60.34	62.52	57.71	60.69	85.80	84.64	80.46	84.27
α: 0.6 β: 1.0	64.33	66.96	58.40	64.20	85.94	85.15	80.89	84.61
α: 0.8 β: 1.0	69.59	72.99	58.89	68.81	87.00	87.35	80.91	85.92
α: 1.0 β: 0.2	73.25	75.31	62.83	71.99	86.71	86.05	81.67	85.44
α: 1.0 β: 0.4	74.25	77.25	61.98	73.00	86.51	85.80	80.95	85.11
α: 1.0 β: 0.6	73.30	77.16	60.39	72.26	87.17	87.10	80.32	85.77
α: 1.0 β: 0.8	72.40	76.63	59.44	71.50	86.46	87.69	80.68	85.80
α: 1.0 β: 1.0	71.71	76.45	58.76	71.02	86.26	87.13	80.41	85.44

表 4.2.6 對數能量尺度重刻法 II 於人工設定參數值之測試(實驗 2)

實驗二：我們利用線性迴歸的曲線擬合方法來取得α與β適當解。實驗上主要根據 Aurora-2.0 實驗語料庫中的兩組訓練語料，分別是乾淨環境(Clean)下之訓練語料和複合情境(Multi)下之訓練語料，由於此兩組訓練語料在語料內容上存在相對應的關係，因此吾人將此對應的關係利用曲線擬合法求得α與β解。此外在情境下之訓練語料共分為 5dB、10dB、15dB、20dB 與混合(Multi)四種不同訊噪比程度的噪音干擾情況，所以α與β參數可以分別求出在五組不同噪音干擾下的適當解與一組混和所有噪音干擾情況的適當解，其次如上所述的α與β解必須同時使用到兩組訓練語料，故實驗結果中表 4.2.7 於乾淨環境訓練模式，此時只針對測試語料做調整，而表 4.2.8 複合情境訓練模式則將α與β參數同時對訓練語料和測試語料做處理。另一方面如表 4.2.9 與表 4.2.10 我們比較在不同情況下的α與β參數對於不同噪音干擾程度的正確率效果，我們期望當參數取得的噪音干擾程度與測試環境相同的情形下可以有較好的辨識率，但很可惜的在此相同情況下的辨識率並不沒有因為噪音干擾程度一樣而特別提升，然而我們發現當 05dB 平均值的正確率在乾淨環境訓練模式下明顯變差，而複合情境訓練模式下的正確率則差異不大。最後結果顯示乾淨環境訓練模式如表 4.2.7，當訓練環境設定為 20dB 情境下所求得的α與β參數效果最佳，而複合情境訓練模式為 5dB 情境下所求得的參數效果最佳。

乾淨環境訓練模式					
訊噪比	參數值	測試組 A	測試組 B	測試組 C	總平均
05 dB	α: 0.90 β: 0.71	64.23	67.69	50.27	62.82
10 dB	α: 0.95 β: 0.43	70.11	73.17	56.23	68.56
15 dB	α: 0.98 β: 0.32	72.19	75.22	58.70	70.70
20 dB	α: 0.98 β: 0.25	73.26	75.85	60.11	71.66
Multi	α: 0.98 β: 0.34	71.24	74.36	57.91	69.82

表 4.2.7 對數能量尺度重刻法 II 於曲線擬合法之參數實驗(乾淨環境訓練模式)

複合情境訓練模式					
訊噪比	參數值	測試組 A	測試組 B	測試組 C	總平均
05 dB	α: 0.90 β: 0.71	86.61	86.38	80.98	85.39
10 dB	α: 0.95 β: 0.43	86.03	85.49	80.93	84.79
15 dB	α: 0.98 β: 0.32	86.42	85.71	81.18	85.09
20 dB	α: 0.98 β: 0.25	86.47	85.72	81.15	85.11
Multi	α: 0.98 β: 0.34	86.26	85.61	81.16	84.98

表 4.2.8 對數能量尺度重刻法 II 於曲線擬合法之參數實驗(複合情境訓練模式)

乾淨環境訓練模式								
訊噪比	參數值	Clean	20dB	15dB	10dB	05dB	00dB	0~20dB 平均
05 dB	α: 0.90 β: 0.71	98.42	89.98	81.82	68.07	47.70	26.55	62.82
10 dB	α: 0.95 β: 0.43	98.86	94.57	88.57	76.07	54.47	29.12	68.56
15 dB	α: 0.98 β: 0.32	98.91	95.96	90.89	79.15	57.34	30.17	70.70
20 dB	α: 0.98 β: 0.25	98.98	96.54	92.05	80.95	58.67	30.11	71.66
Multi	α: 0.98 β: 0.34	98.90	95.56	90.02	77.98	55.91	29.63	69.82

表 4.2.9 對數能量尺度重刻法 II 於不同訊噪比干擾下實驗(乾淨環境訓練模式)

複合情境訓練模式								
訊噪比	參數值	Clean	20dB	15dB	10dB	05dB	00dB	0~20dB 平均
05 dB	α: 0.90 β: 0.71	98.41	97.68	96.64	93.52	83.60	55.52	85.39
10 dB	α: 0.95 β: 0.43	98.43	97.30	96.00	92.35	82.33	56.00	84.79
15 dB	α: 0.98 β: 0.32	98.46	97.37	96.02	92.47	82.65	56.93	85.09
20 dB	α: 0.98 β: 0.25	98.42	97.33	96.01	92.52	82.62	57.05	85.11
Multi	α: 0.98 β: 0.34	98.40	97.22	95.90	92.36	82.65	56.78	84.98

表 4.2.10 對數能量尺度重刻法 II 於不同訊噪比干擾下實驗(複合情境訓練模式)

綜合比較：

本小節將綜合比較文獻參考的方法與吾人所提出的對數能量尺度重刻法，並且比較多項式擬合統計圖等化法 (Polynomial-Fit Histogram Equalization, PHEQ) [Lin et al. 2006]於能量維度的效果，多項式擬合統計圖等化法主要精神可以視為是利用一個轉換函數(Transformation Function)，此函數能將測試語句的語音特徵向量

