對數能量特徵正規化於語音辨識之進一步研究

陳鴻彬
台師大資工所
james@csie.ntnu.edu.tw

林士翔
台師大資教所
69308027@cc.ntnu.edu.tw

陳柏琳
台師大資工所
berlin@csie.ntnu.edu.tw

摘要

本論文主要探討對數能量正規化技術於語音辨識之應用。現今語音辨識系統常會因語音訊號受環境雜訊干擾而產生某種程度上的影響；因此，語音強健式(Speech Robustness)技術的發展長久以來一直被視為一個非常重要的研究領域，過去已有許多方法成功地被提出，可以在消除或抵抗雜訊上有不錯的效果。其中，不少研究指出語音能量參數的正規化對於雜訊環境下語音辨識影響甚鉅。因此，在本論文我們提出一種新的語音能量特徵正規化方法－對數能量尺度重刻(Log Energy Rescaling, LER)，以使用對數轉換函數方式來對語音對數能量參數作正規化。同時，我們亦與目前幾種廣泛被使用的語音能量特徵正規化方法作比較。實驗語料庫是採用由歐洲電信標準協會所發行的AURORA-2 語料；實驗結果初步地證實本論文所提出之方法，能有效減少語音能量受雜訊干擾所造成的失真情形，進而提昇辨識效果。

關鍵詞：自動語音辨識、語音強健技術、對數能量特徵。

1. 前言

現今自動語音辨識(Automatic Speech Recognition, ASR)系統在語音訊號不受噪音干擾的實驗室環境下，通常可獲得不錯的辨識效果，但若應用至實際日常生活環境中，則往往會因為環境中複雜因素的影響，造成訓練環境與測試環境存在環境不匹配(Environmental Mismatch)的差異，使得系統辨識效能大幅度降低。環境中複雜因素包括背景噪音(Background Noise)、錄音設備本身產生的噪音或是通道效應(Channel Effect)等。正因如此，語音強健(Speech Robustness)技術長久以來一直被視為重要的研究課題，主要是希望藉由對訊號本身、語音特徵參數或是模型參數做適當的處理與調整，以減緩雜訊干擾的影響、降低訓練環境與測試環境不匹配的情形、提升語音訊號及語音特徵參數本身的強健性，進而提高系統辨識效能。

環境中干擾語音訊號的雜訊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)。加成性噪音為錄製語音時，原始語音與背景噪音以線性加成(Linearly Additive)的關係同時被收錄進去，例如周遭人聊天的聲音或是機器設備所發出的噪音等；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一來表示。

語音強健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雜訊對語音訊號的影響，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[11]：

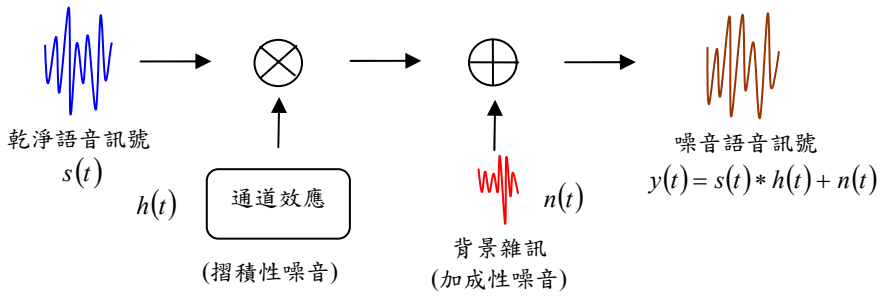
1. 語音強化技術(Speech Enhancement)
目的在於提升語音訊號本身的品質，通常是假設語音訊號與雜訊訊號二者在統計上是不相關(Uncorrelated)，希望能由觀察到的雜訊語音(Noisy Speech)重建出乾淨語音(Clean Speech)訊號。常見的技術有頻譜消去法(Spectral Subtraction, SS)[6]、維爾濾波器(Wiener Filter, WF)[10]等。

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)[7]、倒頻譜正規化法(Cepstral Mean and Variance Normalization, CVN)[1]等。

