國立中山大學資訊工程學系

碩士論文

Department of Computer Science and Engineering

National Sun Yat-sen University

Master Thesis

能量特徵重刻對噪音強健性語音辨識之影響

Rescaled Energy Cepstral Coefficients for Noise-robust Speech Recognition

研究生： 許妙鸞

Miau-Luan Hsu

指導教授： 陳嘉平 博士

Dr. Chia-Ping Chen

中華民國一０一年六月

June 2012

摘要

我們針對前端特徵擷取對噪音強健性的影響，提出資料導向之能量特徵重刻法(Data-driven energy cepstral coefficients rescaling, DECCR)對能量特徵參數進行重刻。目的是希望能夠減少含噪音與乾淨語料之特徵值的差異性。然而特徵向量中的對數能量與倒頻譜參數 c_0 對辨識率的效能影響最大，因此我們將針對此兩種能量特徵參數做進一步的處理。

我們以一般辨識系統常用之梅爾倒頻譜參數與 Teager 能量倒頻譜參數爲基礎，再利用資料導向之能量特徵重刻法重刻能量特徵參數。而資料導向之能量特徵重刻法主要分爲語音動態偵測、分段對數尺度函數以及參數搜尋法三個部分。我們使用語音動態偵測判斷語音與非語音出現的段落，並利用分段對數尺度函數給予不同尺度的權重，在函數中所用到的係數則是利用參數搜尋法得到。

最後，我們採用AURORA 2.0 語料庫來驗證此方法之成效，從實驗結果發現，DECCR-MFCC 與DECCR-TECC 相對於MFCC 與TECC 之相對改善率皆有重大的改善。

關鍵詞： Teager 能量, 強健性語音辨識, gamma-tone 濾波器, 能量重刻, 資料導向

ABSTRACT

In noise-robust speech recognition, font-end feature extraction task is an important processing. For this reason, we propose data-driven energy cepstral coefficients rescaling (DECCR) to rescale energy feature coefficients. The aim of DECCR is to reduce the differences of clean and noisy energy feature coefficients. However, log energy and cepstral coefficients of the zeroth order are the most important coefficients for recognition performance. The following processing will focus on energy feature coefficients rescaling.

We use DECCR to rescale the output features of MFCC (Mel-frequency cepstral coefficients) and TECC (Teager energy cepstral coefficients). The DECCR method is divided into three parts, respectively, voice activity detection (VAD), piecewise log rescaling function and parameter searching algorithm. We use VAD to detect the voice and non-voice segments, and use piecewise log rescaling function to get different rescaling weights. The parameters of piecewise log rescaling function is obtained by parameter searching algorithm.

Finally, the output features are evaluated on the Aurora 2.0 database. From the results of DECCR-MFCC and DECCR-TECC, we found that the relative improvements over the baseline MFCC and TECC features are statistically significant.

Keyword: Teager energy, noise-robust speech recognition, gamma-tone filters, energy rescale, data-driven

Contents

List of Tables

List of Figures

Chapter 1

介紹

1.1 研究動機與目的

語音辨識最主要的目的是希望讓機器聽懂人類說話的聲音，進而操控機器執行相對應的工作。現今已經有許多語音辨識方面的應用，例如: 谷歌語音搜尋(google voice search)、讀寫機、聲控家電等等，但是在有限的語料所訓練出來的聲學模型中，要辨識龐大的使用人潮，辨識效能將其差無比。而造成效能低落的因素有語者差異、背景噪音等，我們暫且不論語者間的差異性，將目標鎖定爲背景噪音，我們可以發現在毫無噪音的環境下，可以快速並且準確地得到辨識結果，相對地，在充滿背景噪音的環境下，系統將無法準確辨識，甚至無法辨識，因此我們希望可以藉由觀察語音的特性，擷取出具有噪音強健性之特徵參數來增加辨識效能。

本論文中，我們以常見的 MFCC 與 [1]所提出之TECC 爲基準，結合 [2] 所提出之對數能量尺度重刻的方法，藉由能量爲語音辨識之重要指標的特性 [3, 4, 5, 6, 7]，調整能量參數(對數能量與 c_0序列)，目的是希望乾淨與含噪音特徵之能量參數在重刻後的差值最小化。然而，對數能量尺度重刻不論是在語音(speech)或非語音的部分(non-speech)皆使用相同的重刻尺度。我們針對這個部分，提出資料導向之能量特徵重刻法(Data-driven energy cepstral coefficients rescaling, DECCR)，此方法主要分爲語音動態偵測(Voice activity detection, VAD)、分段對數尺度函數(Piecewise log rescaling function) 與參數搜尋法(parameter searching algorithm)三個部分。我們利用語音動態偵測的方法偵測語音與非語音出現的時間點，並利用分段對數尺度函數對能量參數做不同尺度的重刻。此外，分段對數尺度函數所使用之參數是由參數搜尋法決定。由實驗

結果證實本論文所提出之方法，能有效減少語音能量特徵受到噪音干擾所造成的失真，進而提升辨識效果。

1.2 背景

現今自動語音辨識的技術已經相當成熟 [8]。但是在訓練與測試語料不匹配的情況下，辨識率將快速下滑。就背景噪音而言，是造成環境不匹配的主要因素，路人講話聲、汽車行駛的聲音...等等皆是不希望出現在錄製語料中的噪音，另一種噪音則是來自於錄製設備的差異。除此之外，尚有通道效應(Channel effects)、說話方式(Speaking styles)、語者差異(Speaker variations)...等影響因素。正因如此，噪音強健性長久以來被視爲一個重要的研究課題。

爲了改善上述之環境上的不匹配，目前已經有許多方法被提出，例如: [9, 10]。然而，依據方法的本質可分爲前端處理與後端聲學模型的調適。前端處理的部分又可分爲兩種類型:

1. **語音強化技術(Speech enhancement techniques)**
   語音強化技術 [11, 12]目的在於希望能夠藉由觀察含噪音之語音還原出乾淨語音訊號，以提升語音訊號本身的品質。常見的方法有維爾濾波器(Wiener filter) [13]、頻譜消去法(Spectral subtraction) [14]等。

2. **強健性語音特徵(Robust speech features)**
   強健性語音特徵的目的則是擷取出語音訊號中不受環境變化干擾而失真的強健性語音特徵參數。常見方法有倒頻譜平均消去法(Cepstral mean subtraction, CMS) [15]、倒頻譜正規化法(Cepstral mean and variance normalization, CVN)、頻譜熵值特徵(Spectral entropy feature) [16]等。

而後端處理的部分爲聲學模型的調適(Acoustic model adaptation)，主要藉由針對目標背景雜訊調整聲學模型，期望調適後的模型可以適用於新的環境。常見的方法有最大相似度線性迴歸法(Maximum likelihood linear regression, MLLR) [17]、最大事後機率法則(Maximum a posteriori, MAP) [18]。

本論文所提出之特徵擷取方法是屬於強健性語音特徵。我們以梅爾倒頻譜參數(Mel-frequency cepstral coefficients, MFCC)與 [1] 所提出之 Teager 能量倒頻譜參

數(Teager energy cepstral coefficients, TECC)爲基礎，結合能量特徵重刻的方法，縮短乾淨與含噪音語料之能量特徵值的差距。此方法使我們的辨識系統更具有噪音強健性。

1.3 論文架構

以下爲本論文的基本架構，第2章將介紹梅爾倒頻譜參數與 Teager 倒頻譜參數的特徵擷取方法，第3章爲說明對數能量尺度重刻與我們所提出的資料導向之能量特徵重刻法，第4章爲實驗，包括辨識系統設定、語料庫的介紹、效能評估方法以及實驗結果，最後爲結論與未來展望。