每一維的統計分佈分別轉換至先前已從訓練語句中定義好的對應參考分佈。綜合實驗結果如下表，表 4.2.11 為乾淨環境訓練模式，表中得知所有的方法與基礎實驗比較都有明顯的進步。但各組分別觀察，發現結果當中的測試組 C 的效果表現有好有壞，且好的進步效果有限。總平均結果則是以對數能量尺度重刻法Ⅰ的進步率為最高到 34.7%，其他方法則在 30%左右或更低一些。在表 4.2.12 則為複合情境訓練模式，結果顯示對數能量尺度重刻法Ⅰ仍然有比較好的進步率，可以高達 9.0%。最後綜觀兩種環境的訓練模式來看，實驗中測試組 C 的 MIRS 通道加成性噪音干擾，在結果表現上明顯是所有方法都無法有效提升辨識率。

乾淨環境訓練模式					
0~20dB 平均	測試組 A	測試組 B	測試組 C	總平均	進步率
Baseline	58.94	58.48	59.97	58.96	
FES	70.60	71.20	60.90	68.90	24.23
LEDRN Linear	73.11	75.47	58.65	71.16	29.73
LEDRN Non-Linear	74.18	76.45	62.52	72.75	33.61
SLENⅡ	69.93	74.59	55.85	68.98	24.41
PHEQ on Energy	72.45	76.03	61.50	71.69	31.03
LERNⅠ(M = 100)	74.36	76.72	63.83	73.20	34.70
LERN Ⅱ (α: 0.98 β: 0.25)	73.26	75.85	60.11	71.66	30.95

表 4.2.11 乾淨環境訓練模式下綜合實驗結果

複合情境訓練模式					
0~20dB 平均	測試組 A	測試組 B	測試組 C	總平均	進步率
Baseline	85.22	83.99	80.67	83.82	
FES	81.93	82.74	75.12	80.89	-18.10
SLENⅡ	82.95	84.34	75.53	82.02	-11.11
PHEQ on Energy	86.65	86.03	79.27	84.92	6.82
LERNⅠ(M = 100)	86.31	86.27	81.22	85.28	9.00
LERN Ⅱ (α: 0.98 β: 0.25)	86.47	85.72	81.15	85.11	7.95

表 4.2.12 複合情境訓練模式下綜合實驗結果

4.3 音框對數能量正規化與倒頻譜正規化法之加成性

4.3.1 倒頻譜正規化法(Cepstral Mean and Variance Normalization, CMVN)

倒頻譜正規化法[Viikki and Laurila 1998]主要是減去倒頻譜特徵參數的平均值並針對特徵向量的標準差做正規化。假設一句話經特徵擷取後為一連串倒頻譜特徵向量 $C=\{C_1,C_2,\cdots C_t,\cdots,C_T\}$,$t=1,\cdots,T$，C_t 代表這語句的第 t 個特徵向量，T 為這語句的總特徵向量個數。最後在倒頻譜上經過倒頻譜正規化法處理後，可以適度的減少由不同的通道所造成不匹配影響。則倒頻譜正規化法求得的新特徵向量為 $\hat{C}_t$ ，如式(4.3.1)：

$$\hat{C}_t=\frac{C_t[n]-\mu[n]}{S[n]} \quad ,t=1,\cdots,T \tag{4.3.1}$$

其中

$$\mu[n]=\frac{1}{T}\sum_{t=1}^{T}C_t[n] \tag{4.3.2}$$

$$S[n]=\sqrt{\sum_{t=1}^{T}(C_t[n]-\mu[n])^2/T} \tag{4.3.3}$$

倒頻譜正規化法除了減少通道效應所造成的干擾外，同時也正規化語音特徵的機率分布，使各維度的語音特徵機率分布能夠標準化。也因為倒頻譜正規化法在語音辨識技術中效果佳並已被廣泛使用，所以特別將音框對數能量正規化於倒頻譜正規化法實驗其加成性的結果。

4.3.2 實驗結果

實驗結果如表 4.3.1 與表 4.3.2，表中我們比較基礎實驗(Baseline)結果、對數能量尺度重刻法 I (LERNI)結果、倒頻譜正規化法(CMVN) 結果和對數能量尺度重刻法 I 加上倒頻譜正規化法(LERNI+CMVN)的結果，從總平均正確率來看，在乾淨環境訓練模式下我們得知當倒頻譜正規化法加上對數能量尺度重刻法 I 的結果會有最佳的正確率，並且從各組別中可以發現對數能量尺度重刻法 I 對於不同的噪音環境的干擾加上倒頻譜正規化法都可以有正確率加成的作用，並沒有因為不同的噪音而下降辨識率。其次比較在複合情境訓練模式下的總平均正確率，隨然沒有特別提高辨識率，但正確率同樣高達 90.01%，顯示在複合情境訓練模式下對數能量尺度重刻法 I 亦不會嚴重干擾倒頻譜正規化法的效果。

乾淨環境訓練模式				
方法	測試組 A	測試組 B	測試組 C	總平均
Baseline	58.94	58.48	59.97	58.96
LERN	74.36	76.72	63.83	73.20
CMVN	77.27	80.40	72.83	77.64
LERNI+CMVN	80.41	82.98	76.63	80.68

表 4.3.1 對數能量尺度重刻法 I 與倒頻譜正規化法之加成性實驗(乾淨環境訓練模式)

複合情境訓練模式				
方法	測試組 A	測試組 B	測試組 C	總平均
Baseline	85.22	83.99	80.67	83.82
LERN	86.31	86.27	81.22	85.28
CMVN	90.30	90.50	88.48	90.01
LERNI+CMVN	90.46	90.42	88.33	90.01

表 4.3.2 對數能量尺度重刻法 I 與倒頻譜正規化法之加成性實驗(複合情境訓練模式)