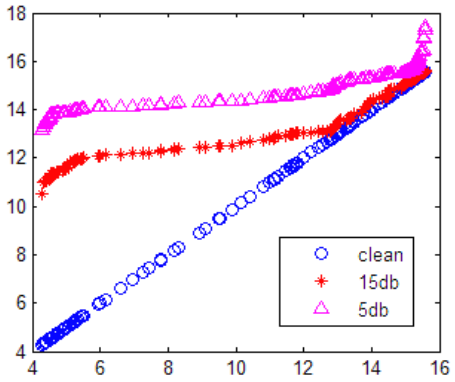
3. 聲學模型調適(Acoustic Model Adaptation)
藉由少量的調適語料(Adaptation Data)對由乾淨語音所訓練而成的聲學模型作調整，主要調整聲學模型中機率分佈的參數，如平均值向量(Mean Vector)或共變異矩陣(Covariance Matrix)。期望調適後的模型可以適用於新的環境，以降低環境不匹配的現象。常見的技術有最大事後機率法則(Maximum a Posteriori, MAP)[4]、最大相似度線性回歸法(Maximum Likelihood Linear Regression, MLLR)[2]等。

在本論文我們提出一種新的語音能量特徵正規化方法－對數能量尺度重刻(Log Energy Rescaling, LER)，以使用對數轉換函數方式來對語音對數能量參數作正規化，此一方法基本上是屬於上述第二類強健性語音特徵處理。同時，我們亦與目前幾種廣泛被使用的語音能量特徵正規化方法作比較。實驗語料庫是採用由歐洲電信標準協會所發行的 AURORA-2 語料；實驗結果初步地證實本論文所提出之方法，能有效減少語音能量受雜訊干擾所造成的失真情形，進而提昇辨識效果。

本論文後續章節安排如下：第二章節將闡述我們所提出之對數能量尺度重刻方法；第三章介紹目



圖一 雜訊干擾示意圖



圖二 加成性噪音對語音特徵參數的影響

前幾種廣泛被使用的語音能量特徵正規化方法；第四章為實驗與討論；第五章為結論。

2. 對數能量尺度重刻(Log Energy Rescaling)

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零。另一方面，當一段語句受到不同的錄音環境影響時，其語音特徵向量序列的對數能量所對應維度特徵值的變化是最容易被觀察出。因此，我們可以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，來試圖重建出乾淨的語音對數能量特徵，亦即對於語音對數能量特徵進行正規化。

