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

資料驅動能量特徵調整於雜訊性語音辨識

Data-driven Rescaling of Energy Features 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

中華民國一０一年七月

July 2012

Acknowledgments

時光飛逝，在中山資訊工程研究所的生活已接近尾聲。兩年來的研究所生涯，用無數的歡笑與淚水交織纏繞出的甜蜜回憶，將深刻地烙印於我心深處。兩年的研究所求學期間，首先要感謝指導教授陳嘉平老師的栽培。於學術研究與論文撰寫方面，感謝老師孜孜不倦的指導，讓學生學習到許多寫作的技巧。於生活心態方面，老師總是秉持著不畏艱難與持之以恆的學習態度，使我對自己的人生有了新的啓發與收穫。於身體健康方面，感謝老師創立體育課的活動，適當的運動可以促進身體健康，使頭腦更聰明。於此，特致上由衷的謝意。

接著我要感謝秉豐、志宇和予宏，陪伴我度過研究所生涯，彼此互相討論與鼓勵，使得我的生活過得一點也不孤單。以及要感謝銘冠、宗芃、又壬等學弟，因爲有你們的陪伴，使得我的研究所生活更多采多姿，有你們的幫忙，才可以讓我的論文研究更爲順利與完善。這兩年在實驗室的相處，不論是在課業、研究或生活上都得到許多扶持。

最後還要感謝我的家人，感謝你們一直以來的關心與支持，讓我能夠專注於課業與研究。我將這份喜悅與您們分享！

摘要

本論文主要探討能量特徵重刻技術對雜訊性語音辨識的影響。語音辨識系統常會受到環境雜訊的影響而導致辨識效能低落，使得語音強健性技術長久以來被視爲一個非常重要的研究課題。然而過去有不少研究指出語音能量特徵對於雜訊環境下的語音辨識影響甚鉅，因此我們提出資料驅動能量特徵重刻法 (Data-driven energy features rescaling, DEFR) 對能量特徵作進一步的調整。此方法分爲語音活動偵測、分段對數尺度函數以及參數搜尋法三個部分。目的是希望能夠減少雜訊與乾淨語音特徵值的差異性。我們將此方法應用在梅爾倒頻譜參數與 Teager 能量倒頻譜參數上，並且和均值消去法與均值正規化法作比較。我們採用 Aurora 2.0 與 Aurora 3.0 語料庫來驗證此方法之成效，由實驗結果證實本論文所提出之方法，能夠有效地提升辨識率。

關鍵詞： Teager 能量，強健性語音辨識，能量重刻，資料驅動，語音活動偵測

ABSTRACT

In this paper, we investigate rescaling of energy features for noise-robust speech recognition. The performance of the speech recognition system will degrade very quickly by the influence of environmental noise. As a result, speech robustness technique has become an important research issue for a long time. However, many studies have pointed out that the impact of speech recognition under the noisy environment is enormous. Therefore, we proposed the data-driven energy features rescaling (DEFR) to adjust the features. The method is divided into three parts, that are voice activity detection (VAD), piecewise log rescaling function and parameter searching algorithm. The purpose is to reduce the difference of noisy and clean speech features. We apply this method on Mel-frequency cepstral coefficients (MFCC) and Teager energy cepstral coefficients (TECC), and we compare the proposed method with mean subtraction (MS) and mean and variance normalization (MVN). We use the Aurora 2.0 and Aurora 3.0 databases to evaluate the performance. From the experimental results, we proved that the proposed method can effectively improve the recognition accuracy.

Keyword: Teager energy, noise-robust speech recognition, energy rescale, data-driven, voice activity detection

Contents

List of Tables

List of Figures

Chapter 1

介紹

1.1 研究動機與目的

近年來，科技快速發展，在硬體上許多輕薄短小的智慧型電子設備不斷的被開發出來，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爲主的互動方式。因此語音成爲人類與智慧型電子設備間最主要的人機介面，更進一步的帶動自動語音辨識 (automatic speech recognition, ASR) 技術的發展。目前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，要辨識龐大使用人潮錄製的語音訊號，辨識效能將奇差無比。而造成效能低落的主要因素有語者差異、背景噪音等等，我們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。

我們以常見的梅爾倒頻譜參數 (Mel-frequency cepstral coefficients, MFCC) 與 Teager 能量倒頻譜參數 (Teager energy cepstral coefficients, TECC) [1] 爲基準，並根據能量爲語音辨識之重要指標的特性 [2, 3, 4, 5]，結合對數能量尺度重刻 (Log energy rescaling, LER) [6] 來調整能量特徵。目的是希望能夠減少語音能量特徵受雜訊干擾所造成的失眞。然而，對數能量尺度重刻不論是在語音 (speech) 或非語音 (non-speech) 的段落皆使用相同的尺度，導致在雜訊能量大時，非語音段落的雜訊能量值下降幅度過小，因此我們提出資料驅動能量特徵重刻法 (data-driven energy features rescaling, DEFR) 來解決此問題。我們使用語音活動偵測 (voice activity detection, VAD) [7] 判

斷語音與非語音出現的時間點，再利用分段對數尺度函數 (piecewise log rescaling function) 對能量特徵做不同尺度的重刻，以減緩雜訊干擾的影響，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。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) [8]、說話方式 (speaking styles)、語者差異 (speaker variations) [9] 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。