Chapter 2

特徵參數擷取

本章節主要針對實驗中所使用之特徵擷取方法做詳細說明。第2.1小節將說明梅爾倒頻譜參數 (Mel-frequency Cepstral Coefficients, MFCC) 的特徵擷取流程。第2.2小節則說明 Teager 能量倒頻譜參數 (Teager Energy Cepstrum Coefficients, TECC) 的實作方法。

2.1 梅爾倒頻譜參數

由於梅爾倒頻譜參數充分的考慮人耳在不同頻率的聽覺特性，成爲自動語音辨識中最常使用的特徵參數。梅爾倒頻譜參數的流程如圖2.1，其說明如下:

1. 音框化(Framing)：由於語音訊號是時變的訊號，我們無法以線性非時變的方法分析長時間(Long-Term)的語音訊號特徵。因此我們藉由音框化將其分割爲短時間(Short-Term)的訊號，使得語音訊號具備暫時穩定的特性。然而爲了避免相鄰兩音框的變化過大，我們會讓相鄰音框之間有重疊的區域。

2. 預強調(Pre-emphasis)：爲一高通濾波器(High-Pass Filter)，主要功用是加強聲波高頻的能量。人類說話的聲音受到聲帶及嘴唇的效應，產生的語音在高頻部分會有衰減的特性，透過預強調可以補償語音信號受到發音系統所壓抑的高頻部分，其如方程式2.1

$$s_{pe}[n] = s[n] - \alpha\, s[n-1], \alpha = 0.97 \tag{2.1}$$

其中 $s_{pe}[n]$ 爲輸出訊號，$s[n]$爲原始輸入訊號，α爲預強調的參數。

3. 漢明窗(hamming window)：每個音框根據固定時間點切割會造成音框邊緣出現訊號不連續的現象，此現象會造成音框經由快速傅立葉轉換後高頻部分產生雜訊。爲了降低雜訊的產生，我們將預強調後的音框做快速傅立葉轉換前會先乘上一個漢明窗，以增加音框左右兩端的連續性。如方程式2.2

$$s_{hw}[n] = w[n]s_{pe}[n], \tag{2.2}$$

其中

$$w[n] \triangleq 0.54 - 0.46\cos\left(2\pi\left(\frac{n+0.5}{N}\right)\right)$$

4. 快速傅立葉轉換(Fast Fourier Transform, FFT)：訊號的特性從時域上是很難找出特性，所以我們通常會將它轉換到頻域，觀察各個頻帶間能量的分佈，並藉由能量的分佈找出代表不同語音的特性。轉換方程式爲

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} s_{hw}[n]e^{-jn\omega} \tag{2.3}$$

5. 梅爾頻譜濾波器(Mel-frequency Filter Bank)：人耳對於頻率的變化在高頻與低頻時的敏感度不同，在低頻時人耳的感受會比較敏銳，此時對頻率變化的感受就會呈線性的。而當頻率變化位於高頻部分，人耳的感受就會越來越粗糙，當頻率大於1 KHz時，人耳對於頻率的感受就會呈現對數變化。梅爾頻率的目的即是模擬此種現象，方程式2.4爲梅爾頻率和一般頻率f 的關係式。

$$Mel(f) = 2595\log_{10}(1 + \frac{f}{700}) \tag{2.4}$$

三角濾波器部分，經由研究結果顯示人耳聽覺神經不只接受單一特定頻率的刺激，而是會受到一定範圍內頻率的影響，距離某特定頻率越遠其影響越小。所以在通過梅爾頻率的對數轉換後，必須再通過M個三角濾波器的處理，使得三角濾波器在梅爾頻率上平均分佈來模擬人耳的聽覺特性，而三角濾波器的公式如下

$$H_m[k] = \begin{cases} 0 & ,k < f[m-1] \\ \frac{k-f[m-1]}{f[m]-f[m-1]} & ,f[m-1] \le k < f[m] \\ \frac{f[m+1]-k}{f[m+1]-f[m]} & ,f[m] \le k \le f[m+1] \\ 0 & ,k > f[m+1] \end{cases} \tag{2.5}$$

其中$f[m]$ 爲第m個三角濾波器的中心點，$H_m[k]$爲頻率k 在第m個三角濾波器的權重(Weight)，N 爲音框大小。而$f[m]$則是由方程式2.6求得。

$$f[m] = Mel^{-1}\left(Mel(f_l) + m \times \frac{Mel(f_h) - Mel(f_l)}{M+1}\right), \tag{2.6}$$

其中f_l 爲M 個三角濾波器中最低的頻率，f_h 爲M 個三角濾波器中最高的頻率。本論文所使用之f_l、f_h與M的爲

$$f_l = 64, \quad f_h = 4000, \quad M = 23$$

6. 對數轉換(Logarithm)：音波振動透過空氣經由外耳與中耳藉由三小聽骨傳遞到後方的內耳，在傳遞的過程中，造成能量的損失，而能量最主要影響的將是人耳對於音量大小的解析度，因此我們透過對數運算對音量壓縮，除去語音訊號在相位(phase)上的變化。

7. 離散餘弦轉換(Discrete Cosine Transform, DCT)：在對數轉換後經由離散餘弦轉換的目的是希望將訊號轉換爲倒頻譜係數。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，增加辨識效能。方程式如2.7

$$C[n] = \sum_{m=1}^{M} \cos\left[\frac{\pi n\,(m-0.5)}{M}\right] S_{log}[m] \tag{2.7}$$

其中S_{log} 爲梅爾濾波器組中的資料取對數運算的輸出，$C[n]$爲MFCC 特徵向量，M 爲濾波器的個數。

2.2 Teager 能量倒頻譜參數

本小節主要說明 Teager 能量倒頻譜參數的擷取方法，流程如圖2.1。從圖中我們可以看出 Teager 能量倒頻譜參數與梅爾倒頻譜參數特徵擷取的實作步驟，最主要的差別在於 Teager 能量倒頻譜參數使用 gamma-tone filter (GTF) 取代梅爾倒頻譜參數所使用之三角濾波器來過濾每個頻帶間的能量，並且利用 Teager 能量評估法(Teager Energy Estimation)對通過濾波器之能量做進一步的估測，以得到更精確的能量值。以下我們將在第2.2.1小節與第2.2.2小節說明 gamma-tone 濾波器 與 Teager 能量評估法(Teager energy estimation)。

2.2.1 Gamma-tone 濾波器

在人類聽覺系統中，耳蝸是相當重要的器官，而基底膜(Basilar membrane)則是耳蝸接收聲音最重要的組織，對聲音訊號的振幅與頻率都有不同的響應。其功能就像一個帶通濾波器(Bandpass filter)，對於聲音的高頻，最大振幅會靠近基底膜的底部(base)；相反地，對於聲音的低頻，最大振福會出現在基底膜的頂部(apex)。Gammatone 濾波器的設計理念就是在模擬基底膜對於頻率選擇與頻譜分析的特性。而一個連續時間的 GTF 之脈衝響應(impluse response)如圖2.2所示，其方程式爲

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_c t + \phi), \tag{2.8}$$

其中 a 爲振幅(amplitude)，n 爲濾波器的階數，b 爲濾波器的帶寬(bandwith)，方程式爲

$$b = b_1 ERB(f_c) \tag{2.9}$$

其中 b_1 生理常數，$ERB(f_c)$ 爲等效矩形帶寬模型(Equivalent Rectangular Bandwidth, ERB)，會隨著中心頻率的改變而得到對應的帶寬來提高濾波器的效能，其式子爲

$$\begin{aligned} ERB(f_c) &= \frac{\int |G(\omega_c)|^2 d\omega}{|G(\omega_c)|^2} \\ &= 6.23(\frac{f_c}{1000})^2 + 93.39(\frac{f_c}{1000}) + 28.52 \end{aligned} \tag{2.10}$$