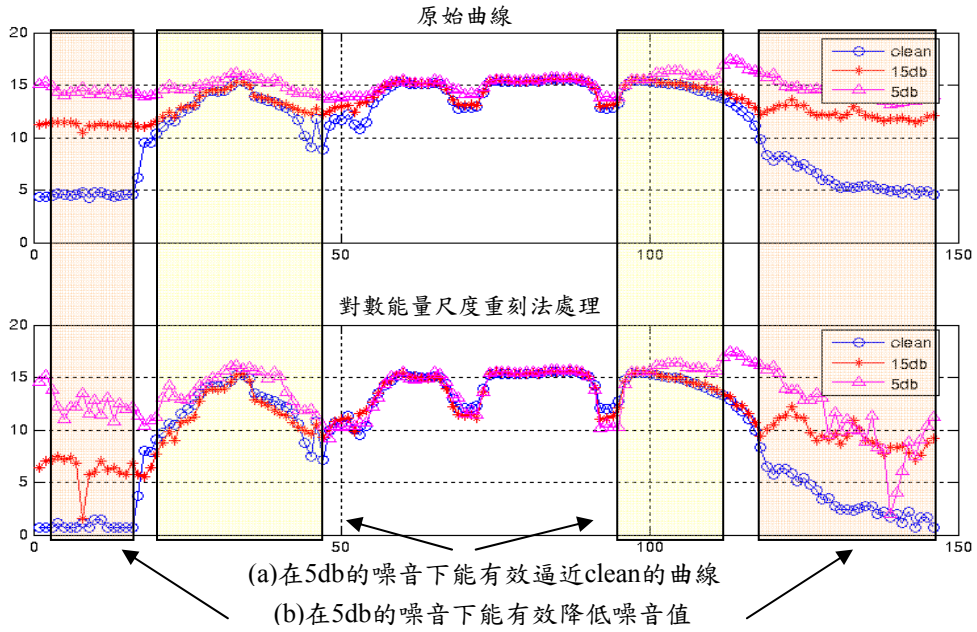
雜訊環境對於語句的對數能量特徵的影響可用圖二來作說明：其中藍色(圓形)點是以乾淨語句每一音框(Frame)的對數能量同時當作X軸與Y軸座標值所繪出；紅色(星形)點是以乾淨語句每一音框的對數能量當作X座標值，以及加上訊噪比(Signal-to-Noise Ratio, SNR)15dB雜訊後的對數能量當作Y軸座標值所繪出；粉紅色(三角形)點是以乾淨語句每一音框的對數能量當作X座標值，以及加上訊噪比 5dB 雜訊後的對數能量當作Y軸座標值所繪出。由圖二可得知，當受到雜訊影響時將會使得對數能量產生非線性的失真：在對數能量較高的音框僅有輕微的影響；相反地，在對數能量較低的音框則會有較為嚴重的影響，整個對數能量值被提高，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)若以對數能量來區隔愈顯不易。或者，我們亦可用圖三上圖來說明雜訊環境對於語句對數能量特徵的影響：其中X座標表示連續的語音音框；Y軸座標代表在每一音框的對數能量；藍色(圓形)曲線代表原始乾淨語音的對數能量；紅色(星形)與粉紅色(三角形)曲線分別代表訊噪比為 15 與 5dB 的雜訊語音的對數能量。由圖三亦可看出對數能量較低的音框，受雜訊的影響較為較嚴重，反之亦然。值得注意的是，在(a)區間與(b)區間，原本是屬於非語音區間，但受到雜訊的干擾後，使得其對數能量相對提升許多。因此，我們認為上述情形可能就是造成乾淨和雜訊語音訊號二者間對數能量統計特性差異的主要原因之一。

基於上述的觀察，本論文提出一種簡易有效的語音能量特徵正規化方法－對數能量尺度重刻(Log Energy Rescaling, LER)，以使用對數轉換函數(如圖三所示)方式來對語音對數能量作正規化。其基本原理是將特徵能量值乘上其所處分位差(Quantile)區間對應的對數轉換函數值，而此一對數函數值介於 0 到 1 之間。因此，經此一對數能量尺度重刻處理過後，將會使原來對數能量值較低的語音段落(或音框)其對數能量值越低、對數能量值較高的語音段落(或音框)其對數能量值幾乎維持不變，試圖讓雜訊語句在經正規化後與原始乾淨語句有相近的對數能量值域，其具體作法如下。首先，從每一語句(測試及訓練語句)所有音框 T 中找出最大對數能量值 $LE\_\max$ 以及最小對數能量值 $LE\_\min$ ：

$$E\_max = \underset{1\le i\le T}{Max}\, E[i],\qquad E\_min = \underset{1\le i\le T}{Min}\, E[i] \tag{1}$$

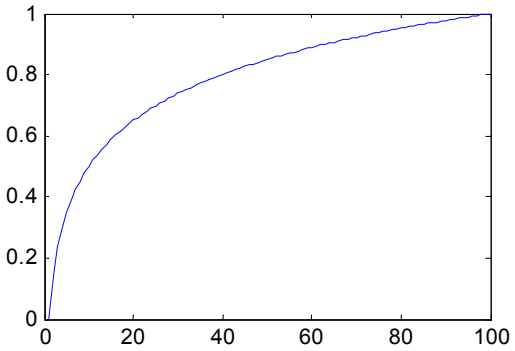
然後根據 $LE\_\max$ 及 $LE\_\min$ 決定對數能量值域範圍，並將此一範圍等份成 M 個分位差，每個分位差寬度 L 可表示成：

$$L = \frac{E\_max - E\_min}{M} \tag{2}$$



圖三 對數能量尺度重刻法處理前與處理後示意圖(語音內容為：139O)