近年來越來越多學者投入語音強健這方面的研究，因此也有越來越多的語音強健技術被提出。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 [10, 11] 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。而非調整辨識系統模型或特徵參數，希望藉由語音與雜訊所呈現不同的統計特性，來重建乾淨的語音訊號或是特徵參數。然而在語音強化的過程中，往往也能順帶提升辨識的正確率。常見的方法有維爾濾波器 (Wiener filter, WF) [12]、頻譜消去法 (spectral subtraction, SS) [13]、訊噪比波行處理 (signal waveform processing, SWP) [14]、噪音遮罩法 (noise masking, NM) [15] 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眞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) [16]、倒頻譜正規化法 (cepstral mean and variance normalization, CVN) [17]、聲道長度正規化法 (vocal tract length normalization, VTLN) [18]、統計圖等化法 (histogram equalization, HEQ) [19]、特徵空間旋轉法 (feature space rotation, FSR)、頻譜熵值特徵 (spectral entropy feature) [20] 等。

而後端處理的部分爲聲學模型的調適 (acoustic model adaptation)，主要藉由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。這類的方法優點是僅需要少量的語料就能對聲學模型進行調適；缺點是在進行即時調適時，計算量過大。常見的方法有最大相似度線性迴歸法 (maximum likelihood linear regression, MLLR) [21]、最大事後機率法則 (maximum a posteriori, MAP) [22]，以及平行模型合併法 (parallel model combination, PMC) [23]。而在本論文，我們主要的研究是朝強健性語音特徵的方向進行。

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 章爲說明我們所提出的資料驅動能量特徵重刻法，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

本章節將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的特徵擷取流程。第 2.2 小節則說明 Teager 能量倒頻譜參數的實作方法。

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。而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，我們很難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)

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

其中 $s_{pe}[n]$ 爲預強調處理過的輸出訊號，$s[n]$ 爲原始輸入訊號，α 爲預強調的參數 (本論文中設定爲 0.97)。

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，而這種現象會造成音框經由快速傅立葉轉換後產生高頻雜訊。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 $w[n]$ 爲漢明窗，式子如下

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

其中 N 音框取樣數。

4. 快速傅立葉轉換 (fast Fourier transform, FFT) ：語音訊號在時域上的變化十分快速且隨著時間不斷的改變，因此從時域上是很難觀察出它的特性。但是在頻域上短時間內的語音訊號是呈現週期性的，所以通常會經由快速傅立葉轉換將語音訊號轉換至頻域，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。而快速傅立葉轉換方程式爲

$$X(k) = \sum_{n=0}^{N-1} s_{hw}[n]\exp\left(-j\frac{2\pi kn}{N}\right),\ 0 \le k < N. \tag{2.3}$$

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，梅爾頻率和一般頻率 f 的關係式爲

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

6. 對數壓縮 (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。在此我們將梅爾頻率濾波器組中的能量取對數壓縮表示爲 $s_{\log}[m]$。

7. 離散餘弦轉換 (discrete cosine transform, DCT) ：在對數壓縮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。其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，增加辨識效率。方程式如下

$$c_i = \sum_{m=1}^{M} s_{\log}[m] \cos\left[\frac{\pi i\,(m-0.5)}{M}\right], \tag{2.5}$$

其中 c_i 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 濾波器 (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。

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)。GTF 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 (impulse response) 如圖 2.2 所示，其方程式爲

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_c t + \phi), \tag{2.6}$$

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

$$b = b_1 ERB(f_c), \tag{2.7}$$

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

$$\begin{aligned} ERB(f_c) &= \frac{\int |G(\omega_c)|^2 d\omega}{|G(\omega_c)|^2} \\ &= 6.23(\frac{f_c}{1000})^2 + 93.39(\frac{f_c}{1000}) + 28.52, \end{aligned} \tag{2.8}$$

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。

本論文中 GTF 所使用之梅爾中心頻率的計算公式爲

$$f_c[m] = Mel^{-1}\left(Mel(f_l) + m \times \frac{Mel(f_h) - Mel(f_l)}{M+1}\right), \tag{2.9}$$

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。在論文中，我們將 f_l、f_h 與 M 設定爲

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

然而，連續時間 GTF 函數是無法直接實作的，因此我們採用 [24] 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 (連續域)，得到 $G(s)$ 爲

$$G(s) = \frac{\left[s+b+\left(\sqrt{2}+1\right)\omega_c\right]\left[s+b-\left(\sqrt{2}+1\right)\omega_c\right]\left[s+b+\left(\sqrt{2}-1\right)\omega_c\right]\left[s+b-\left(\sqrt{2}-1\right)\omega_c\right]}{\left[\left(s+b+j\omega_c\right)\left(s+b-j\omega_c\right)\right]^4}. \tag{2.10}$$

再由 s 域轉換至 z 域 (離散域) 的關係爲 $z = e^{sT}$ (其中 T 爲取樣週期)，並令 $\cos(\omega_c T) = a_1$、$\sin(\omega_c T) = a_2$、$e^{-bT} = a_3$，得到 $G(z)$ 爲

$$\begin{aligned} G(z) = &\frac{T - Ta_3\left[a_1 + \left(\sqrt{2}+1\right)a_2\right]z^{-1}}{1 - 2a_1a_3z^{-1} + a_3^2z^{-2}} \times \frac{T - Ta_3\left[a_1 - \left(\sqrt{2}+1\right)a_2\right]z^{-1}}{1 - 2a_1a_3z^{-1} + a_3^2z^{-2}} \times \\ &\frac{T - Ta_3\left[a_1 + \left(\sqrt{2}-1\right)a_2\right]z^{-1}}{1 - 2a_1a_3z^{-1} + a_3^2z^{-2}} \times \frac{T - Ta_3\left[a_1 - \left(\sqrt{2}-1\right)a_2\right]z^{-1}}{1 - 2a_1a_3z^{-1} + a_3^2z^{-2}}. \end{aligned} \tag{2.11}$$

將上式簡化後，其表示式爲

$$G(z) = \frac{\sum_{j=0}^{5} num_j z^{-j}}{\sum_{i=0}^{9} den_i z^{-i}}. \tag{2.12}$$