其中$G(\omega_c)$ 爲$g(t)$ 傅立葉轉換後的頻率響應(frequency response)，如圖2.3所示，而$|G(\omega_c)|$ 爲在中心頻率之帶通濾波器的最大振福。

在本論文中，我們設定

$$a = 1, \quad \phi = 0, \quad n = 4$$

本論文中gamma-tone 濾波器所使用之梅爾中心頻率的計算公式爲方程式2.6，其濾波器的數目、最高頻率與最低頻率的設定皆與梅爾倒頻譜參數中的三角濾波器之設定相同。

然而，連續時間GTF 函數是無法直接實作的，因此我們採用[19]提出之方法進行轉換。首先使用拉普拉斯轉換法(Laplace transform)將 $g(t)$ 轉換至 s 域(s domain爲連續域)，再將 s 域轉換至 z 域(離散域)，得到 $G(z)$ 爲

$$G(z) = \frac{\sum_{j=0}^{5} a_j Z^{-j}}{\sum_{i=0}^{9} b_i Z^{-i}} \tag{2.11}$$

最後利用z轉換求得離散時間 gamma-tone 濾波器之常係數線性差分方程(Linear Constant-Coefficient Difference Equation, LCCDE)進行實作。

2.2.2 Teager 能量評估法

Teager 能量評估法是一種非線性能量計算的方法，其主要目的在於增強語音訊號與噪音之間的能量差別。將語音中穩定或半穩定(語音部分)予以強化，並且使不穩定的訊號(雜訊部分)的能量值衰減。其連續時間的方程式表示爲

$$\psi\left[x(t)\right] = \left[\frac{d}{dt}x(t)\right]^2 - x(t)\left[\frac{d^2}{dt^2}x(t)\right] \tag{2.12}$$

將方程式2.12轉換爲離散型式，式子如下

$$\psi\left[x(n)\right] = \left[x(n)\right]^2 - x(n+1)x(n-1) \tag{2.13}$$

我們發現方程式2.13在處理一個音框前後兩端邊緣的取樣點會有超出邊界的問題，因此將式子修改爲

$$x[n] = \begin{cases} (x[n])^2 - x[n]\cdot x[n+1], & \text{if } n = 0 \\ (x[n])^2 - x[n]\cdot x[n-1], & \text{if } n = N-1 \\ (x[n])^2 - x[n+1]\cdot x[n-1], & \text{otherwise} \end{cases} \tag{2.14}$$

本論文中，我們將 gamma-tone 濾波器各個頻帶的能量皆使用 Teager 能量評估法降低噪音的能量，使我們所擷取的特徵參數能夠更加具有噪音強健性。

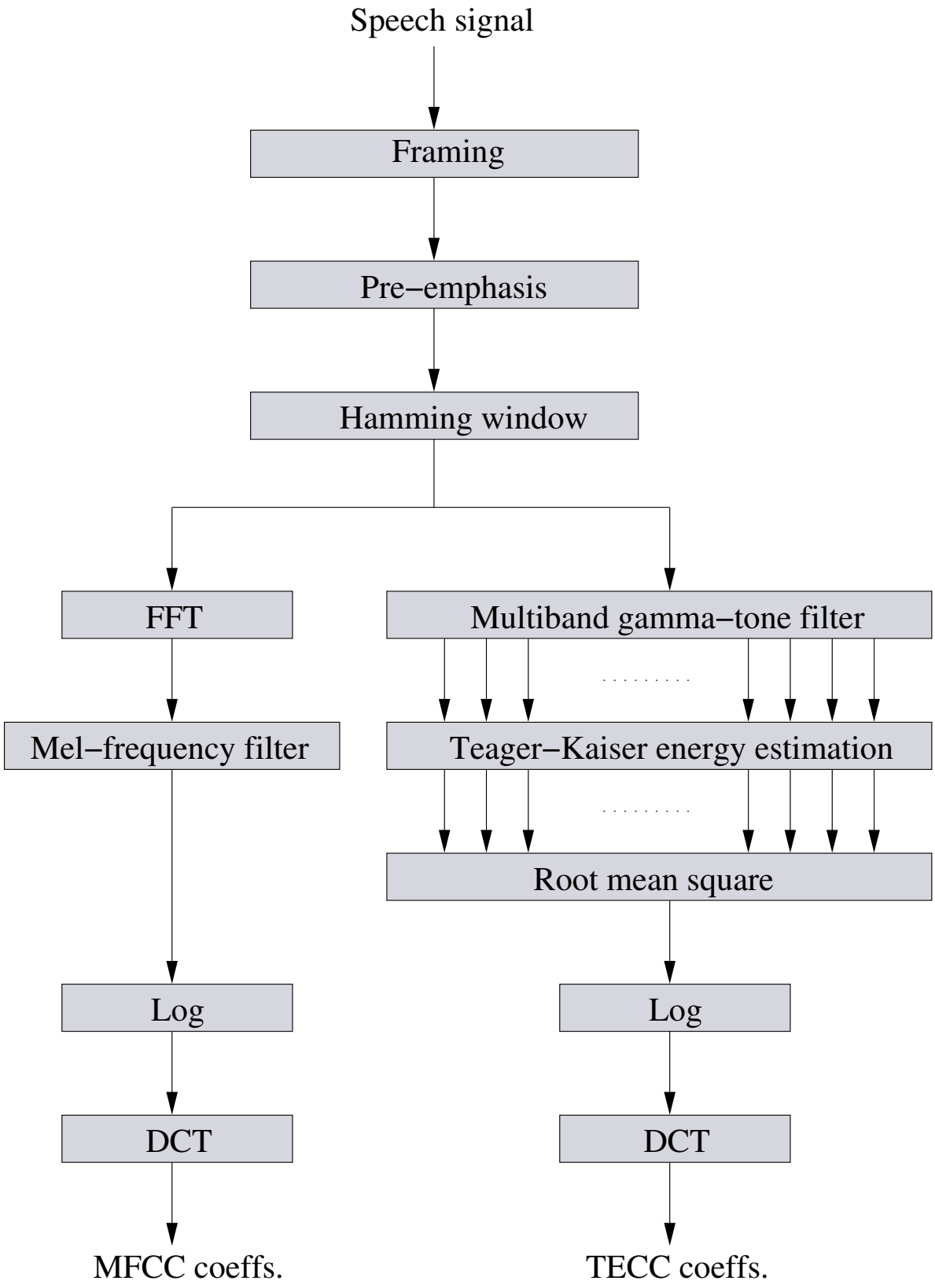

圖 2.1: MFCC 與 TECC 特徵擷取之訊號處理流程

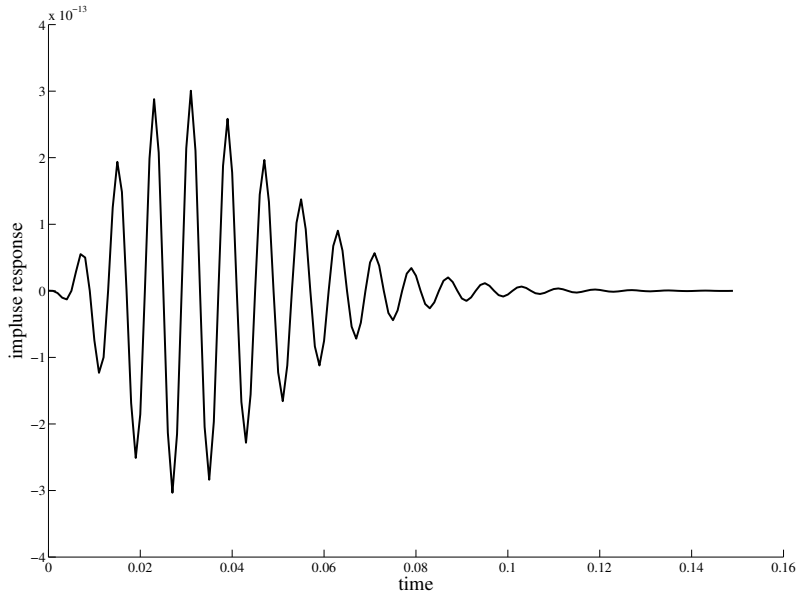

圖 2.2: Gamma-tone 濾波器之脈衝響應(中心頻率爲 1000 Hz)