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。

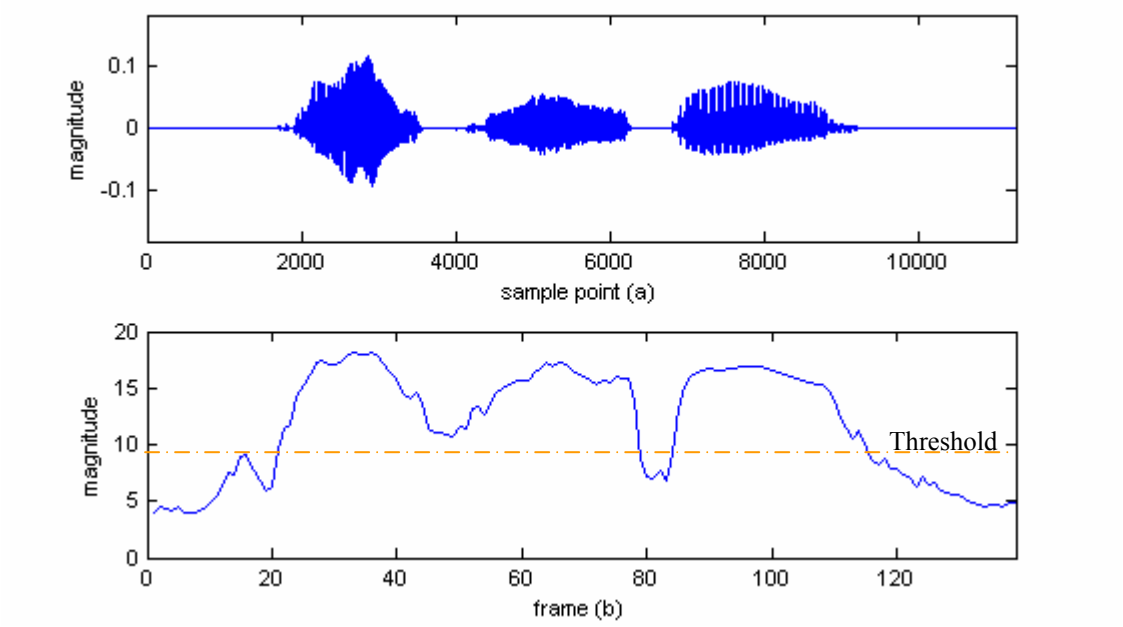

圖 5.1.1 音框能量偵測法圖示

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。

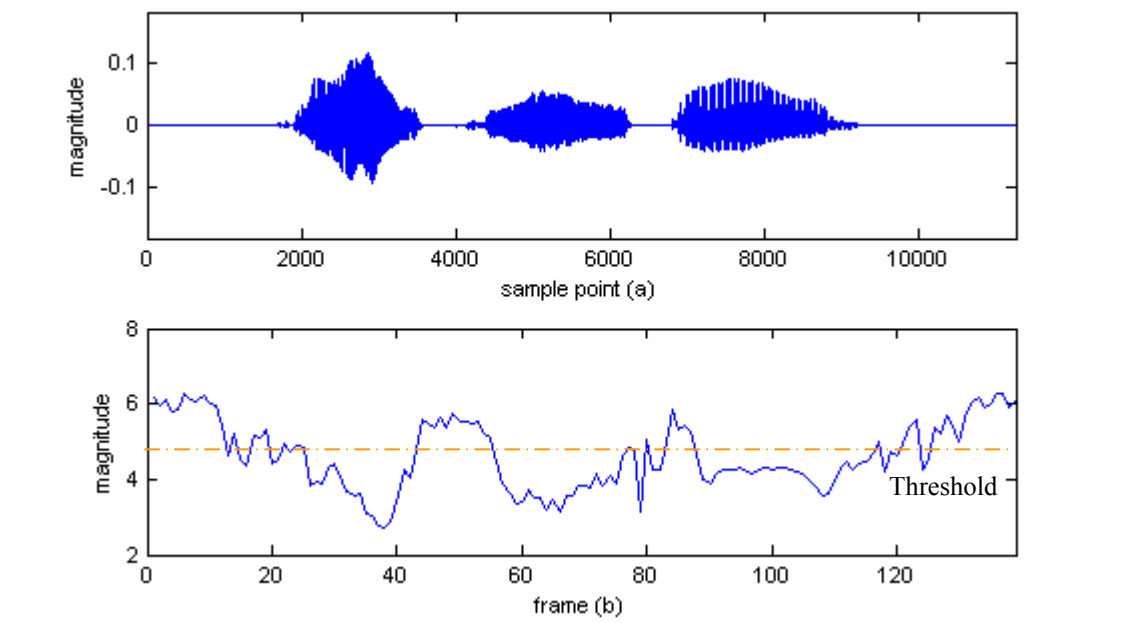

圖 5.1.2 音框熵值法圖示

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。

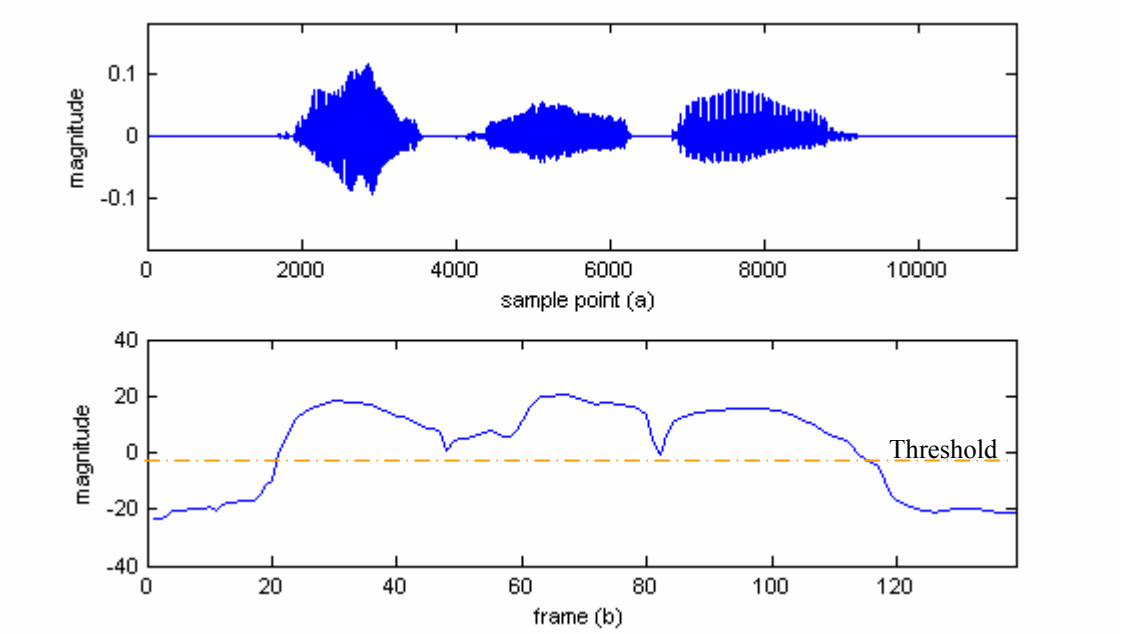

圖 5.1.3 長期頻譜差異值圖示

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(LERNⅠ)，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。

