第一章　　序論

1.1　研究動機

人類在幾千年的演化過程中，智慧不斷的累積傳承，因此過去文明變遷和人類演化的步伐是一致的。而如今科技進化的速度，卻早已大大的超越了人類演化的速度。日常生活中可以使用的多媒體影音資訊越來越多，例如廣播電視節目，語音信件，演講錄影和數位典藏等。這些多媒體資訊可以從網路上大量地取得，成為傳統文字資訊外社會大眾廣泛使用的資訊來源。顯而易見的是，在上述的絕大部份多媒體中，語音可以說是最具語意的主要內涵之一。例如觀賞電視節目或電影的時候，如果沒有影像只有聲音，雖然無法完整的享受到娛樂的效果，但是對於內容或多或少還是能夠體會的；然而若是沒有聲音只有影像，除了無法感受到娛樂的效果外，對於內容恐怕能夠了解的部份也就更少了。此外在全球造成驚人的商機，也吸引了相當多的企業在該領域積極地投資的電子消費產品，例如 MP3 隨身聽，或者是蘋果電腦企業最近所推出的多功能隨身碟 ipod，普遍的來說體積已越來越小。然而我們人類演化的速度卻遠遠跟不上科技發展的速度，也就是說，手掌並不會跟著越變越小。因此在產品體積的設計考量上，勢必需要在攜帶的便利性以及操作的便利性上做個取捨。

由於輕巧簡便已逐漸成為科技產品外型的設計趨勢，很明顯地，繼續利用雙手來做為人類與產品之間溝通的橋樑已慢慢地不再合適。此外自古以來，語音一直都是人類最自然也最直接的溝通方式，若能改用語音來做為人類和科技產品之間的溝通橋樑，除了具備友善且有效的優點之外，更能省去繁雜的操作手續，也間接的消除了老一輩人對新科技產品的恐懼，而能讓更多人一起來分享高科技所帶來的便利和喜悅。因此吾人相信以語音做為人機之間的溝通管道，將會是未來科技的發展趨勢。

1.2 研究目的

就目前的語音辨識的技術而言，似乎還有一段路要走才能達到理想中的境界。語音辨識通常會遭受到一些複雜的因素干擾，諸如背景噪音 (Background Noise)，以及通道效應 (Channel Effect) 等諸多因素，使得辨識系統始終無法發揮最佳的效用，而辨識率往往也差強人意。因此長久以來，語音強健 (Speech Robustness) 一直都是為大家所重視的研究領域，並希望藉由相關抗噪音的技術的處理，可以提升語音訊號的強健性，降低噪音等相關因素對於訊號本身的影響，進而提升辨識的結果 [Claude *et al.* 1996]。

一般而言，噪音可以分為兩種類型。一種是加成性噪音（Additive Noise），另一種則是摺積性噪音（Convolutional Noise）。所謂的加成性噪音，好比在錄音的時候，除了乾淨語音訊號外所同時收錄的音源。這種噪音在日常生活中很容易接觸到，諸如火車進出站台所發出的聲響，或者是選舉造勢場合中激情的群眾所產生的聲音等，都可視為是加成性噪音的類型。而所謂的摺積性噪音，通常指的就是通道效應（Channel Effect），也就是說，語音訊號在經由不同的傳輸管道後造成改變的現象。例如有時候拿麥克風高歌時，聽到自己聲音反而會有種陌生的感覺。近年來，有越來越多的學者投入精力在研究關於語音強健這方面的領域，因此也有越來越多的語音強健技術被提出。而這些技術的目標都是相同的，也就是希望能藉此來提升語音的強健性，進而提升辨識率，以讓語音這個技術能夠更廣泛的應用在日常生活各方面。根據這些技術的本質，大致上可以分為以下三個方向，而在本篇論文，吾人主要的研究是朝第二個和第三個方向來進行。

1. 語音模型調適技術（Speech Model Adaptation）

這種方法處理的對象是聲學模型（Acoustic Models），而非語音的特徵參數（Feature Parameters）。主要的方式是依照辨識當時的環境來調整聲學模型中機率分佈，像是平均值向量（Mean Vector）和共變異矩陣（Covariance Matrix），