在本論文我們初步將分位差個數 M 設為 100，而每



圖四 對數表曲線

個分位差 m 對應的對數轉換函數值為：

$$W(m)=\frac{\log(m)}{\log(M)} \tag{3}$$

另一方面，每個音框 i 對數能量所落至分位差的索引可表示成：

$$Index_i=\left\lfloor \frac{E[i]-E\_min}{L} \right\rfloor \tag{4}$$

因此音框 i 經正規化後的對數能量 $\hat{E}[i]$ 可以表示成：

$$\hat{E}[i]=E[i]\times W(Index_i) \tag{5}$$

從圖三的上圖看出，未經處理的雜訊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圖差異甚大。尤其以 5db 雜訊干擾的情況下，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，不同訊噪比狀況下對數能量值曲線圖差異縮小了，我們可以容易地由圖三下圖的(a)區間與(b)區間分別觀察出我們所提出對數能量尺度重刻方法能在某種程度上將雜訊語音對數能量曲線逼近乾淨語音對數能量曲線。

3. 其它常見之能量特徵正規化方法

依據第二章的觀察可得知，通常受到雜訊影響時將會使得語音訊號的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嚴重的影響，整個對數能量值被提高。在下面幾小節，我們將簡介幾種過去被提出用以重建乾淨語音對數能量的方法。

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

音框能量消去法主要假設雜訊語音為語音訊號與雜訊訊號加成的結果，因此若要得到乾淨的語音訊號能量，只須要將含有雜訊的語音能量扣掉雜訊能量，特別是針對測試語料每一語句的任何音框來作處理。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 的雜訊語音能量、乾淨語音能量以及雜訊本身的能量，而三者間有式(6)的關係：

$$E_y[i]\cong E_x[i]+E_n[i] \tag{6}$$

然而，$E_n[i]$ 事先並未能知道，必需透過特定方式估測而得。在本論文，我們進一步假設每一語句的前 K 音框為雜訊音框，以取得估計值 $\hat{E}_n$ 來取代 $E_n[i]$，如式(7)所示：

$$\hat{E}_n = \frac{1}{K}\sum_{i=1}^{K} E_y[i] \tag{7}$$

在得到 $\hat{E}_n$ 估計值後，我們採用雜訊遮蔽法以防止 $E_n[i]-\hat{E}_n$ 出現負值來取得近似的乾淨語音訊號能量，如下(8)：

$$E_x[i] = \begin{cases} Ey[i]-\alpha\hat{E}n & , Ey[i] > \hat{E}n \\ \beta Ey[i] & , Ey[i] \le \hat{E}n \end{cases} \tag{8}$$

其中 α 為過度估測因子，β 為底限因子。

3.2 對數能量動態範圍接近法(1) (log Energy Dynamic Range Normalization, LERN1)

對數能量動態範圍接近法(1)的目標是針對於乾淨的語音訊號做預處理，讓乾淨語音訊號的對數能量可以逼近雜訊語音訊號的對數能量，使得經由對數能量動態範圍接近法(1)處理過的訓練語料所練出的語音模型會與測試的雜訊語音訊號兩者間互相匹配。

對數能量動態範圍接近法(1) 的考慮是：語音訊號能量曲線波峰部位，不容易受到雜訊影響；而在波谷部位會受到雜訊嚴重干擾，以致於乾淨語音與測試語音在能量特徵上有不匹配的現象。因此利用一個線性處理，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。具體作法如下。對每一則訓練語句的所有 T 個音框能量中找出最大對數能量值 $LE\_\max$ 以及最小對數能量值 $LE\_\min$ ：

$$\begin{aligned} LE\_\max &= \underset{1\le i\le T}{Max}\log E[i] \\ LE\_\min &= \underset{1\le i\le T}{Min}\log E[i] \end{aligned} \tag{9}$$