最後利用 z 轉換求得離散時間 GTF 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

$$\psi\left[x(t)\right] = \left[\frac{d}{dt}x(t)\right]^2 - x(t)\left[\frac{d^2}{dt^2}x(t)\right]. \tag{2.13}$$

將方程式(2.13) 轉換爲離散型式，式子如下

$$\psi\left[x(n)\right] = \left[x(n)\right]^2 - x(n+1)x(n-1). \tag{2.14}$$

我們發現方程式(2.14) 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

$$x[n] = \begin{cases} (x[n])^2 - x[n]\cdot x[n+1], & \text{if } n = 0 \\ (x[n])^2 - x[n]\cdot x[n-1], & \text{if } n = N-1 \\ (x[n])^2 - x[n+1]\cdot x[n-1], & \text{otherwise.} \end{cases} \tag{2.15}$$

本論文中，我們將 GTF 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。

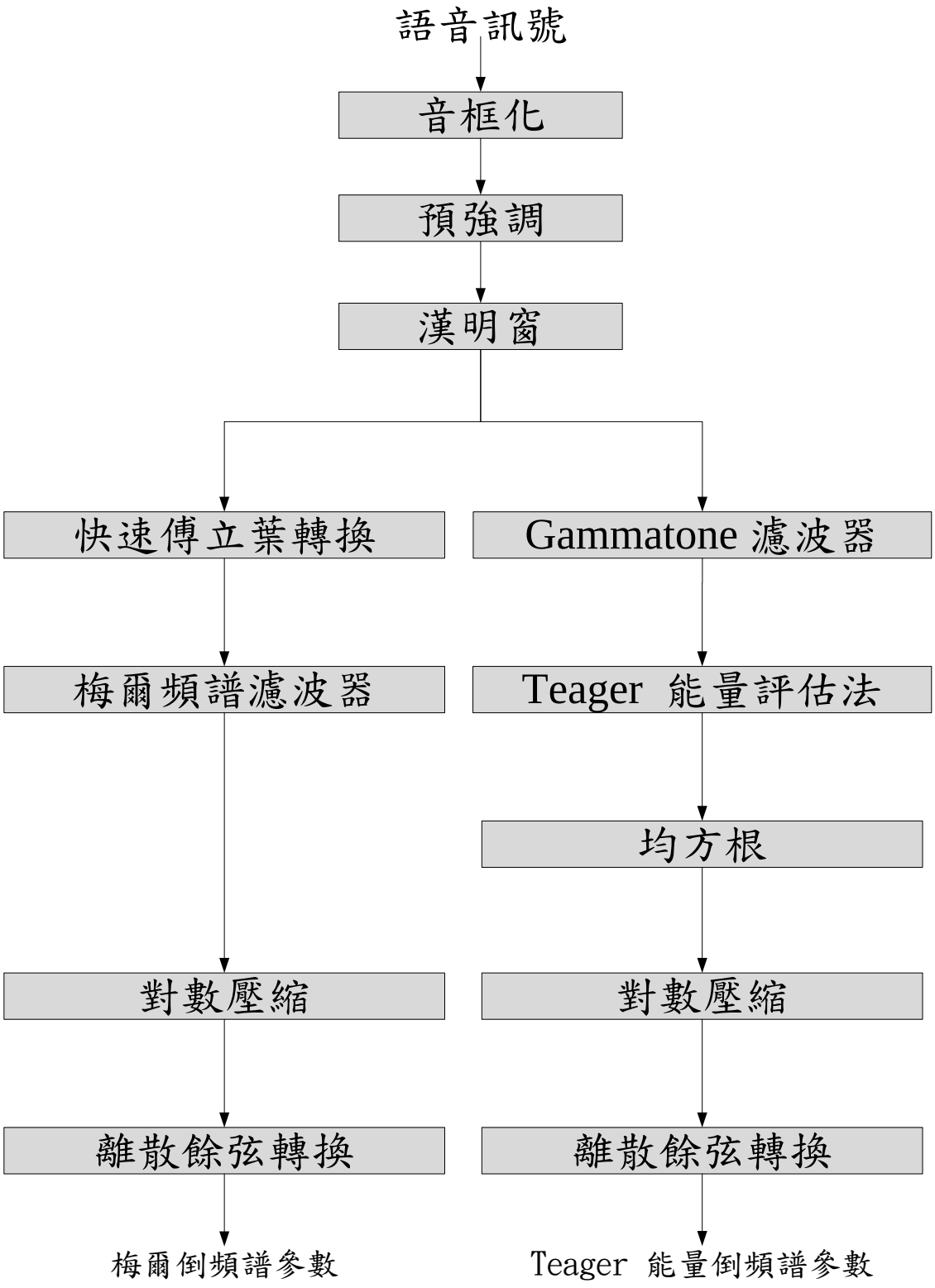

圖 2.1: 梅爾倒頻譜參數與 Teager 能量倒頻譜參數的擷取流程

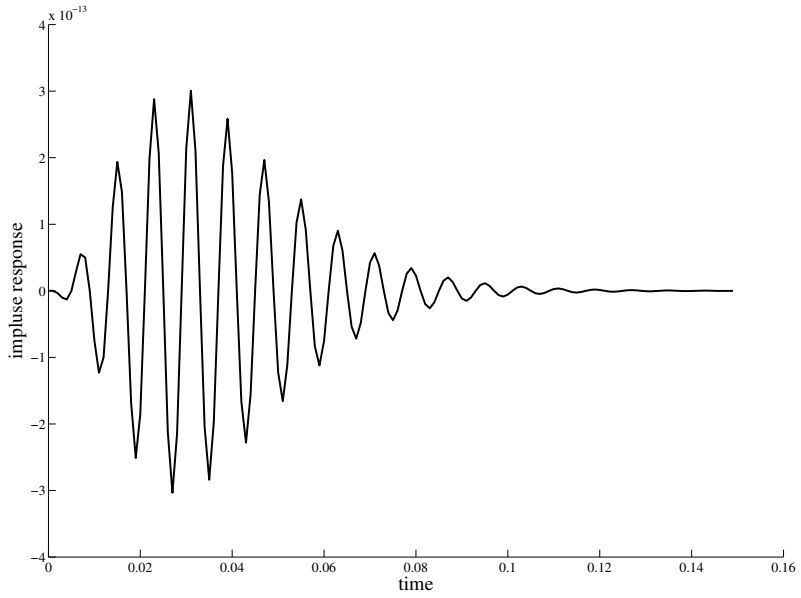

圖 2.2: Gamma-tone 濾波器之脈衝響應(中心頻率爲 1000 Hz)