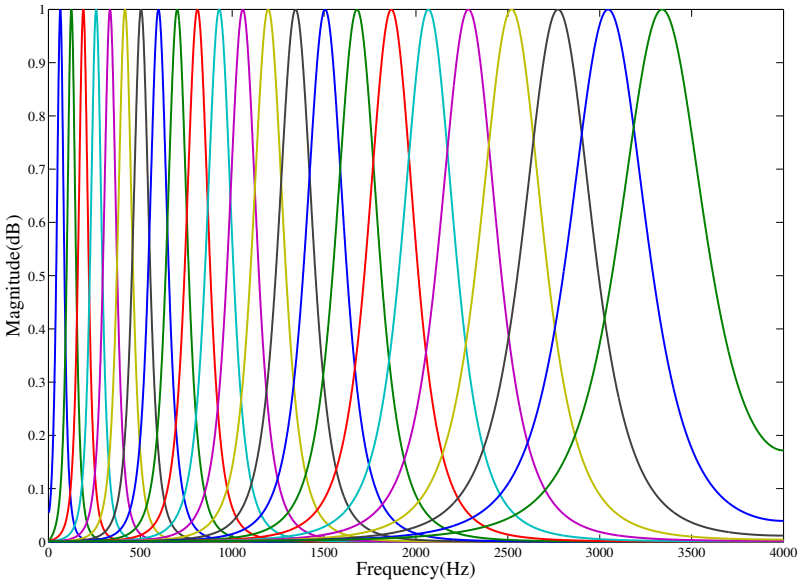

圖 2.3: Gamma-tone 濾波器之頻率響應

Chapter 3

能量特徵重刻

本章節將語音訊號利用第2章節所介紹的訊號處理方式，擷取出梅爾倒頻譜參數與 Teager 能量倒頻譜參數，對其對數能量(Log energy)或第零維的倒頻譜係數(c_0)進行重刻，流程如圖3.1。能量特徵參數重刻的主要目的是希望能夠使得含噪音語料與乾淨語料之特徵參數的 c_0 能夠盡可能的相近，以提高辨識的正確率。我們以 [2]所提出之對數能量尺度重刻法(Log Energy Rescaling, LER)爲基礎，做進一步的改善。資料導向之能量特徵重刻法(Data-driven energy cepstral coefficients rescaling, DECCR) 爲我們所提出之新方法。以下我們分別在第3.1與3.2小節詳細說明對數能量尺度重刻法與資料導向之能量特徵重刻法。

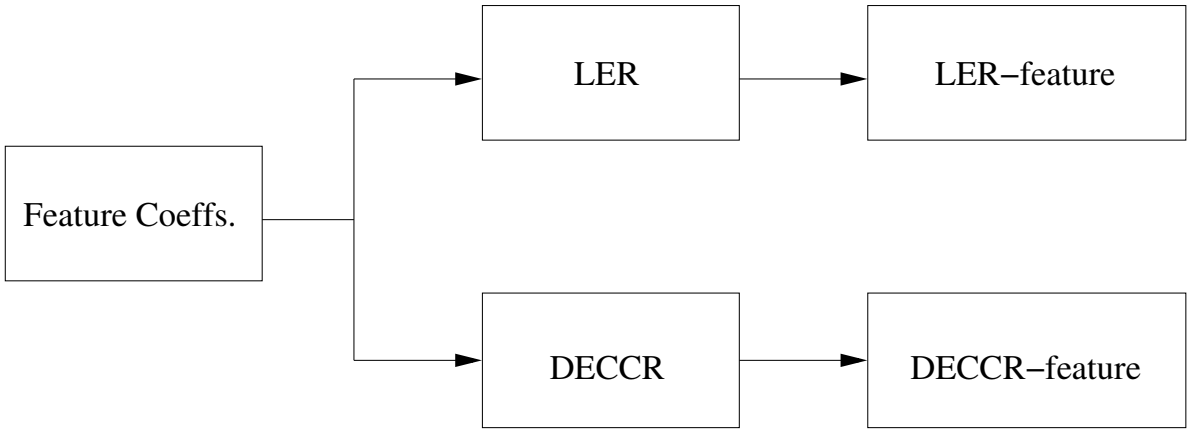

圖 3.1: 重刻特徵參數流程圖

3.1 對數能量尺度重刻法

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；反之若非語音出現期對數能量特徵值則會接近於零。[2]基於此現象提出語音能量特徵正規化方法-對數能量尺度重刻法，以使用對數轉換函數方式對語音對數能量作重刻。對數能量尺度重刻法之步驟說明如下，首先，找出每一語句所有音框 i 最大能量值 E_{max} 以及最小能量值 E_{min} ，將此一範圍分成 M 等份，而每個音框 i 對數能量所落至分位差 m_i 的索引爲

$$m_i = \left\lfloor \frac{E_i - E_{min}}{E_{max} - E_{min}} \times M \right\rfloor, \tag{3.1}$$

其中 E_i 爲第 i 個音框的對數能量，我們將 m_i 對數轉換函數(如圖3.2)

$$W(m_i) = \frac{log(m_i)}{log(M)} \tag{3.2}$$

得到能量重刻之權重 $W(m_i)$ ，因此重刻後的能量表示爲

$$\widehat{E_i} = E_i \times W(m_i) \tag{3.3}$$

經由對數能量尺度重刻所得到新的特徵參數，我們稱爲LER-freature。例如:將梅爾倒頻譜參數做對數能量尺度重刻所得到的特徵稱之爲LER-MFCC。

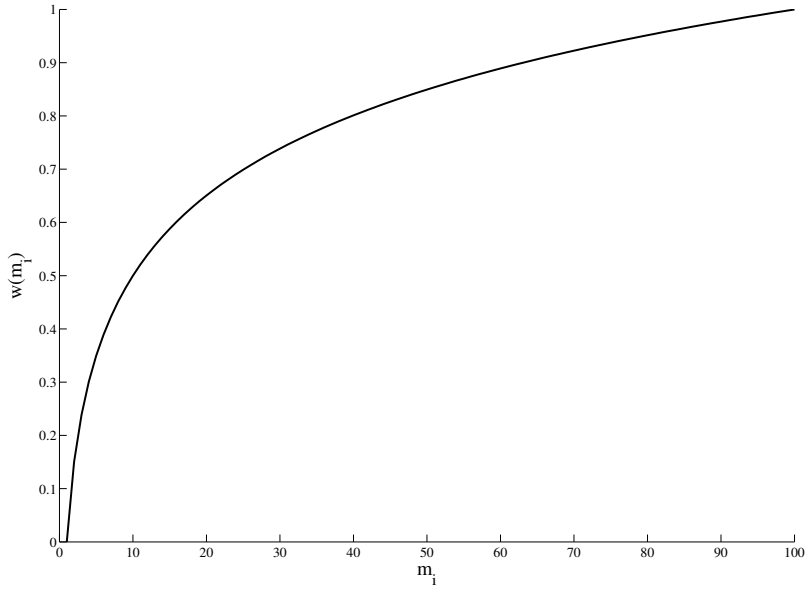

圖 3.2: 對數轉換函數

3.2 資料導向之能量特徵重刻方法

能量特徵重刻在強健性語音辨識的領域中是屬於特徵補償的一種。主要的目的是希望透過重刻特徵參數減少乾淨與含噪音特徵兩者的差異性，現今已有許多相關的方法被提出，例如: 一般常用的統計圖等化法(histogram equalization) 與方差正規化(variance normalization) 都可視爲特徵重刻的變形。由此可見，語音特徵的能量對辨識來說是很重要的因素。然而過去也已經有對數能量特徵重刻的方法被提出 [2]，其基礎的想法爲一段含噪音的語音有語音的段落，能量值偏高，而非語音(non-speech) 的部分，能量值偏低。本論文繼承了這個概念，並提出資料導向之能量特徵重刻方法對特徵參數做進一步的處理。