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 $D.R.$ ，定義出測試語音檔要調整的最小音框能量值 $T\_\min$ ：

$$\begin{aligned} &D.R. = 10\times\frac{LE\_\max}{T\_\min} \\ &\text{or } T\_\min = \alpha\times LE\_\max \end{aligned} \tag{10}$$

然後，我們可以針對每句訓練語句的每個音框的對數能量作更新：

$$\begin{aligned} &if \quad LE\_\min < T\_\min \\ &then \\ &\log\hat{E}[i] = \log E[i] + \frac{T\_\min - LE\_\min}{LE\_\max - LE\_\min}\left(LE\_\max - \log E[i]\right) \end{aligned} \tag{11}$$

其中 $\log\hat{E}[i]$ 為更新後之對數能量。由式(11)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接近法(1)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。

3.3 對數能量動態範圍接近法(2) (log Energy Dynamic Range Normalization, LERN2)

對數能量動態接近法(1)則是以對乾淨的語音訊號作處理，然而對數能量動態範圍接近法(2)選擇將前述方法做測試語料和訓練語料的反向假設修正，使其需要修正的處理對象為受雜訊干擾的測試語音訊號，並將原本方法(1)的線性的處理方式改變成利用非線性方式做處理，使訓練與測試的訊號能有更好的匹配效果，期使雜訊語音訊號在時間軸上有波谷的地方更為接近乾淨語音訊號。其具體作法如下：從每一則測試語句的所有音框能量中找出最大音框能量 $LE\_\max$ 以及最小音框能量 $LE\_\min$ ，如式(9)所示。利用事前預期的測試語音之噪音干擾大小決定一個動態能量的範圍 $D.R.$ 值，定義出要調整的目標最小能量音框 $T\_\min$ 如式(10)。最後根據 $T\_\min$ 針對每則測試語音檔的每個音框對數能量作更新：

$$\begin{aligned} \hat{E}[i] = E[i] &- \frac{LE\_\min - T\_\min}{\log(LE\_\max) - \log(LE\_\min)} \\ &\times\left(\log(LE\_\max) - \log(E[i])\right) \end{aligned} \tag{12}$$

其中 $\hat{E}[i]$ 為更新後之對數能量。

3.4 靜音音框對數能量正規化法(1) (Silence Log-Energy Normalization, SLEN1)

靜音音框對數能量正規化法(1)是一個非常簡單而有效的方法，主要利用語音的端點偵測(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，由於雜訊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。由參考的實驗證明若只保留波峰部分經由辨識依然可以得到不錯的辨識率。因此大膽假設，對能量特徵而言最重要的是語音能量曲線而非能量失真的距離，也就是說一段語音整體的能量曲線比音框能量的降低或昇高失真距離還重要，若可以保留清晰的曲線，就可以得好的辨識率。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雜訊的干擾，因此正規化處理過後，乾淨的語音訊號能量波形將會與受雜訊影響的語音能量的曲線相似。

具體作法是利用雜訊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(1)。靜音音框對數能量正規化法(1)主要利用能量來判斷語音與非語音的門檻值，並且利用式(14)把非語音音框正規化為一個常數。數學式如下：

$$\tau = 1.2\times\frac{1}{T}\sum_{i=1}^{T}\log E[i] \tag{13}$$

$$\log \hat{E}[i]_{i=1\ldots T} = \begin{cases} \log E[i] & \text{if } E[i] \geq \tau \\ \Phi & \text{if } E[i] < \tau \end{cases} \quad (14)$$

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能量，τ 為門檻值而 Φ 為一個常數。

3.5 靜音音框對數能量正規化法(2) (Silence Log-Energy Normalization, SLEN2)