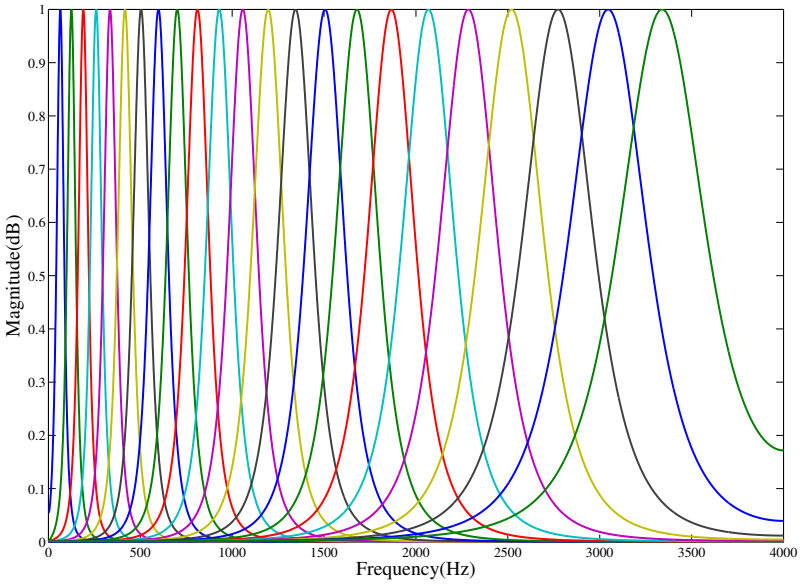

圖 2.3: Gamma-tone 濾波器之頻率響應

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來擷取特徵參數，並對其對數能量 (Log energy) 或第零階的倒頻譜係數 (c_0) 進行重刻，流程如圖 3.1。此作法主要目的是希望能夠使得雜訊語音的能量特徵值近似於乾淨語音，以提高辨識的準確度。而本論文所提出之 DEFR 的作法是參考自 [6] 所提出之 LER，並加以修改與創新。我們將在第 3.1 小節對 DEFR 作詳細說明。

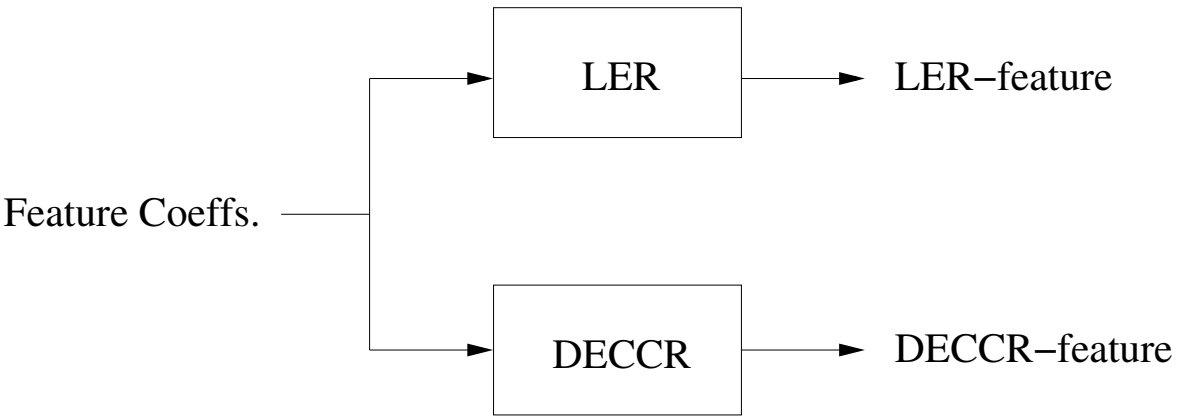

圖 3.1: 重刻特徵參數流程圖

3.1 資料驅動能量特徵重刻法

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調整能量特徵以減少乾淨與含噪音特徵兩者的差異性。我們觀察能量特徵的變化

發現，一段雜訊語音中有語音的段落，其能量值偏高；反之，非語音的部分，能量值偏低。基於此現象，本論文提出 DEFR 的方法對能量特徵作進一步的處理。

此方法主要分爲語音活動偵測、分段對數尺度函數與參數搜尋法三個部分。我們利用 [7] 所提出的低頻譜之語音活動偵測來取得語音與非語音出現的時間點，並利用分段對數尺度函數決定能量參數調整的權重值。而分段對數尺度函數所使用的參數是由參數搜尋法自動決定。以下我們就這三個部分做詳細說明。

3.1.1 低頻譜之語音活動偵測

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。因此發展了語音活動偵測來判斷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 $[0, 50\text{Hz}]$ 之間都有相當比例的能量。藉由此特性，我們計算每個音框的低頻帶頻譜能量，並根據能量值，判斷該語音音框是否爲純雜訊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

$$X^{(i)}(f_k) = X^{(i)}[k] = \sum_{n=0}^{N-1} x_i[n] e^{-j\frac{2\pi kn}{K}}, 0 \le k \le K-1, \tag{3.1}$$