此方法主要分爲語音動態偵測(Voice activity detection, VAD)、分段對數尺度函數(Piecewise log rescaling function) 與參數搜尋法(parameter searching algorithm)三個部分。我們利用[20]所提出的 VAD 來偵測語音與非語音出現的時間點，並利用分段對數尺度函數對能量參數做不同尺度的重刻。其分段對數尺度函數所使用之參數是由參數搜尋法決定。以下我們就這三個部分做詳細說明。

3.2.1 語音動態偵測

在語音訊號處理系統中，語音訊號常常受到環境噪音的影響，使得系統效能低落。因此發展了語音動態偵測(voice activity detection, VAD)來偵測訊號中語音與非語音的位置。我們觀察語音在各個頻帶間能量的變化，發現無論任何種類的噪音，在頻帶$[0, 50Hz]$之間都有相當比例的能量。藉由此特性，我們計算每個音框在此低頻帶的頻譜能量，並根據此能量值，判斷該語音音框是否爲純噪音音框或是含語音的音框。其做法詳細說明如下，首先假設我們有一段語句(utterance) u，對每個音框 i 取離散傅立葉轉換(Discrete Fourier Transform, DFT)，式子爲

$$X^{(i)}(f_k) = X^{(i)}[k] = \sum_{n=0}^{N-1} x_i[n] e^{-j\frac{2\pi kn}{K}}, 0 \le k \le K-1 \tag{3.4}$$

其中 f_k 爲頻率，其值爲

$$f_k = \frac{F_s}{2K} k \tag{3.5}$$

其中 F_s 爲取樣頻率， K 爲離散傅立葉的點數。在此

$$F_s = 8000, \quad K = 256$$

我們利用方程式3.4可以得到每個頻率 f_k 的能量值，因此我們定義出頻帶$[0, 50Hz]$之頻譜能量計算的方式爲

$$Y_{[F_L,F_U]}^{(i)} = \sum_{F_L \le f_k \le F_U} \mid X_m(f_k) \mid \tag{3.6}$$

根據方程式3.6，我們計算每個音框之低頻帶頻譜強度，即 0 至 50 Hz以內的低頻帶頻譜強度如下

$$Y_{[0,50]}^{(i)} = \sum_{0 \le f_k \le 50} \mid X_m(f_k) \mid \tag{3.7}$$

接著以一段語音前 P 個音框之低頻帶頻譜強度的平均爲門檻值，其計算公式如下

$$\theta = \frac{1}{P} \sum_{i=0}^{P-1} Y_{[0,50]}^{(i)} \tag{3.8}$$

其中 P 爲純噪音的音框數，經由實驗結果將 P 設定爲 6 時效果最好。我們將每個音框低頻帶內的頻譜能量 $Y_{[0,50]}^{(i)}$ 與門檻值 $theta$ 做比較，若 $Y_{[0,50]}^{(i)} \le \theta$ 則將其歸類爲非語音音框，反之，則屬於語音音框。判斷式如下:

$$\text{第}i\text{個音框} = \begin{cases} Y_{[0,50]}^{(i)} \le \theta, & \text{非語音音框} \\ Y_{[0,50]}^{(i)} > \theta, & \text{語音音框} \end{cases} \tag{3.9}$$

本論文使用[20]所提出的方法來偵測語音與非語音出現的時間點，使得我們所提出之能量重刻的方法可以更加精確。而語音動態偵測的正確性，我們將強制對齊(force alignment)的方法實作在Aurora 2.0語料庫，其實驗結果爲語音與非語音訊號出現的時間點，我們將此結果是爲語音動態偵測的參考答案。接著與上述低頻帶能量之語音動態偵測器所得到的結果進行比對，正確率如下表

表 3.1: 低頻譜能量之語音動態偵測器的準確度

Train	Set A	Set B	Set C
73.52	69.38	67.69	69.29

3.2.2 分段對數尺度函數

分段對數尺度函數是爲了使含語音的能量能夠更確實的保留原有的能量值，而非語音的能量能夠大幅度的下降，以增加語音與非語音能量的差異性。因此我們結合第3.2.1小節所介紹的語音動態偵測方法對能量就不同尺度的重刻。首先，假設我們有一段語句(utterance) u 的特徵向量，其處理過程如下

- 從每一語句 u 的 c_0 序列中找出最大 c_0 值 M_u 與最小 c_0 值 m_u。
- 考慮每個音框 i 之特徵值 $c_0[i]$，定義 $r[i]$ 爲

$$r[i] = \frac{c_0[i] - m_u}{M_u - m_u}$$

很明顯地，我們會得到

$$0 \leq r[i] \leq 1$$

- 進行重刻運算後的特徵爲

$$\tilde{c}_0[i] = w[i]c_0[i] \tag{3.10}$$

其中$w[i]$ 爲音框i 之權重。假設語句中的噪音屬於穩態的(quasi-stationary)，高能量的段落包含語音的可能性會相當高。因此，這意指

$$\begin{aligned} r[i] \approx 1 \quad &\longrightarrow \quad w[i] \approx 1, \\ r[i] \approx 0 \quad &\longrightarrow \quad w[i] \approx 0. \end{aligned} \tag{3.11}$$

已經有很多方法實作式子 3.11的想法。我們提出分段對數尺度函數實現上述的想法，函數爲

$$w[i] = \begin{cases} \left[\frac{\log(r[i]\times M)}{\log(M)}\right]^{\alpha_1}, & Y^{(i)}_{[0,50]} \leq \theta \\ \left[\frac{\log(r[i]\times M)}{\log(M)}\right]^{\alpha_2}, & Y^{(i)}_{[0,50]} > \theta \end{cases} \tag{3.12}$$

其中 $Y^{(i)}_{[0,50]}$ 爲 0 至 50 Hz頻譜的能量，$theta$ 爲語音與非語音的門檻值。式子 3.12中的參數 M 由經驗法則將其設定爲100，而 α_1 及 α_2 是經由最小化平行語料庫之訓練集(parallel training data sets) 整體的失眞決定，我們採用第 3.2.3節的參數搜尋法取得最佳的參數，其函數如圖。除訓練資料使用此重刻參數外，測試資料亦使用此參數進行重刻。