靜音音框對數能量正規化法(2)，同靜音音框對數能量正規化法(1)的假設，而判斷語音與非語音的門檻值方法改用變動音框位移(Variable Frame Rate)來做。借用變動音框位移(VFR)來作為雜訊音框的選擇方式。式(15)為計算變動音框位移之方法：

$$\log y[i] = \frac{1}{2}\left(\log E[i+1] - \log y[i-1]\right) \quad (15)$$

$\log y[i]$ 為 $\log E[i]$ 所對應的輸出; $\log E[i]$ 為每個音框的對數能量值。若計算出的 $\log y[i]$ 值小於門檻值 τ 就認定為非語音音框，而 τ 的計算值與整個演算法如下式(16)與(17)：

$$\tau = \frac{1}{T}\sum_{i=1}^{T} \log y[i] \quad (16)$$

$$\log \hat{E}[i]_{i=1\ldots T} = \begin{cases} \log E[i] & \text{if } y[i] \geq \tau \\ \Phi & \text{otherwise} \end{cases} \quad (17)$$

3.6 動態音框能量搜尋法 (Energy Search Based Variable Frame Rate Analysis)

動態音框能量搜尋法[5]，主要目的是從原本的對數能量特徵值找出一個較具有代表特性的一階差量特徵值來取代原本的對數能量。希望藉此方法能夠改善噪音對能量干擾的影響結果。作法上則是找出每一個音框之後的鄰近音框對數能量與當前音框對數能量的最大一階差量，此一階差量既用來取代原本的對數能量。而這裡所指的鄰近音框則被設定在一個固定範圍之間，並在此固定範圍之間尋找出一個與目前音框對數能量具有最大的一階差量。由參考的實驗結果得知，利用這一階差量特徵值取代原本對數能量特徵值能夠有更好的辨識效果。演算法如下式(18)：

$$\Delta E_m = \underset{K_{\min} \leq k \leq K_{\max}}{\arg\max} \frac{\log(E_{m+1}(k)) - \log(E_m)}{k} \quad (18)$$

其中 ΔE_m 為找到的音框能量的最大一階差量值，k 值則將介於目前音框往後的 $K_{\min}$ 與 $K_{\max}$ 的區間內尋找，而 $E_{m+1}(k)$ 即是找到的鄰近音框能量值。

4. 實驗與討論

4.1 實驗架構與設定

本論文實驗所使用的語料庫 Aurora-2 是由歐洲電信標準協會（European Telecommunications Standards Institute, ESTI）所發行[3]，其本身為一套含有雜訊的連續英文數字語料，其中雜訊包含八種來源不同的加成性噪音和二種不同特性的通道效應。語料庫中的加成性噪音包括機場（Airport）、人聲（Babble）、汽車（Car）、展覽會館（Exhibition）、餐廳（Restaurant）、地下鐵（Subway）、街道（Street）及火車站（Train Station），且依照不同訊噪比（Signal-to-Noise Ratio, SNR）各自加入乾淨的語音裡，訊噪比包括 20dB、15dB、10dB、5dB、0dB 和-5dB；通道效應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雜訊以及加成性雜訊之種類不同，Aurora-2 分為三組測試群組 Set A、Set B 和 Set C，Set A 所呈現的雜訊是屬於穩性（Stationary）雜訊，Set B 則是非穩性（Nonstationary）雜訊，Set C 除了穩性與非穩性雜訊外，還使用與訓練語料不同的通道效應。

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，其中包含 16 個狀態（State），並且每個狀態是利用 3 個高斯混合分佈（Gaussian Mixture Distribution）表示。另外靜音模型的部份有二種模型，一個為靜音（Silence）模型，包含三個狀態，用來表示語句開始跟結束時的靜音；另一個為間歇（Pause）模型，包含六個狀態，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都是使用 HTK 工具套件[9]完成。

前端處理方面（Front-End Processing），本論文的基礎實驗是採用梅爾倒頻譜係數（Mel-Frequency Cepstral Coefficients, MFCCs）作為語音特徵參數，取樣音框長度（Frame Length）為 25 毫秒，音框間距（Frame Shift）為 10 毫秒，每個音框的資訊是以 39 維表示，其中包含 12 維的梅爾倒頻譜係數以及一維的對數能量（Log Energy），同時會對 13 維特徵參數取其相對的一階差量係數（Delta Coefficient）和二階差量係數（Acceleration Coefficient）。