以降低環境不匹配的現象，使得原始的模型經由調適後可以在新的環境下也能有好的表現。這類的方法優點是僅需要少量的調適語料就能對聲學模型進行調適；缺點是在進行即時調適的時候，計算量並不小。常見的方法有最大後機率法則（Maximum A Posterior, MAP）[Lee *et al.* 1983]，最大相似度線性回歸法（Maximum Likelihood Linear Regression, MLLR） [Leggetter 1995]，以及平行模型合併法（Parallel Model Combination, PMC）[Gales *et al.* 1993 ; Hung *et al.* 2002]等。

2. 語音強化技術（Speech Enhancement Techniques）

這類的技術最主要的目的，在於提升語音訊號本身的品質，而非調整辨識系統模型或特徵參數，希望藉由語音和噪音所呈現不同的統計特性，來重建乾淨的聲音訊號或是特徵參數，使受到噪音干擾的語音聽起來會比較接近無噪音環境下的語音。然而實驗證明，在語音強化的過程中，往往也能順帶提升辨識的正確率。常見的技術有：頻譜消去法（Spectral Subtraction, SS）[Boll 1979]，維爾濾波器（Wiener Filter, WF）[Ephraim 1992]，訊噪比波型處理（Signal Waveform Processing, SWP）[Macho 2001]，噪音遮罩法（Noise Masking, NM）[Mellor *et al.* 1992]等。

3. 強健性語音特徵（Robust Speech Features）

這類的技術的主要目的，在於擷取出語音訊號中比較具有強健性的特徵值，也就是不會隨著週遭環境變化而導致有太多失真的參數。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 Normalization, CN）[Viikki 1998]，聲道長度正規化（Vocal Tract Length Normalization, VTLN） [Rose *et al.* 1996]，統計圖等化法（Histogram Equalization, HEQ）[Korkmazsky 2004]，特徵空間旋轉法（Feature Space Rotation, FSR） [Molau 2003]等。

此外如鑑別性分析（Discriminant Analysis）這方面的技術，主要是利用原始資料類別的統計資訊，例如類別共變異矩陣等，以求得轉換矩陣（Transformation Matrix），將原本的特徵向量投影到新的特徵空間，以得到較具鑑別力的語音特徵，並同時抑制噪音的影響。常見的技術有線性鑑別分析（Linear Discriminant Analysis, LDA）[Duda and Hart 1973]，異質性鑑別分析（Heteroscedastic Discriminant Analysis, HLDA）[Gales 1999]，最小分類錯誤評估線性鑑別分析（Minimum Phone Error based Linear Discriminant Analysis, MPE- LDA）[Zhang *et al.* 2005]，彈性鑑別分析（Flexible Discriminant Analysis, FDA）[Hastie *et al.* 1993]，核函數線性鑑別分析（Kernel LDA, KLDA） [Weston *et al.* 1999]等。

除此之外，在最近這兩三年也有學者嘗試以語音頻譜熵值（Spectral Entropy）來作為語音特徵參數 [Mistra *et al.* 2004]，目的是希望藉由熵值來描述訊號中乾淨語音和噪音之間的分佈情形。相關的實驗證明，頻譜熵值特徵在某些條件下，確實是具備抗噪的能力 [Sivadas *et al.* 2005]。

上述三個方向中，第一個方向是屬於語音辨識系統裡後端處理（Back-end Processing）的部份 [Zhu *et al.* 2005]，目的是讓語音辨識器中的隱藏式馬可夫模型（Hidden Mokov Model, HMM）更能適用於實際辨識環境。而其它兩個方向則是屬於前端處理（Front-end Processing）的部份，目的是對語音訊號特徵擷取（Feature Extraction）作改進，以去除環境噪音對於語音訊號的影響。近年來有越來越多的研究，是結合上述不同方向來增加語音的強健性。例如先強化語音，再對語音參數做強健性處理；或者是先求取強健性語音參數後，再接著對聲學模型進行調適。

1.3 研究內容

本篇論文主旨在於對查表式統計圖等化法（Table Based Histogram Equalization, THEQ）和分位差統計圖等化法（Quantile Based Histogram Equalization, QHEQ）

作深入探討，以及與其他強健性語音特徵做結合，諸如倒頻譜消去法（Cepstral Mean Subtraction, CMS），倒頻譜正規化法（Cepstral Normalization, CN），高階倒頻譜正規化法（Higher Order Cepstral Moment Normalization, HOCMN），頻譜熵值特徵（Spectral Entropy Feature, SE）等。並且進一步的和語音強化技術做結合，如兩階段式維爾濾波器（Two Stage Wiener Filter, TWF），而最後整合出一套新的語音強健技術。例如先進行查表式統計圖等化法，再進行高階倒頻譜正規化法；或者先進行兩階段式維爾濾波器，再進行查表式統計圖等化法。