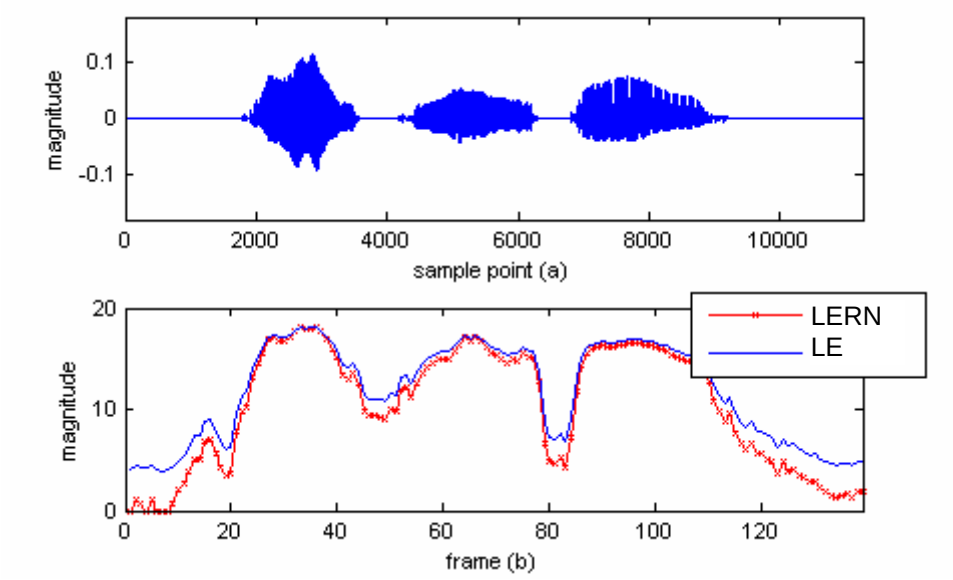

圖 5.3.1 尺度重刻法於音框能量偵測圖示

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

第六章 對數能量為基礎之語音正規化於中文大詞彙連續語音辨識系統

6.1 中文大詞彙連續語音辨識系統

中文大詞彙連續語音辨識系統所使用的語料庫為 MATBN 電視新聞語料[Wang et al. 2005]，為中央研究院資訊所口語小組[SLG]耗時三年與公共電視台[PTS]合作錄製完成。每天一個小時的公視晚間新聞深度報導，收錄了 200 天(一天約一小時)的電視新聞語料，其中包含 2001 年的新聞 30 小時、2002 年 146 小時及 2003 年 24 小時。本論文初步地選擇採訪記者語料作為實驗語材，其中包含 25.5 小時的訓練語料(5,774 句)，供聲學模型訓練之用；1.5 小時的評估語料(292 句)，供辨識評估之用。其中男女語料各半。

在以下小節中將扼要介紹台灣師範大學資工所目前所發展的新聞語音辨識系統，主要包括前端處理、聲學模型訓練、詞典的建立(Lexicon Construction)、語言模型訓練和詞彙樹複製搜尋(Tree-Copy Search)等部分[Chen et al. 2004, 2005]。

(1) 前端處理

實驗中使用梅爾倒頻譜特徵作為語音訊號的特徵參數。在求取梅爾倒頻譜特徵時，將語音資料切割成一連串部分重疊的音框，每一個音框由 12 維的梅爾倒頻譜特徵與 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 13 維的梅爾倒頻譜特徵是由 18 個梅爾頻譜上濾波器組(Filter Banks)的輸出經餘弦轉換求得。同時，對數能量為基礎之語音正規化法將使用於對數能量特徵維度上。

(2) 聲學模型

聲學模型採用連續密度隱藏式馬可夫模型(Continuous Density Hidden Markov Model, CDHMM)，模型的總數量有 151 個，其中包含了 1 個靜音模型，112 個聲母模型，以及 38 個韻母模型。每個模型的狀態數分別為 3 至 6 個不等，每個狀態皆為高斯混合分佈，其中每個高斯混合分佈的分佈個數分別為 1 至 128 個不等。此外，聲母和韻母共有 403 種不同的音節組合。

(3) 詞典及語言模型訓練

本系統使用詞雙連以及詞三連語言模型(Word Bigram and Trigram Language Models)，並以從中央通訊社(Central News Agency, CNA)2001 與 2002 年所收集到的約一億七千萬個中文字語料作為背景語言模型訓練時的訓練資料[LDC]。語音辨識詞典部分則包含有七萬二千個字詞。另外，在本論文中的語言模型使用 Katz 語言模型平滑技術[Katz 1987]，在訓練時是採用 SRL Language Modeling Toolkit (SRILM)，它是一套相當方便且容易使用的語言模型研究工具軟體[SRILM]。

(4) 詞彙述複製搜尋

本系統的大詞彙連續語音辨識方法是採用由左至右(Left-to-right)、音框同步(Frame-synchronous)的詞彙樹複製搜尋方式[Aubert 2002]。在詞彙樹中每個分枝(Arc)代表一個 INITIAL 或 FINAL 的隱藏式馬可夫模型，由樹根(Root)到任一個樹梢(Leaf)的路徑代表一個詞或一些發音相同的詞，路徑上的分枝就是代表這個詞或這些詞會使用到的隱藏式馬可夫模型。具體來說，所採用的詞彙樹複製搜尋演算法，搜尋時每個音框會同時存在數棵詞彙樹複製(Tree Copies)，每個詞彙樹代表不同的語言模型歷史或限制(Language Model History or Constraint)。在每個音框中，若有不完全路徑已抵達樹梢時，代表一個完整詞已可被產生；同時，不同棵詞彙樹複製間已抵達樹梢的不完全路徑，若具有相同的語言模型歷史，則會