我們以對數能量尺度重刻與資料導向之能量特徵重刻法重刻第2章所提出之梅爾倒頻譜參數與 Teager 能量倒頻譜參數，其比較圖分別爲圖 3.4與圖 3.5。從圖中，我們可以很清楚地看出梅爾倒頻譜參數與 Teager 能量倒頻譜參數再重刻運算後，明顯地減少了乾淨與含噪音語句之差異。而圖中，梅爾倒頻譜參數與 Teager 能量倒頻譜參數由參數搜尋法得到之重刻參數 α_1 與 α_2 爲

$$\alpha_1 = 1.3, \quad \alpha_2 = 1.0$$

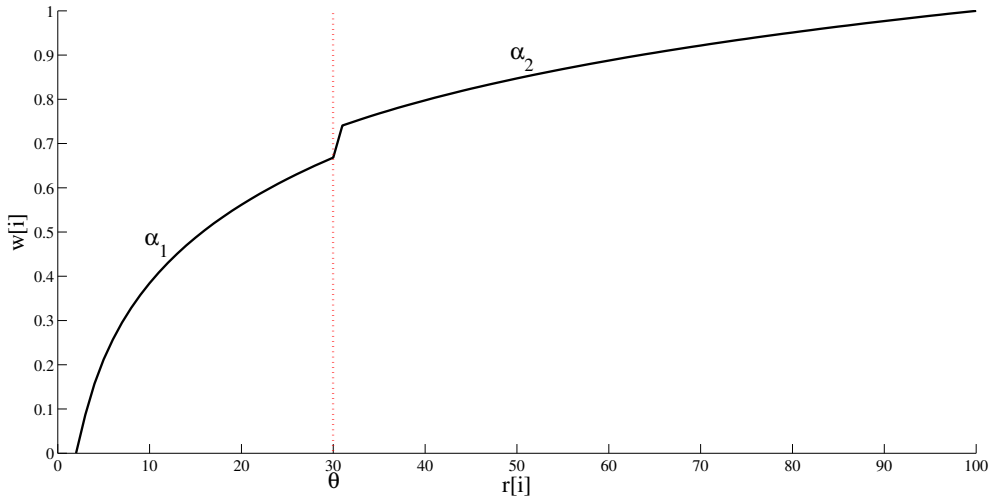

圖 3.3: 分段對數尺度函數

3.2.3 參數搜尋法

本小節所提出的參數搜尋法(parameter searching algorithm)，目的爲尋找第3.2.2小節所使用之參數 α_1 及 α_2，使得乾淨與含噪音語句之特徵值的差距最小化。而此方法必須滿足

$$1 \leq \alpha_2 < \alpha_1 \leq 2,$$

我們希望非語音的特徵值可以快速下降，所以令 $\alpha_1 > \alpha_2$ ，使低於門檻值之特徵值下降速率提高。演算法 1爲參數搜尋法的虛擬程式碼(pseudo code)。程式碼的搜尋法則爲將所有訓練語料重刻後的乾淨與相對應之含噪音之特徵值差值相加，而 α_1 與 α_2 的值每次皆增加 0.1，我們觀察所有的案例 (case)後，選出差值總和最小值時的α_1 及 α_2 爲最佳參數。

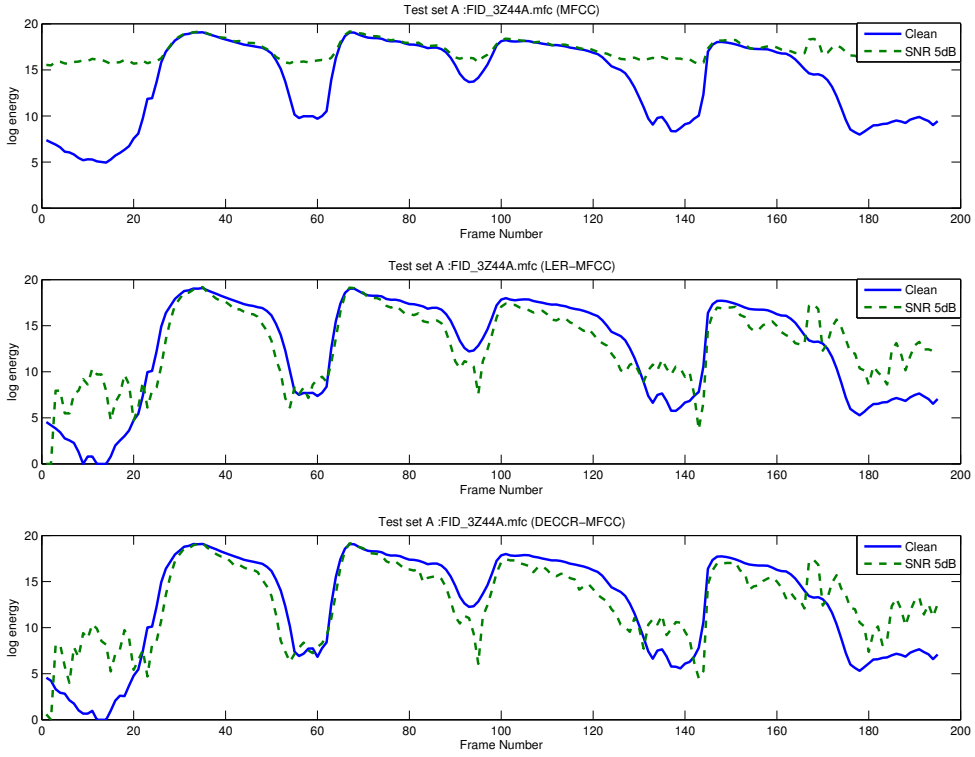

圖 3.4: 一對平行語句之MFCC(上)、LER-MFCC(中) 與DECCR-MFCC(下) 對數能量序列的比較，語句的ID 爲FID_3ZZ4A.08。

```
for α1 = 1.1; α1 <= 2; α1+ = 0.1 do
    for α2 = 1; α2 < α1; α2+ = 0.1 do
        dist = 0.0;
        for u = 1;u ≤ NumofUtt; u++ do
            rc0_u^clean = Rescale c0 of clean feature for utterance u;
            rc0_u^multi = Rescale c0 of multi feature for utterance u;
            dist += | rc0_u^clean − rc0_u^multi |;
        end
        if(min(dist)) Record dist,α1,α2;
    end
end
```

Algorithm 1: 參數搜尋法之虛擬碼。圖中 NumofUtt 爲乾淨與含噪音之訓練語料的對應組數目。

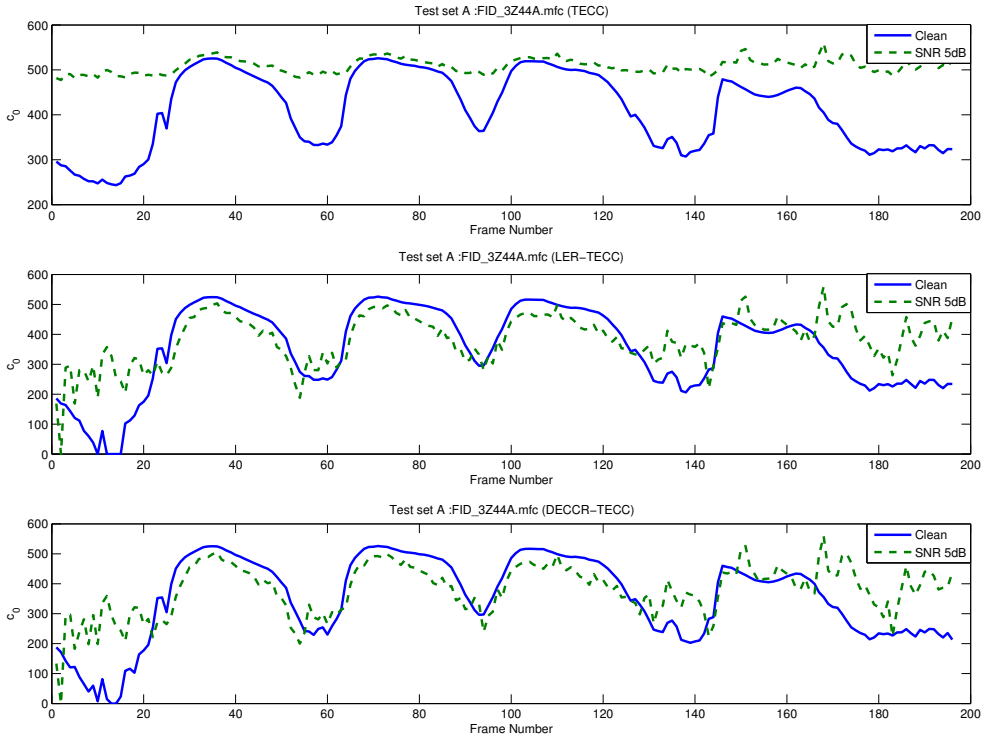

圖 3.5: 一對平行語句之TECC(上)、LER-TECC(中) 與DECCR-TECC(下) c_0 特徵序列的比較，語句的ID 爲FID_3ZZ4A.08。