其中關於查表式統計圖等化法，吾人嘗試將參考分佈根據語音的特性分為二至三類，實驗結果證實將參考分佈做分類對於辨識率的確能有所提升。此外吾人也嘗試對頻譜熵值特徵作線性鑑別分析，再與傳統語音特徵參數合併，成為新的特徵參數。

1.4 研究成果

本論文主要的研究貢獻有二：一是將查表式統計圖等化法加以改良，二是對頻譜熵值特徵加以改良。第一個貢獻主要是以查表式統計圖等化法為主，並與其它相關語音強健技術結合來提升語音的強健性，最後將查表式統計圖等化法加以改良為改良式統計圖等化法（Modified Histogram Equalization with Two Classes, MHEQ-2），也就是將參考分佈依據音框的種類分為靜音（Silence）和語音（Speech）。或者更進階的將語音細分為聲母（Initial）和韻母（Final）（Modified Histogram Equalization with Three Classes, MHEQ-3）。雖然將參考分佈分為兩類的作法，國外的學者已經有研究過 [Molau 2003]，不過在國內似乎鮮少有學者將其作用在中文上並進行深入的探討。而吾人根據中文的特性將語音再分為聲母和韻母兩類，經實驗證明的確可以明顯地提升噪音環境下的辨識率。在乾淨語音訓練模式下，相對辨識率分別提升了 6.56% 和 5.75%，將近是查表式統計圖等化法所提升 3.31%的兩倍。而在噪音環境下，MHEQ-3 則表現的比 MHEQ-2 稍好，

不同訊噪比下的平均相對辨識率提升了 80.83%。

第二個貢獻在於對擷取出的頻譜熵值特徵以線性鑑別分析（Linear Discriminant Analysis, LDA）的技術進行降維（Dimension Reduction），再與傳統的語音特徵參數結合來作為新的語音特徵參數。實驗結果證明，加入頻譜熵值後的新的特徵參數的確加強了語音的強健性，在乾淨的環境下，辨識率也相對了提升近 1.00%。再者，若將新的特徵參數和本論文另一個研究主題（THEQ）作結合，更可以達到加成性的效果，平均相對辨識率提升至 5.19%，比單獨使用 THEQ 所提升的 3.31% 還要高了 1.88%。

1.5 論文大綱

本論文的大綱如下：

第二章：主要在回顧與研究有關的語音強健方法，大致上分為五個部份：（一）傳統語音倒頻譜正規化法。（二）探討各種不同的等化法。（三）主要是介紹強化語音的方法。（四）介紹熵值。（五）介紹高階倒頻譜正規化法（Higher Order Cepstral Moment Normalization, HOCMN）[Hsu 2004]。

第三章：主要是介紹實驗用的語料與相關實驗設定。本論文中我們採用兩種語音資料庫，一種是廣為諸多學者所使用的 AURORA 2.0 [ETSI Website : http://www.elda.org/article20.html]，資料內容是由英語發音的連續數字字串；一種是由廣播新聞中所擷取出來的語音檔案，資料是關於華語廣播新聞的新聞內容。吾人決定採用兩種語音資料庫的原因有二：一是為了與國際研究接軌，二是為了能將本論文所發展的技術實際用於中文語音辨識的相關研究。

第四章：關於基礎實驗結果，包含了語音強化法以及強健性語音參數在不同資料庫下的表現，以及其結合之後的效果。首先是探討查表式統計圖等化法，分別在

頻譜與倒頻譜上作用的優劣。接著是探討查表式統計圖等化法對能量維（Log-Energy or C0）進行等化法的好壞。再來探討各式統計圖等化法的差異性，並與其他強健性技術和語音強化法結合後的效果。最後比較上述諸多方式的優劣。

第五章：將試圖對傳統的統計圖等化法做改良，把語音依據不同的特性分為二類或三類，並根據各自的種類，給予不同的統計圖資料來做為等化的方式，並與上述諸多技術結合進而得到新的架構，並探討其效用。

第六章：將頻譜熵值加以改良，也就是先對其進行線性鑑別分析後，再與傳統梅爾倒頻譜特徵係數結合，成為新的特徵參數，並且進而和查表式統計圖等化法結合，更加地提升語音的強健性。

第七章：關於本論文的研究結論，以及未來研究方向。

最後是參考文獻。