進行再結合(Recombination)，保留最大分數者，並以它們的語言模型歷史為標註，產生新的一棵詞彙樹複製，或加入到一棵已存在且具有相同語言模型歷史的詞彙數複製中。另一方面，由於存活的隱藏式馬可夫模型狀態節點可能會隨音框數呈指數倍增加，因此必須以光束剪裁(Beam Pruning)技術適當地剪裁分數較低的狀態節點或不完全路徑。在本系統採用詞單連語言模型前看(Word Unigram Language Model Look-ahead)技術，對每一個詞彙樹複製內部狀態節點，會以其所在分枝(或隱藏式馬可夫模型)之可能拜訪樹梢節點中具最大詞單連語言模型機率，做為該內部狀態節點的語言模型前看分數。此外，在每個音框，會記錄存活的詞彙樹複製樹梢節點中分數較高者的相關資訊(這些樹梢節點本身代表著可能的候選詞)，諸如它們的語言模型歷史、對應候選詞開始與結束的音框以及搜尋時聲學解碼的分數，然後再依此資訊建立起一個詞圖(如圖 6.1)。並在詞圖上使用更高階的語言模型，如詞三連(Trigram)、詞四連(Fourgram)語言模型等，重新進行一次詞圖動態規劃搜尋(Word Graph Rescoring)，找出最佳的文句。在本系統中，詞彙樹複製搜尋階段是使用詞雙連語言模型，而在詞圖搜尋階段則是使用詞三連語言模型。

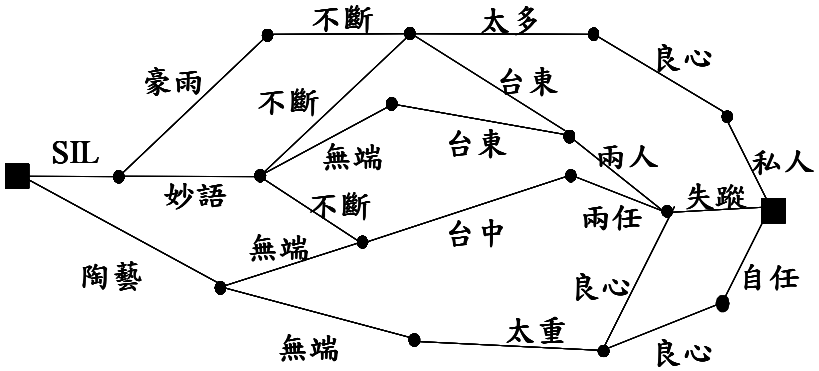

圖 6.1 所有可能的文句組合之詞圖

6.2 中文大詞彙連續語音辨識實驗

在中文大詞彙連續語音辨識結果，我們主要觀察對數能量特徵在經過對數能量尺度重刻法Ⅰ對於中文大詞彙系統的辨識效果，但由於實驗數據如表 6.2.1 中顯示，雖然在正確率的提升與基礎實驗比較只有少許的進步，不容易察覺是否有顯著的提升效果，因此在這裡我們將利用美國國家標準和科技機構 National Institute of Standards and Technology(NIST)所採用的 Matched Pairs Sentence-Segment Word Error (MAPSSWE)[Gillick and Cox 1989]統計式信心度估計的方法來測試對數能量尺度重刻法的效果。MAPSSWE 信心度估測的方法主要須將正確答案當基準，其次比較標準參考方法的答案與新方法所得到的答案。最後當信心度估測中的 *P* 值(*P*-values)檢定機率小於 0.05 即可已說明新的方法有顯著(Significance)的進步。

中文大詞彙連續語音辨識實驗結果：

根據第四章的實驗結果，我們得知對數能量尺度重刻法於歐洲電信標準協[ETSI]會所發行的連續英文數字語料庫有明顯的進步，但 Aurora-2.0 語料實際上只是小詞彙語料庫，對於真實環境下的語音辨識結果並無法明顯表示新技術的效果，因此我們進一步的將對數能量尺度重刻法實做於中文大詞彙連續語音辨識系統中，實驗數據如表 6.2.1，表中分別比較對數能量尺度重刻法Ⅰ在不同尺度設定值的結果，同時也比較對數能量尺度重刻法Ⅱ不同的設定值。此外，我們利用 MAPSSWE 信心度估測的方法檢測在不同的尺度設定值的 *P* 值(*P*-values)，方法Ⅰ中我們設定的尺度大小有 80、100、200、250、500 和 1000，實驗結果顯示當尺度 *M*=500 的時候有最佳的提升，在音節(Syllable)、字(Character)和詞(Word)的正確率都分別有 0.87%、0.98%和 0.82%，信心水準皆高於 99.9%，而在字的正確率部分則大多都可以有顯著的進步表現。方法Ⅱ中的 α 和 β 參數值則是採用表 4.2.6 和表 4.2.7 的最佳值，結果顯示對數能量尺度重刻法Ⅱ對於中文大詞彙連續

語音辨識系統的辨識率沒有實際的幫助，在所有正確率上幾乎都變差，只有當α與β參數分別設定為 1.0 和 0.2 的時候在詞(Word)的部分比基礎實驗多出 0.02%的正確率。最後我們進一步將對數能量尺度重刻法Ⅰ加上倒頻譜正規化法(CMVN)比較，如表 6.2.2，雖然在信心度評估上沒有顯著的進步，但在音節、字和詞的正確率上確實都有進步。

WG	Baseline (MFCC)	LERNI						LERNⅡ	
		M=80	*M*=100	*M*=200	*M*=250	*M*=500	*M*=1000	α : 1.0 β : 0.2	α : 0.98 β : 0.25
Syllable (Acc)	77.03	77.81	77.57	77.52	77.45	**77.90**	77.76	77.01	76.76
Syllable *P* -value		Significance	0.01	0.01	0.01	Significance	Significance	0.5517	0.9394
Character (Acc)	69.30	70.17	70.00	69.88	69.74	**70.28**	70.10	69.18	68.76
Character *P* -value		Significance	Significance	Significance	0.01	Significance	Significance	0.7611	0.9989
Word (Acc)	60.38	60.98	60.95	60.82	60.73	**61.20**	61.04	60.40	59.59
Word *P*-value		0.01	0.01	0.05	0.09	Significance	0.01	0.4721	0.9990