Chapter 4

實驗

4.1 辨識系統設定

本實驗分別以梅爾倒頻譜參數與 Teager 能量倒頻譜參數爲基準(baseline)，對其對數能量(Log energy, LE)與倒頻譜參數 c_0 序列使用均值消去法(mean subtraction, MS)、均值正規化法(mean and variance normalization, MV)、對數能量尺度重刻法以及資料導向之能量特徵重刻方法做進一步處理。而梅爾倒頻譜參數之特徵向量是由 $c_1, ..., c_{12}$ 及對數能量所組成，其特徵擷取工具爲 Aurora Frontend WI007。Teager 能量倒頻譜參數之特徵向量則由 $c_1, ..., c_{12}$ 及 c_0 所組成，其特徵參數使用第2.2小節的方法得到。在訓練與測試階段時，再加入 velocity (delta) 及 acceleration (delta-delta) 的特徵。而特徵擷取所使用之音框長度 (frame length) 爲 25 ms，音框間距 (frame shift) 爲10 ms，濾波器頻帶的數目爲 23。

後端的聲學模型是使用標準的Aurora 評估方式 [21]。其中數字部分由 1 ∼ 9 及 zero、oh 所組成，每個數字模型包含 16 個狀態(state)，每個狀態是利用 3 個高斯混合分布(Gaussian Mixture Distribution) 表示。另外靜音(silence) 模型包含 3 個狀態，而間歇(short-pause) 模型包含一個狀態，此兩種模型使用六個高斯混合分布表示。

4.2 實驗語料

本論文所用之語料庫爲歐洲電信標準協會(European Telecommunication Standard Institute, ETSI)發行的Aurora 2.0 語料庫。Aurora 2.0 語料庫爲連續英文數字語料，內容

是以美國成年男女所錄製的乾淨環境連續數字，再加上噪音與通道效應，被廣泛的使用在噪音強健性的評估上。其噪音類型可分爲加成性噪音 (additive noise) 與摺積性噪音 (convolutional noise)，加成性噪音分別在地下鐵(subway)、細語(babble)、汽車(car)、宴會(exhibition)、餐廳(resturant) 、街道(street)、機場(airport)與火車站(train station)等八種不同場所錄製而成。而摺積性噪音則是從以上八種噪音中取出兩種，並依每5 dB 爲區間之不同訊噪比 (signalto-noise ration, SNR) 及乾淨語料，分爲七種不同的噪音程度。訓練語料分爲乾淨與含噪音(multi)兩種訓練集，這兩種訓練集語料的個數皆爲8440句，而測試語料每種噪音各1001句。

4.3 效能評估方法

Aurora 2 的實驗結果分爲乾淨(clean-train)和含噪音(multi-train)之訓練集合，測試資料集(test set)則是使用訊噪比(signal-to-noise ratio, SNR) $0-20$dB 的語料。實驗結果如表 4.1、 4.2、 4.3、 4.4 所示。Aurora 2 的測試資料分爲testa、testb 與testc，其測試語料的數目比例爲2:2:1。表中平均 (Avg) 之欄位是測試資料集依照測試語料(test data)數目的比例計算得到，Aurora 2 之計算式子爲

$$\textbf{Avg} = \frac{\textbf{Set A} * 2 + \textbf{Set B} * 2 + \textbf{Set C} * 1}{5}, \tag{4.1}$$

此外，欄位 $rimp$ 是先利用4.1計算各個特徵的平均，再分別以 MFCC 與 TECC 的詞錯率爲基準 (baseline) 計算相對改善率 (relative improvement)，式子如下

$$rimp = \frac{S_c - S_b}{100 - S_b} \times 100\% \tag{4.2}$$

其中 S_c 是我們想要與基準比較之平均值，S_b 爲對應的基準。

4.4 實驗結果

本小節將會對強健性語音的辨識實驗結果做說明。我們將實驗區分爲clean-train與multi-train 兩種利用不同訓練資料集所得到的聲學模型，藉此分析能量參數對語音辨識上的影響。實驗結果證明本論文所提出之方法可以有效地提升噪音環境下語音的辨識率，降低噪音對語音的干擾。其分析如下：

1. Clean-train之實驗結果我們以 MFCC 與 TECC 爲基準，對特徵參數使用 MS、MV、LER 及 DECCR 等方法重刻特徵值。從表4.1與表4.2的實驗結果，我們發現 MFCC 與 TECC 經過 MS 後，雖然其語音訊號只做平移，卻成功的提升 19.30% 與 29.34% ，這證明了音檔前後非語音音框的能量對辨識率的影響甚大，而 MV、 LER 與 DECCR 皆對語音與非語音音框之能量做進一步的處理，使乾淨與含噪音之能量參數值能夠盡可能地減少差距。從實驗結果證實 MV、 LER 與 DECCR 的方法大幅度的突破 MS 的辨識率，其中由 DECCR 方法重刻之特徵參數效果最好。此外，我們綜合比較 DECCR-MFCC 與 DECCR-TECC 之辨識結果， DECCR-TECC 的結果明顯的優於 DECCR-MFCC ，因此證實了 DECCR 的方法應用在 TECC 上效果非常卓越。

2. Multi-train之實驗結果Multi-train 的實驗設定皆與clean-train 相同，其主要差別爲訓練語料庫之音檔包含了含噪音的訓練語料。我們比較表4.1、4.2、4.3與表4.4的實驗結果，可以很明顯地看出訓練聲學模型時，加入了含噪音的訓練語料，使聲學模型更具有噪音強健性。然而在這樣的訓練環境下，特徵參數的擷取方式依然存在極大的影響力，其中以 MV 的效果最好。

綜合上述結果，雖然 DECCR 的辨識結果在 multi-train 中略低於 MS 與 MV ，但是其實驗結果與MFCC、TECC相比，相對改善率有極大幅度的提升，因此我們可以證實 DECCR 具有非常傑出的噪音強健性。

表 4.1: Aurora 2.0 之詞辨識率。聲學模型訓練的語料爲乾淨語料。實驗結果爲訊噪比0-20 dB 之平均。Avg 爲Set A、Set B 及Set C之平均，rimp爲相對改善率，以 MFCC 爲基準，利用方程式4.2計算得到。

Feature	Set A	Set B	Set C	Avg.	Rel. imp.
MFCC	61.34	55.75	66.14	60.06	-
MFCC+MS	66.18	70.81	64.88	67.77	19.30
MFCC+MV	70.18	70.77	66.37	69.65	24.01
LER-MFCC	74.60	74.51	65.23	72.69	31.62
DECCR-MFCC	75.52	75.58	65.77	73.59	33.88

表 4.2: Aurora 2.0 之詞辨識率。聲學模型訓練的語料爲乾淨語料。實驗結果爲訊噪比0-20 dB 之平均。Avg 爲Set A、Set B 及Set C之平均，rimp爲相對改善率，以 TECC 爲基準，利用方程式4.2計算得到。

Feature	Set A	Set B	Set C	Avg.	Rel. imp.
TECC	55.55	51.79	65.30	56.00	-
TECC+MS	66.92	71.52	67.67	68.91	29.34
TECC+MV	74.91	75.38	76.03	75.32	43.91
LER-TECC	77.36	77.40	66.97	75.30	43.86
DECCR-TECC	79.09	80.19	72.15	78.15	50.34

