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)。

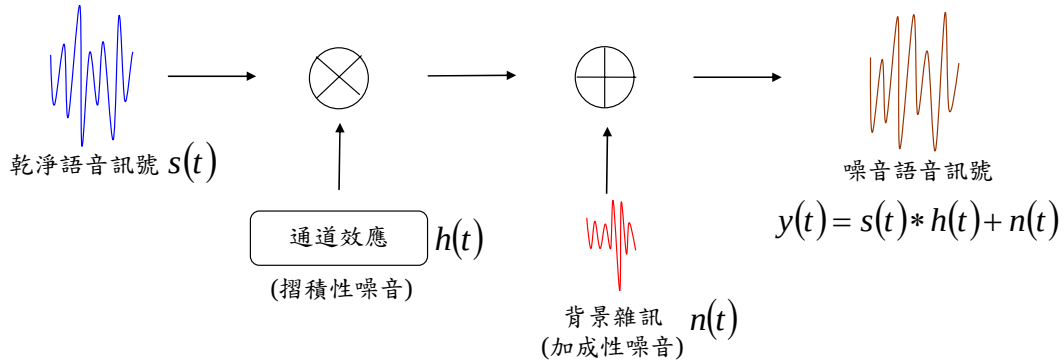

圖 1.3.1 噪音干擾示意圖

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。