表 6.2.1 對數能量尺度重刻法於中文大詞彙系統實驗結果

表中 WG 為詞圖搜尋，Syllable、Character 和 Word 為音節、字和詞的正確率(Acc)，*M* 代表尺度大小，significance(顯著性)則表示為 *P*-value < 0.001，信心水準高於 99.9%以上。

WG	Baseline + CMVN	LERNI + CMVN		
		M=100	M=500	M=1000
Syllable (Acc)	79.90	79.98	79.94	79.92
Syllable *P* -value		0.31	0.40	0.46
Character (Acc)	72.49	72.55	72.63	72.62
Character *P* -value		0.35	0.19	0.21
Word (Acc)	63.41	63.63	63.72	63.73
Word *P*-value		0.17	0.09	0.08

表 6.2.2 對數能量尺度重刻法與倒頻譜正規化法之加成性實驗結果

第七章 結論與未來展望

結論：

近年來自動語音辨識器已經越來越受到世人的重視，諸如手機的語音撥號功能、門禁自動開關辨識、銀行語音對話系統和電信業者推出的電話秘書語音轉簡訊功能等等，在語音應用系統這方面的產品其相關業者也不斷的推陳出新，正因為如此大眾對語音辨識功能的應用有更大的期待。

然而不論是業界或學界其對於語音辨識功能效果的高標準是一致的，也就是目標為百分之百不會錯的語音辨識率效果。因此本論文對於完美的辨識率目標作為最重要的研究方向，所以進一步的期望能降低噪音對語音訊號所造成的影響，達到提升辨識率是本文討論的重點之一。在研究方法中藉由觀察語句的語音對數能量特徵在不同雜訊環境下的變化，我們試圖尋找一個重建乾淨的語音對數能量特徵的方法。故吾人提出以「對數能量尺度重刻法」來減少噪音的影響，此一方法能簡單且有效地對付不同的環境雜訊干擾，並且可以容易的修正噪音所造成的異常高峰或波谷所造成部份特徵值被過度放大或縮小的特殊情形，亦即是對語音對數能量特徵進行尺度正規化。目前經由實驗數據證明此方法在歐洲電信標準協會(ETSI)發行的 Aurora-2.0 語料庫上的辨識率比傳統梅爾倒頻譜方法的平均詞正確率還要高出 12.51%的提升，並且將此方法實做於中文大詞彙連續語音辨識系統，證明在大詞彙的語料庫中於音節、字和詞的正確率依然都有提升效果，正確率分別提高 0.87%、0.98%和 0.82%。此外，本文討論的另一重點是放在語音端點偵測上，原因就如參考文獻中所提及的效果一樣，若能精確判斷語音訊號之正確端點位置使成為僅包含語音訊號的語句段落，則自然的在非語音訊號部份就不需要作語音辨識，一方面可以降低辨識器的負擔，另一方面也就不會有辨識錯誤的問題發生，進而幫助提高正確的語音辨識效果。

未來展望：

目前階段，對數能量尺度重刻法初步地只針對於音框對數能量來處理。未來應該可以嘗試使用此方法，將對數能量尺度重刻法應用到語音特徵的每一特徵維度上。進一步的研究則可以從圖 7.1 觀察，我們發現實際上在特徵擷取過程中的梅爾濾波器組(Mel-frequency Filterbank)後 23 維濾波器輸出值強度也有如同我們對音框對數能量觀察時的相似現象，亦即在非語音訊號的段落，原本音框上的維濾波器輸出值大小應該偏低，但該維濾波器輸出值卻因為受到噪音的干擾而增加維濾波器輸出值大小的特性。因此我們大膽假設其實尺度重刻的方法對於梅爾濾波器組輸出值也能夠有相同正規化的效果，進而提高語音辨識率。

然而對數能量尺度重刻法的另一個問題是在於尺度(Scale)大小的決定，在本論文中的尺度大小，現階段都是以測試的方式來找出最佳尺度，但理論上應該和對數能量的數值有一定的關聯性。不同尺度大小於對數轉換函數如圖 7.1 所示，圖中我們比較尺度為 50、100、250、500 和 1000 的關係。藉由圖 7.2 的比較得知當尺度較大的時候，等分區間由左到右，對數轉換函數從零開始便快速的提升到對數轉換函數值 0.8 附近，因此音框對數能量經過轉換函數後可以保留較多的等分有較高的對數能量值。此外，轉換函數也可能讓轉換後的對數能量過小造成對數能量曲線不連續的現象，如圖 7.3 橢圓區域範圍中的對數能量曲線，會有突然的波谷(Valley)出現，因此未來的研究範圍可以進階的討論平滑化方法。

綜觀上述的幾個未來方向，可以歸納為以下三點：

1. 梅爾三角濾波器組後的頻率能量當受到環境噪音干擾的能量改變現象相似於語音對數能量上的變化。因此對數能量尺度重刻法應該也能夠適當的減少噪音對頻率能量的影響。
2. 對數能量尺度重刻法的尺度大小設定會影響語音辨識率的效果。
3. 對數轉換函數會造成對數能量過小而產生不連續的現象，可以進階的研究平滑化處理技術。