表 4.3: Aurora 2.0 之詞辨識率。聲學模型訓練的語料爲含噪音之語料。實驗結果爲訊噪比0-20 dB 之平均。Avg 爲Set A、Set B 及Set C之平均，rimp爲相對改善率，以 MFCC 爲基準，利用方程式4.2計算得到。

Feature	Set A	Set B	Set C	Avg.	Rel. imp.
MFCC	87.82	86.27	83.78	86.39	-
MFCC+MS	88.72	87.79	87.25	88.06	12.27
MFCC+MV	89.67	88.07	86.10	88.32	14.18
LER-MFCC	89.36	86.54	85.51	87.46	7.86
DECCR-MFCC	89.39	87.22	85.80	87.80	10.36

表 4.4: Aurora 2.0 之詞辨識率。聲學模型訓練的語料爲含噪音之語料。實驗結果爲訊噪比0-20 dB 之平均。Avg 爲Set A、Set B 及Set C之平均，rimp爲相對改善率，以 TECC 爲基準，利用方程式4.2計算得到。

Feature	Set A	Set B	Set C	Avg.	Rel. imp.
TECC	88.07	87.09	85.74	87.21	-
TECC+MS	89.14	89.33	89.66	89.32	16.50
TECC+MV	90.71	90.34	89.96	90.41	25.02
LER-TECC	90.04	87.61	86.38	88.33	8.76
DECCR-TECC	90.02	89.10	88.10	89.27	16.11

Chapter 5

結論與未來展望

5.1 結論

自動語音辨識系統中，前端特徵擷取的方式對辨識率的影響相當大，然而特徵參數中最重要的參數爲對數能量與倒頻譜參數 c_0 。此兩種參數爲語音訊號在整個各個頻帶間的整體表現，因此極具重要性。在本論文中，我們針對此兩種能量參數提出了資料導向之能量特徵重刻法，並以梅爾倒頻譜參數與 Teager 能量倒頻譜參數爲基準進行重刻。此方法藉由語音動態偵測器偵測出語音與非語音出現之段落，再利用分段對數尺度函數給予不同的權重，重新計算其能量值，以補償語音在噪音環境下的失眞。而分段對數尺度函數的設計理念來自於含噪音之語音其能量特徵值相對較高，非語音處之能量特徵值普遍偏低的特性，希望增加語音與非語音之能量特徵值的差異性，故給予不同尺度的權重。其函數所使用的參數則是由參數搜尋法自動決定。最後，我們採用Aurora 2 語料庫探討此方法在含噪音的環境下是否可以成功達到補償的效果，並且與其他常用的特徵處理方法比較。從實驗結果我們可以看出資料導向之能量特徵重刻法不論是在 clean-train 或 multi-train 中，皆有非常卓越的表現，這也證實了能量參數重刻對於噪音強健性的影響非常大。

5.2 未來展望

資料導向之能量特徵重刻法的實驗結果雖然有大幅度的改善，但是仍然有進步的空間。而其改善的方向可以語音動態偵測爲主要目標，其主要原因爲本論文所使用之語

音動態偵測法只考慮到低頻譜的能量強度，並未考慮到高頻的能量值。因此大部分的能量皆會被誤判爲語音的能量，如此一來，能量重刻所使用的下降尺度將會比較小，而導致噪音的能量值下降幅度不夠，無法鑑別語音與非語音的能量值。但是從另一個角度來看，假設語音的能量被誤判爲非語音的能量比例較高，將導致語音的能量降低過度，失去原有的特性。所以一個準確的語音動態偵測器對本論文的方法是相當重要的，若能提高其準確度，相對的，辨識效能會有更進一步的突破。

Bibliography

[1] D. Dimitriadis, P. Maragos, and A. Potamianos, "On the Effects of Filterbank Design and Energy Computation on Robust Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1504–1516, 2011.

[2] 陳鴻彬, "On the Study of Energy-Based Speech Feature Normalization and Application to Voice Activity Detection," 國立臺灣師範大學碩士論文, 2007.

[3] T. H. Hwang, "Energy contour extraction for in-car speech recognition," in *proceedings of 8th European Conference on Speech Communication and Technology(EUROSPEECH2003)*, 2003.

[4] W. Zhu and D. O'Shaughnessy, "Log-energy dynamic range normalization for robust speech recognition," in *proceedings of 2005 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, vol. 1, pp. 245–249, 2005.

[5] T. H. Hwang and S. C. Chang, "Energy contour enhancement for noisy speech recognition," in *proceedings of 4th International Symposium on Chinese Spoken Language Processing(ISCSLP2004)*, pp. 249 – 252, 2004.

[6] S. M. Ahadi, H. Sheikhzadeh, R. L. Brennan, and G. Freeman, "An energy normalization scheme for improved robustness in speech recognition," in *proceedings of 8th International Conference on Spoken Language Processing(ICSLP2004)*, 2004.

[7] R. Chengalvarayan, "Robust energy normalization using speech/nonspeech discriminator for German connected digit recognition.," in *proceedings of 6th European Conference on Speech Communication and Technology(EUROSPEECH 1999)*, 1999.

[8] P. Krishnamoorthy and S. R. M. Prasanna, "Enhancement of noisy speech by temporal and spectral processing," *The 38th International Speech Communication Association (ISCA)*, vol. 53, pp. 154–174, feb 2011.

[9] C. Garreton, N. B. Yoma, and M. Torres, "Channel Robust Feature Transformation Based on Filter-Bank Energy Filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1082 –1086, 2010.

[10] X. Huang, "Minimizing speaker variation effects for speaker-independent speech recognition," in *proceedings of the workshop on Speech and Natural Language*, pp. 191–196, 1992.

[11] D. Y. Zhao and W. B. Kleijn, "HMM-Based Gain Modeling for Enhancement of Speech in Noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 882 –892, 2007.

[12] J. Ming, R. Srinivasan, and D. Crookes, "A Corpus-Based Approach to Speech Enhancement From Nonstationary Noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 822 –836, May 2011.

[13] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen, "A combined multichannel Wiener filter-based noise reduction and dynamic range compression in hearing aids," *Signal Processing*, vol. 92, no. 2, pp. 417 – 426, 2012.

[14] Y. Denda, T. Nishiura, H. Kawahara, and T. Irino, "Speech recognition with wavelet spectral subtraction in real noisy environment," in *proceedings of 7th International Conference on Signal Processing (ICSP 2004)*, vol. 1, pp. 638–641, 2004.

[15] H. Veisi and H. Sameti, "The integration of principal component analysis and cepstral mean subtraction in parallel model combination for robust speech recognition," *The 17th International Conference on Digital Signal Processing(DSP2011)*, vol. 21, no. 1, pp. 36 – 53, 2011.

[16] H. Misra, S. Ikbal, S. Sivadas, and H. Bourlard, “Multi-resolution spectral entropy feature for robust ASR,” in *proceedings of 2005 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, vol. 1, pp. 253–256, 2005.

[17] P. Raghavan, R. Renomeron, C. Che, D.-S. Yuk, and J. Flanagan, “Speech recognition in a reverberant environment using matched filter array (MFA) processing and linguistic-tree maximum likelihood linear regression (LT-MLLR) adaptation,” in *proceedings of 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, vol. 2, pp. 777 –780 vol.2, mar 1999.

[18] J. L. Gauvain and C. H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 2, pp. 291 –298, apr 1994.

[19] M. Slaney, “An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank,” *Apple Computer Perception Group Tech Rep*, no. 35, 1993.

[20] 杜文祥, ”Study on the Voice Activity Detection Techniques for Robust Speech Feature Extraction,” 國立暨南國際大學碩士論文, 2007.

[21] D. Pearce and H. G. Hirsch, “The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *proceedings of 6th International Conference on Spoken Language Processing(ICSLP2000)*, September 2000.