其中 f_k 爲頻率，其值爲

$$f_k = \frac{F_s}{2K} k, \tag{3.2}$$

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(3.1) 可以得到每個頻率 f_k 的能量值，因此我們定義頻帶 $[F_L, F_H]$ 之頻譜能量計算的方式爲

$$Y^{(i)}_{[F_L, F_U]} = \sum_{F_L \le f_k \le F_U} \mid X_i(f_k) \mid . \tag{3.3}$$

根據方程式(3.3)，我們計算每個音框之低頻帶頻譜強度，即 0 至 50Hz 以內的頻譜強度如下

$$Y^{(i)}_{[0,50]} = \sum_{0 \le f_k \le 50} \mid X_i(f_k) \mid . \tag{3.4}$$

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

$$\theta = \lambda \left(\frac{1}{P} \sum_{i=0}^{P-1} Y_{[0,50]}^{(i)} \right), \tag{3.5}$$

其中 P 爲純噪音的音框數。我們將每個音框低頻帶內的頻譜能量 $Y_{[0,50]}^{(i)}$ 與門檻值 θ 做比較，若 $Y_{[0,50]}^{(i)} \leq \theta$ 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：

$$\text{第}i\text{個音框} = \begin{cases} Y_{[0,50]}^{(i)} \leq \theta, & \text{非語音音框} \\ Y_{[0,50]}^{(i)} > \theta, & \text{語音音框.} \end{cases} \tag{3.6}$$

VAD 的使用讓我們所提出之能量重刻的方法可以更加精確。然而 VAD 的準確度對我們的實驗結果影響甚鉅，因此爲了使 VAD 的準確度提高，我們藉由改變方程式 (3.5) 中的 P 與 λ，使得 VAD 的準確度最佳化。其中 P 與 λ 範圍設定如下

$$6 \leq P \leq 10, \quad 0.1 \leq \lambda \leq 2.0.$$

我們將強制對齊 (force alignment) 的方法實作在 Aurora 2.0 訓練語料庫，其實驗結果爲語音與非語音訊號出現的時間點，我們將此結果視爲參考答案，接著與上述低頻帶能量之語音活動偵測器所得到的結果進行比對。從實驗結果發現，當 $P = 10$ 與 $\lambda = 1.9$ 時 VAD 的準確率最高，其值如下表 3.1。

表 3.1: 低頻譜能量之語音活動偵測器的準確度

clean-train	multi-train
83.92	67.11

3.1.2 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確實地保留原有的能量值，而非語音的能量能夠大幅度地下降，以增加語音與非語音能量的差異性。首先，假設我們有一段語句 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

$$\tilde{c}_0[i] = w[i]c_0[i], \tag{3.7}$$

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

$$\begin{aligned} r[i] \approx 1 \quad &\longrightarrow \quad w[i] \approx 1, \\ r[i] \approx 0 \quad &\longrightarrow \quad w[i] \approx 0. \end{aligned} \tag{3.8}$$

已經有很多方法實作方程式(3.8) 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

$$w[i] = \begin{cases} \left[\frac{\log(r[i]\times M)}{\log(M)}\right]^{\alpha_1}, & Y^{(i)}_{[0,50]} \leq \theta \\ \left[\frac{\log(r[i]\times M)}{\log(M)}\right]^{\alpha_2}, & Y^{(i)}_{[0,50]} > \theta. \end{cases} \tag{3.9}$$

其中 $Y^{(i)}_{[0,50]}$ 爲 0 至 50 Hz頻譜的能量，θ 爲語音與非語音的門檻值。方程式(3.9) 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 3.1.3 節的參數搜尋法取得最佳的參數，其函數如圖 3.3。訓練與測試資料皆使用該參數進行能量重刻。

而 LER 與我們所提出之 DEFR 最大的差別在於 LER 權重的計算是使用對數轉換函數 (如圖 3.2)，其式子如下

$$w[i] = \frac{\log\left(\lfloor r[i] \times M \rfloor\right)}{\log\left(M\right)}. \tag{3.10}$$

我們以 LER 與 DEFR 調整 MFCC 與 TECC 的能量參數值，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 MFCC 與 TECC 再重刻運算後，明顯地減少了乾淨與雜訊語句之差異。

3.1.3 參數搜尋法

本小節所提出的參數搜尋法，目的爲尋找第 3.1.2 小節所使用之參數 α_1 及 α_2，使得乾淨與雜訊語句特徵值的失眞 (distortion) 最小化。此方法必須滿足

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

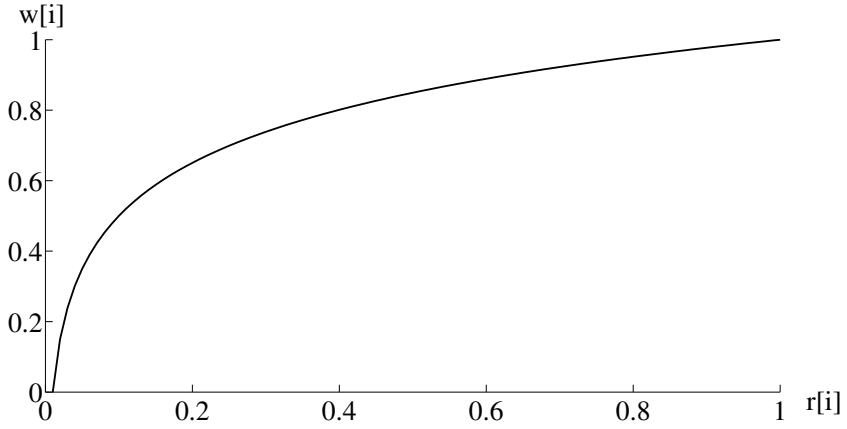

圖 3.2: 對數轉換函數

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。而失眞的計算方法如下式

$$D(\alpha_1, \alpha_2, \theta) = \sum_{u=1}^{U} \sqrt{\sum_{i=0}^{N-1} \left(N^u[i] - C^u[i]\right)^2}, \tag{3.11}$$