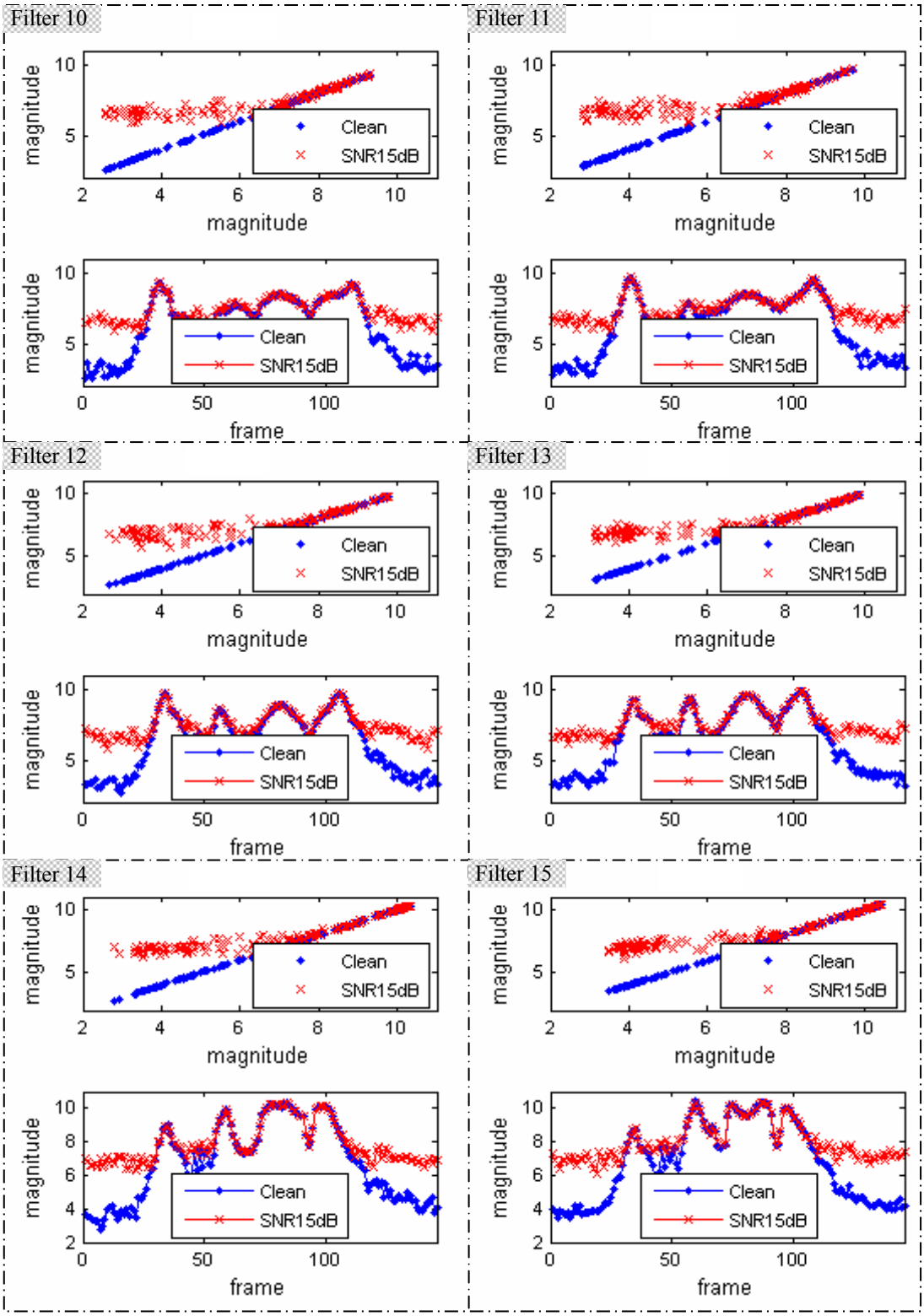

圖 7.1 濾波器(Filter10~ Filter15)受噪音干擾前後比較圖

圖 7.為梅爾三角濾波器組(Filterbank)的強度輸出值，圖中列舉濾波器 10 到濾波器 15 區間的變化情形。第 1、3、5 列為噪音干擾(SNR 15dB)後濾波器輸出值強度對應相同濾波器於乾淨環境下的

比較，第 2、4、6 列則為時間軸上的濾波器輸出值強度變化，語音內容為 139O，該語句的對數能量圖可以參考圖 4.1.4 對數能量示意圖。

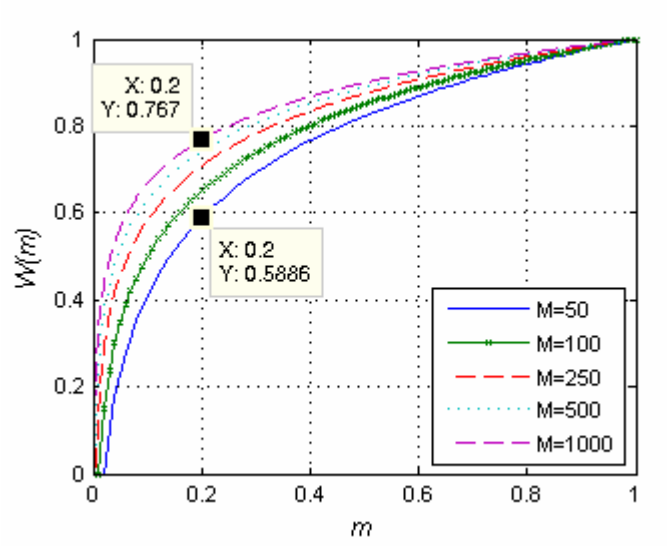

圖 7.2 不同尺度大小之對數轉換函數曲線

如圖 7.2 所示，圖中我們設定尺度為 50、100、250、500 和 1000 的關係，並且將橫軸的各尺度大小正規化到 0～1 的區間，好處在於比較相同等份大小時，容易觀察該等份在不同尺度設定下對數轉換函數值的改變，譬如橫軸刻度為 0.2 的等份大小，則可以發現當尺度為 50 時對數轉換函數值為 0.5886，而當尺度為 1000 時對數轉換函數值為 0.767。因此不同的尺度設定實際上對音框對數能量的影響頗大。

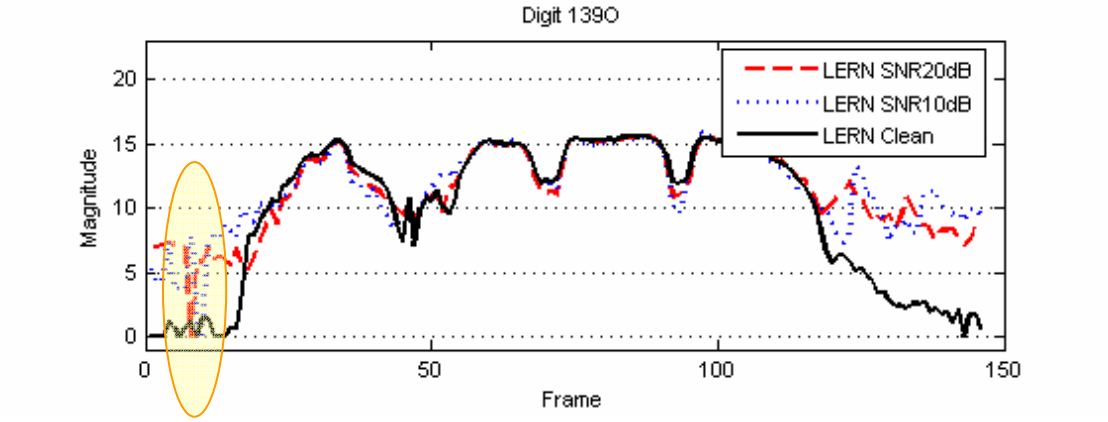

圖 7.3 對數能量尺度重刻法示意圖(語音內容為：139O)

圖 7.3 的橢圓區域範圍中為對數能量尺度重刻法處理後所造成的不連續情形。

參考文獻

[Aubert 2002] X. L. Aubert, “An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition,” Computer Speech and Language, January 2002.