4.2 對數能量刻度調適技術實驗

本小節將探討三組實驗。第一組實驗我們觀察對數能量尺度重刻法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，實驗結果如下表一(使用 100 個分位差)。表格內所呈現的數據皆為平均詞正確率，由 Aurora-2 中三組實驗群組(Sets A, B 及 C)中不同訊噪比(20dB 至 0dB)的辨識結果加總平均而得。結果發現同時使用在訓練語料和測

試語料的情況下有較好的精確率。分析結果後我們認為主要是因為訓練語料在低能量部分並沒有完全逼近靜音情況，而

表一 不同的資料處理對象實

	資料處理對象	平均字精確率 ACC(%)			
		Set A	Set B	Set C	平均
乾淨語料訓練模式	Train + Test	74.35	76.72	63.84	71.64
	Test Only	72.20	75.45	60.58	69.41
複合情境訓練模式	Train + Test	86.31	86.27	81.22	84.60
	Test Only	77.76	81.98	69.86	76.54

表二 不同刻度之對數能量尺度重刻法比較

	Scale	平均字精確率 ACC(%)			
		Set A	Set B	Set C	平均
乾淨語料訓練模式	50	74.10	76.71	63.08	71.30
	100	74.35	76.72	63.84	71.64
	500	73.49	75.32	63.77	70.86
	1000	72.93	74.68	63.58	70.39
複合情境訓練模式	50	86.33	86.25	81.04	84.54
	100	86.31	86.27	81.22	84.60
	500	86.51	85.98	81.66	84.71
	1000	86.51	85.90	81.59	84.67

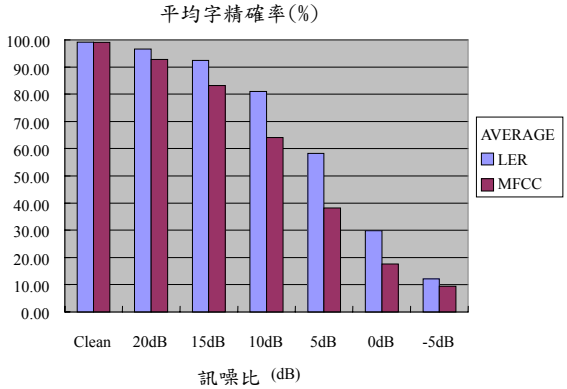
低能量部分仍然會有相當多的噪音值產生。所以訓練語料和測試語料若有同時經過我們的方法處理後會有同樣逼近於零值的匹配效果。

第二組實驗：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 個分位差分別設定為 50、100、500 與 1000 四種尺度做觀察。最後產生的結果如表二。詞正確率同實驗一是由 Aurora-2 中三組實驗群組（Sets A, B 及 C）中不同訊噪比（20dB 至 0dB）的辨識結果加總平均而得。其中以 100 個尺度設定下，乾淨語料訓練模式的效果最好，比傳統 MFCC 作法高出 12.51%的詞正確率。

實驗三中我們則比較傳統 MFCC 方法與使用對數能量尺度重刻法於不同噪音程度的干擾後的詞正確率結果，如圖五所示。在此，對數能量尺度重刻法的分位差設定為 100。實驗結果顯示在不同的訊噪比情況下，對數能量尺度重刻法皆有相對的詞正確率提升作用。

4.3 能量特徵之強健式技術比較

此一小節中我們將比較第二章節所提到的各種方法與對數能量尺度重刻法的比較。其中，3.1 小節音框能量消去法(FES)的參數設定為；$M=5, \alpha=0.95$與$\beta=0.05$。3.2 與 3.3 小節對數能量動態範圍接近法(LERN)動態能量的範圍設定為 $D.R.=17$。3.4 與 3.5 小節靜音音框對數能量正規化法(SLEN)中的常數 Φ 則設定為 0。最後 3.6 小節動態音框能量搜尋分析法(ESVFR)的最小範圍 $K_{\min}$ 設定為 60 而最大範圍 $K_{\max}$ 則設定為 200。實驗結果如下頁表三所示。由實驗數據可以發現我們的方法可以比其他方法更有效的提升詞正確率。