其中 U 表示訓練語句的數目，N 表示語句 u 的音框數，$C^u[i]$ 與 $N^u[i]$ 表示乾淨與含噪音語句 u 的第 i 個音框重刻後的特徵值。

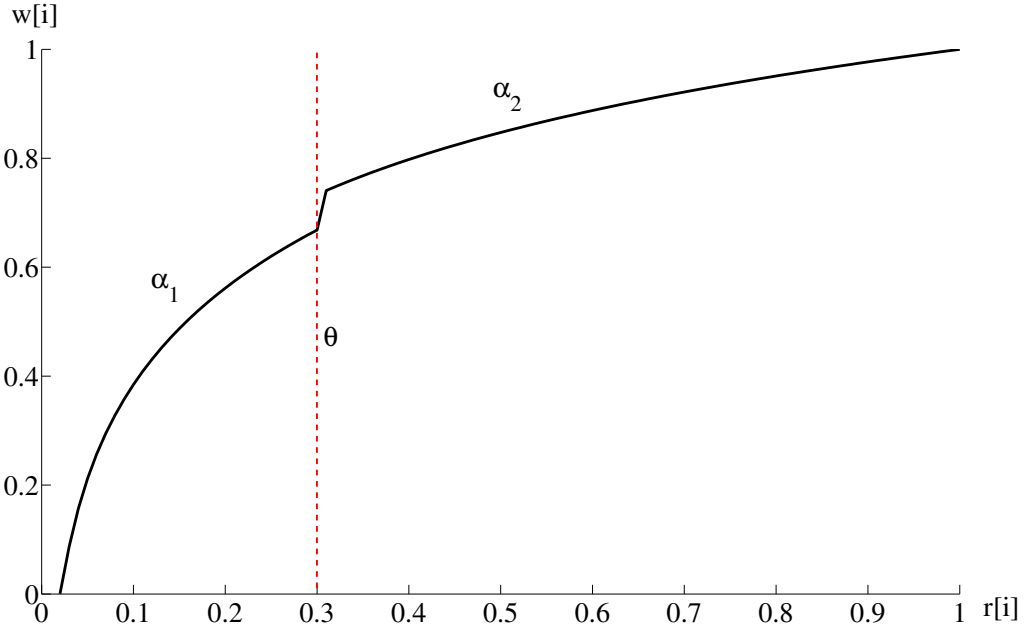

圖 3.3: 分段對數尺度函數。圖中 α_1 和 α_2 分別爲非語音和語音能量所使用之參數

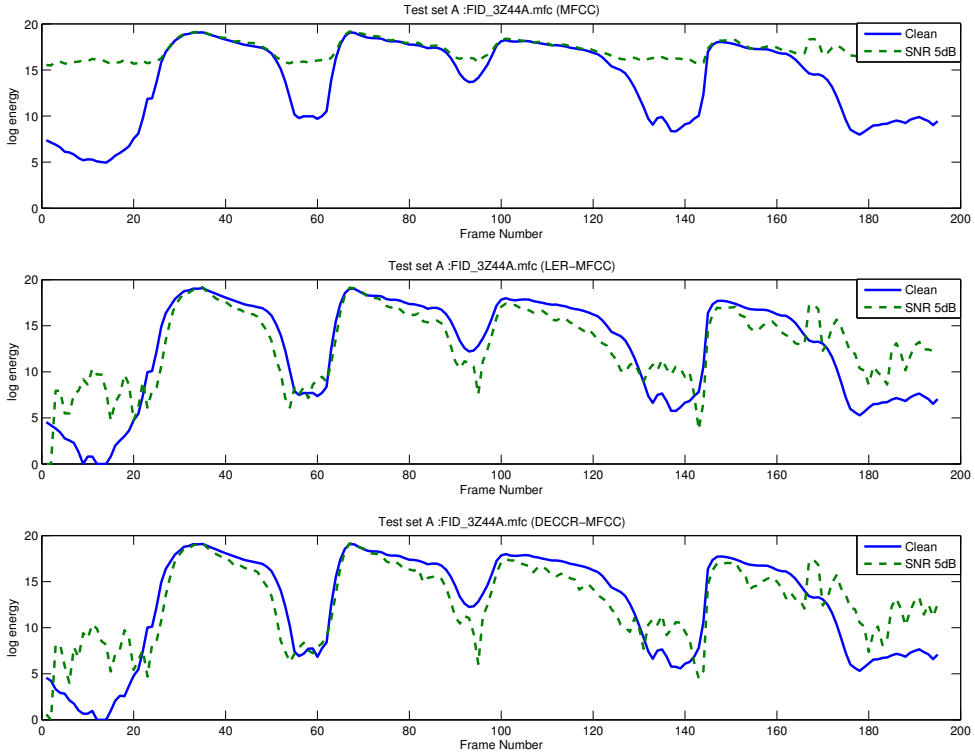

圖 3.4: 一對平行語句之 MFCC (上)、LER-MFCC (中) 與 DEFR-MFCC (下) 對數能量序列的比較，語句的 ID 爲 FID_3ZZ4A.08。