[Boll 1979] S.F. Boll, “Supperssion of Acoutstic Noise in Speech Using Spectral Subtraction,” IEEE Trans. on ASSP, Vol. 27, No. 2, pp. 133-120, 1979.

[Bocchieri and Wilpon 1992] EL Bocchieri, JG Wilpon "Discriminative analysis for feature reduction in automatic speechrecognition," Acoustics, Speech, and Signal Processing, ICASSP 1992.

[Chen et al. 2004] Berlin Chen, Jen-Wei Kuo, Wen-Hung Tsai, “Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription,” in Proc. ICASSP 2004.

[Chen et al. 2005] Berlin Chen, Jen-Wei Kuo, Wen-Huang Tsai, “Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription,” International Journal of Computational Linguistics and Chinese Language Processing, Vol. 10, No. 1, pp. 1-18, March 2005.

[Davis et al. 1980] Davis, S. B. and P Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences." IEEE Transactions on Acoustics, Speech, and Signal Processing 28(4); pp. 357-366, 1980.

[ETSI 2000] H. G. Hirsch, D. Pearce, “The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions,” in Proc. ISCA ITRW ASR 2000.

[Furui 1981] S. Furui, “Cepstral Analysis Techniques for Automatic Speaker Verification,” IEEE Trans. on ASSP, 1981.

[Gauian and Lee 1994] J.L. Gauian and C.H. Lee, “Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,” IEEE Trans. on Speech and Audio Processing, 1994.

[Gomez et al. 2004] R. Gomez, A. Lee, K. Shikano, “Robust Speech Recognition with Spectral Subtraction in low SNR,” in Proc. ICSLP 2004.

[Gillick and Cox 1989] L. Gillick and S. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms", in Proc. ICASSP 89, pp. 532-535.

[Gong 1995] Gong, Y., "Speech Recognition in Noisy Environments:A Survey," Speech Communication 16(3); pp. 261-291.

[Górriz et al. 2006] J.M. G´orriz, J. Ram´ırez, C.G. Puntonet, J.C. Segura, “An Efficient Bispectrum Phase Entropy-based Algorithm for VAD,” in Proc. ICSLP 2006.

[Gillick and Cox 1989] L. Gillick and S. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms", in Proc. ICASSP 89, pp. 532-535. Matched Pairs Sentence-Segment Word Error (MAPSSWE) Test http://www.nist.gov/speech/tests/sigtests/mapsswe.htm.

[Hermansky 1998] Hynek Hermansky, “Should Recognizers Have Ears?”, Speech Communication, 1998.

[Huang and Hon 2001] X. Huang, A. Acero and H. Hon, “Spoken Language Processing: A Guide to Theory, Algorithm and System Development,” Prentice Hall PTR Upper Saddle River, NJ, USA, 2001.

[HTK 2006] S. Young et al., “The HTK Book Version 3.4,” 2006.

[Katz 1987] S. M. Katz, “Estimation of Probabilities from Sparse Data for Other Language Component of a Speech Recognizer,” IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 35, No. 3, pp. 400-401, 1987.

[Leggetter and Woodland 1995] C.J. Leggetter and P.C. Woodland, “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models,” Computer Speech and Language, 1995.

[LDC] Linguistic Data Consortium: http://www.ldc.upenn.edu.

[Lin et al. 2006] Shih-Hsiang Lin, Yao-Ming Yeh, Berlin Chen, "Exploiting Polynomial-Fit Histogram Equalization and Temporal Average for Robust Speech Recognition," the 9th International Conference on Spoken Language Processing (Interspeech - ICSLP 2006), Pittsburgh PA, USA, September 17-21, 2006.

[Misra et al. 2004] H Misra, S Ikbal, H Bourlard, H Hermansky, “Spectral Entropy Based Feature For Robust ASR,” Acoustics, Speech, and Signal Processing, 2004.

[NIST] National Institute of Standards and Technology. http://www.nist.gov/.

[Ramírez 2004] Juan Manuel Górriz, Javier Ramírez, Carlos G. Puntonet, and José Carlos Segura, ”Generalized LRT-Based Voice Activity Detector,” IEEE Signal Processing Letters, Vol. 13, No. 10, October 2006.

[SRILM] A. Stolcke, “SRI language Modeling Toolkit, ” version 1.3.3, http://www.speech.sri.com/projects/srilm/.

[Tai and Hung 2006] Chung-fu Tai and Jeih-weih Hung, “Silence Energy Normalization for Robust Speech Recognition in Additive Noise Environments,” in Proc. ICSLP 2006.

[Viikki and Laurila 1998] O. Viikki, K. Laurila, “Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition,” Speech Communication, Vol. 25, pp. 133-147, August 1998.

[Weizhong and Douglas 2005] Weizhong Zhu and Douglas O’Shaughnessy, ” Log-Energy Dynamic Range Normalizaton for Robust Speech Recognition,” in Proc. ICASSP 2005 pp. 245- 248.

[Wang et al. 2005] Hsin-min Wang, Berlin Chen, Jen-Wei Kuo, and Shih-Sian Cheng, "MATBN: A Mandarin Chinese Broadcast News Corpus," International Journal of Computational Linguistics & Chinese Language Processing, Vol. 10, No. 2, June 2005, pp. 219-236.

[戴仲甫 2006] 戴仲甫, "強健性語音辨認中能量特徵強化及音框選擇之改進技術的研究," 2006.