圖五 比較 MFCC 與 LER 之平均字精確率

4.4 對數能量尺度重刻法與正規化方法結合

最後本論文嘗試將所提出的對數能量尺度重刻法與現有的正規化方法進行結合，而正規化方法分別為倒頻譜正規化法（CVN）和多項式擬合統計圖等化法（Polynomial-Fit Histogram Equalization, PHEQ）[8]。實驗將分別比較單純使用正規化方法的結果與結合對數能量尺度重刻法之後的結果，結果如下頁表四。其中 LER+PHEQ13 表示使用我們的方法後直接加上 PHEQ 方法來操作實驗，而 LER+PHEQ12 則在能量特徵值上只有使用我們的方法，但能量特徵值這一維度上不再使用 PHEQ 方法。實驗結果可以看出我們的方法對於其他正規化方法是有直接的加乘效果。

5. 結論與未來展望

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。因此提出以「對數能量尺度重刻法」來減緩噪音的影響，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。最後經由實驗數據證明次方法比傳統梅爾倒頻譜方法的平均詞正確率還

表三 能量特徵之強健式技術比較

	方法	平均字精確率 ACC(%)			
		Set A	Set B	Set C	平均
乾淨語料訓練模式	MFCC	58.94	58.48	59.97	59.13
	FES	70.60	71.20	60.90	67.57
	LERN1	73.08	75.83	59.85	69.59
	LERN2	69.93	69.87	59.85	66.55
	SLEN1	63.97	68.45	50.48	60.97
	SLEN2	69.93	74.59	55.85	66.79
	ESVFR	68.52	69.65	61.61	66.59
	LER	74.35	76.72	63.84	71.64

要高出 12.51% 的提升，但目前這一個方法只適用於音框能量來處理，相較於百分之百的詞正確率目標仍有很大的空間可以進步。未來我們將嘗試將我們所提出的對數能量調整作法應用到不同的語音辨識問題上。

6. 參考文獻

[1] A. Viikki and K. Laurila, “Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition,” Speech Communication, Vol. 25, 1998.

[2] 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.

[3] 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.

[4] 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.

[5] J Epps, EHC Choi, “An Energy Search Approach to Variable Frame Rate Front-End Processing for Robust ASR,” Interspeech 2005-UROSPEECH.

Speech Using Spectral Subtraction,” IEEE Trans. on

[6] S.F. Boll, “Supperssion of Acoutstic Noise in ASSP, Vol.27, No.2, pp.133-120, 1979.

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

[8] S.H. Lin, Y.M. Yeh and B. Chen, “Exploiting Polynomial-Fit Histogram Equalization and Temporal Average for Robust Speech Recognition,” in Proc. ICSLP 2006.

[9] S. Young et al., “The HTK Book Version 3.3,” 2005.

[10] 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.

[11] Y. Gong, “Speech Recognition in noisy environments: A survey,” Speech communication, Vol.16, 1995.

表四 對數能量尺度重刻法與正規化方法結合

	方法	平均字精確率 ACC(%)			
		Set A	Set B	Set C	平均
乾淨語料訓練模式	PHEQ	79.08	81.88	74.32	78.43
	CMN	77.27	80.40	72.83	76.84
	LER+PHEQ13	79.39	81.59	74.85	78.61
	LER+PHEQ12	80.68	82.74	76.49	79.97
	LER+CMN	80.41	82.98	76.63	80.01
複合情境訓練模式	PHEQ	90.09	90.59	86.86	89.18
	CMN	90.30	90.50	88.48	89.76
	LER+PHEQ13	89.70	90.20	86.66	88.85
	LER+PHEQ12	90.42	90.39	89.26	90.02
	LER+CVN	90.46	90.42	88.33	89.73