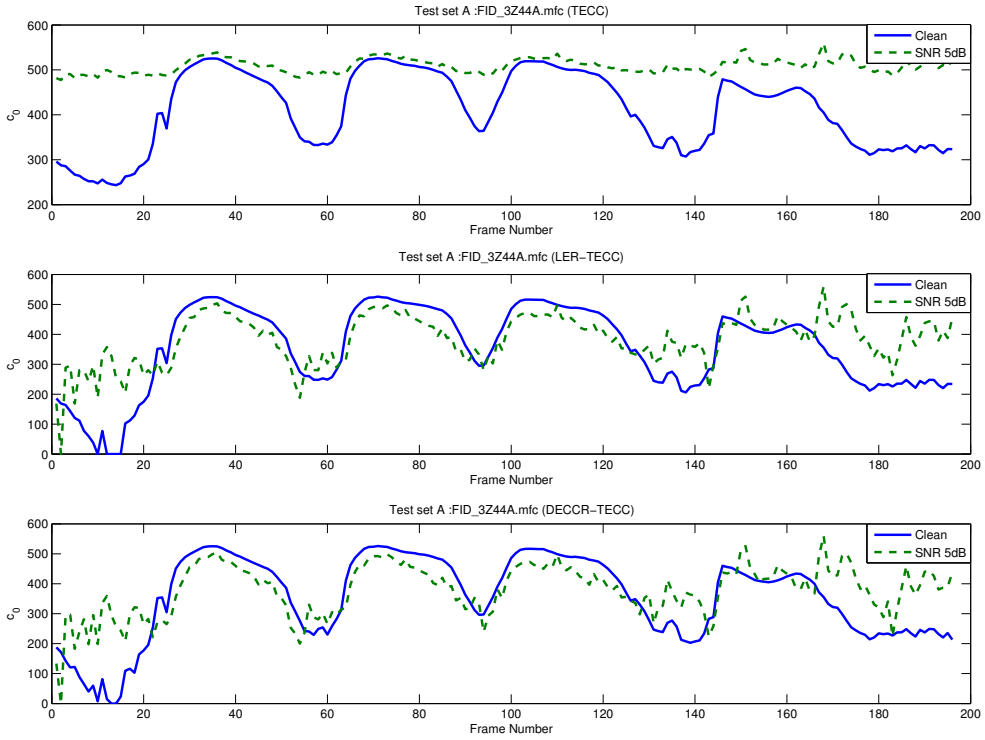

圖 3.5: 一對平行語句之 TECC (上)、LER-TECC (中) 與 DEFR-TECC (下) c_0 特徵序列的比較，語句的 ID 爲 FID_3ZZ4A.08。

Chapter 4

實驗

4.1 辨識系統設定

本實驗所使用的特徵參數爲 MFCC (Aurora frontend wI007)、 TECC 與 advanced front-end (AFE)，其特徵向量是由 $c_1, ..., c_{12}$ 及第 13 維的能量特徵所組成，而能量特徵的使用如表 4.1 的設定。接著我們使用均值消去法 (mean subtraction, MS)、均值正規化法 (mean and variance normalization, MVN)、LER 、 DEFR 以及 MVN ($c_1 \sim c_{12}$) 對能量特徵作進一步處理。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。此外，MVN ($c_1 \sim c_{12}$) 表示只有對特徵向量 $c_1, ..., c_{12}$ 進行的處理。

表 4.1: MFCC、TECC 與 AFE 之能量特徵使用的設定

MFCC	TECC	AFE
Log energy	c_0	Log energy

後端的聲學模型是採用隱藏式馬可夫模型 (hidden Markov model, HMM)，模型內狀態的轉移情形可分爲兩種，一種是停留在原狀態，另一種是由左至右跳到下一個相鄰的狀態。此外，除了數字 1 至 9 分別都有一個相對應的聲學模型外，阿拉伯數字 0 有 zero 和 oh 兩種聲學模型。而每個模型都有 18 個狀態 (states)，包含前後兩個模型間連接用的空狀態，且每個狀態均包含 3 個高斯混合分佈 (Gaussian mixture distributions)。除了數字的聲學模型外，另外還有靜音 (silence) 模型與間歇 (short-

pause) 模型。靜音模型包含 3 個狀態，每個狀態有 6 個高斯混合分佈。而間歇模型包含 1 個狀態，並與靜音模型最中間的狀態共用。

4.2 實驗語料

4.2.1 Aurora 2.0

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 語料庫是由美國成年男女所錄製的乾淨環境連續數字，以人工的方式加入八種不同來源的加成性噪音 (additive noise)，分別爲地下鐵 (subway)、細語 (babble)、汽車 (car)、宴會 (exhibition)、餐廳 (restaurant)、街道 (street)、機場 (airport) 與火車站 (train station) 等，以及不同程度的訊噪比 (signal-to-noise ratio, SNR)，分別爲 -5dB、0dB、5dB、10dB、15dB、20dB 與 clean 等，來觀察噪音對訊號所造成的影響；通道效應則是包含由國際電信聯合會所訂立的兩個標準 G.712 和 MIRS。根據測試語料加入之通道噪音以及加成性噪音的種類不同，可分爲 Set A、Set B 和 Set C 三種測試集合。

4.2.2 Aurora 3.0

Aurora 3.0 語料庫包含四種不同語言，分別爲 Spanish，Finnish，German 和 Danish，其音檔是使用近身 (close-talking) 與手持 (hands-free)兩種不同麥克風錄製而成。而訓練與測試分爲三種不同的場景，分別爲 WM (well-match)，MM (medium-mismatch) 與 HM (high-mismatch)。WM 的訓練與測試集皆採用兩種麥克風及三種訊噪比的語料。MM 之訓練與測試語料只包含手持麥克風所錄製之音檔。HM 則利用手持麥克風錄製之語料爲訓練集，近身麥克風錄製之語料爲測試集。

4.3 效能評估方法

Aurora 2.0 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 SNR $0-20$dB 的語料。實驗結果如表 4.3、4.4 所示。Aurora 2.0 的三個測試集語料的數目比爲 $2:2:1$，而表中平均 (Avg) 的欄位是根據此比例計

算得到，其方程式如下

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

而表 4.5、4.6、4.7、4.8 中的 Avg 欄位則是使用方程式(4.2) 計算得到。

$$\textbf{Avg} = \textbf{HM} \times 0.25 + \textbf{MM} \times 0.35 + \textbf{WM} \times 0.4. \tag{4.2}$$

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

$$rimp = \frac{S_c - S_b}{100 - S_b} \times 100\%, \tag{4.3}$$

其中 S_c 是我們想要與基準比較之平均值，S_b 爲基準的平均。

4.4 實驗結果

本小節將會對實驗結果做分析與說明。而實驗中分段對數函數所使用的 α_1 與 α_2 設定如表 4.4 所示。我們將對實驗結果對 Aurora 2.0 與 Aurora 3.0 個別探討。在 Aurora 2.0 的實驗中，我們將聲學模型分爲 clean-train 與 multi-train，藉此分析能量特徵對語音辨識上的影響。其分析如下：

1. Clean-train 之實驗結果：

   從表4.3 的實驗結果，我們發現 MFCC 與 TECC 經過 MS 後，雖然只有平移語音訊號，相對改善率卻成功的提升爲 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 MVN、LER 與 DEFR 的目的皆爲減少乾淨與雜訊語音之能量特徵的差距。從實驗結果證實 MVN、LER 與 DEFR 的方法大幅度地突破 MS 的辨識率，其中 DEFR 的方法效果最好。此外，爲了使辨識率有更進一步的提升，我們將 DEFR 與 MVN 做結合，使得辨識率突破原有的瓶頸。然而，我們將同樣的方法實作在 AFE 上，發現效果相當不佳，其原因在於 AFE 在訊號處理的過程包含了降噪的部分，因此對其特徵參數做進一步的調整效益不大。綜合比較 DEFR-MFCC 與 DEFR-TECC 之辨識結果，DEFR-TECC 的結果明顯地優於 DEFR-MFCC，進一步證實了 DEFR 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雜訊語音的訓練語料。我們比較表 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，加入了雜訊語音的訓練語料，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。

我們由 Aurora 2.0 的實驗可以很明確地看出能量特徵重刻對辨識效能的影響極大。然而 Aurora 2.0 語料庫是以人工的方式加入噪音，其噪音的變化與現實生活相比還是有所落差。因此我們將同樣的實驗設定實作在 Aurora 3.0 語料庫上，做更進一步的驗證。從表 4.5、4.6、4.7、4.8 中，我們可以發現 MVN 的結果好壞會直接影響 DEFR + MVN ($c_1 \sim c_{12}$) 的辨識率。此外，我們只有使用 DEFR 的方法做特徵擷取，其辨識率與基準相比，有極大幅度的提升。因此我們可以證實 DEFR 具有非常傑出的噪音強健性。

表 4.2: 分段對數函數實作在 Aurora 2.0 與 Aurora 3.0 語料庫上所使用之 α_1 與 α_2 的設定

語料庫	MFCC		TECC		AFE	
	α_1	α_2	α_1	α_2	α_1	α_2
Aurora 2.0	1.9	1.8	1.9	1.8	1.9	1.0
Spanish	1.9	1.8	1.9	1.8	1.9	1.0
Danish	1.9	1.8	1.9	1.7	1.9	1.7
German	1.1	1.0	1.9	1.8	1.1	1.0
Finnish	1.1	1.0	1.9	1.8	1.9	1.0

表 4.3: 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 SNR $0-$ 20dB 之平均。

Feature	Set A	Set B	Set C	Avg.	rimp
MFCC	61.34	55.75	66.14	60.06	-
MFCC+MS	66.18	70.81	64.88	67.77	19.30
MFCC+MVN	70.18	70.77	66.37	69.65	24.01
LER-MFCC	74.60	74.51	65.23	72.69	31.62
DEFR-MFCC	74.92	75.82	64.37	73.17	32.82
LER-MFCC+MVN ($c_1 \sim c_{12}$)	78.40	78.13	76.48	77.91	44.69
DEFR-MFCC+MVN ($c_1 \sim c_{12}$)	79.19	79.26	76.44	78.67	46.59
TECC	55.55	51.79	65.30	56.00	-
TECC+MS	66.92	71.52	67.67	68.91	29.34
TECC+MVN	74.91	75.38	76.03	75.32	43.91
LER-TECC	78.57	79.60	72.77	77.82	49.59
DEFR-TECC	79.24	80.69	71.60	78.29	50.66
LER-TECC+MVN ($c_1 \sim c_{12}$)	80.77	82.26	81.32	81.47	57.89
DEFR-TECC+MVN ($c_1 \sim c_{12}$)	81.10	82.18	81.07	81.53	58.02
AFE	86.69	85.57	82.81	85.47	-
AFE+MS	84.91	85.62	83.39	84.89	-3.99
AFE+MVN	76.80	76.85	74.39	76.34	-62.84
LER-AFE	85.45	85.04	81.09	84.42	-7.23
DEFR-AFE	85.59	85.02	81.29	84.59	-6.06
LER-AFE+MVN ($c_1 \sim c_{12}$)	83.21	83.35	79.70	82.57	-19.96
DEFR-AFE+MVN ($c_1 \sim c_{12}$)	83.23	83.19	79.62	82.50	-20.44

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 SNR $0-20$dB 之平均。

Feature	Set A	Set B	Set C	Avg.	rimp
MFCC	87.82	86.27	83.78	86.39	-
MFCC+MS	88.72	87.79	87.25	88.06	12.27
MFCC+MVN	89.67	88.07	86.10	88.32	14.18
LER-MFCC	89.36	86.54	85.51	87.46	7.86
DEFR-MFCC	89.68	87.68	85.54	88.05	12.20
LER-MFCC+MVN ($c_1 \sim c_{12}$)	90.94	89.84	89.93	90.30	28.73
DEFR-MFCC+MVN ($c_1 \sim c_{12}$)	90.77	89.56	89.38	90.01	26.60
TECC	88.07	87.09	85.74	87.21	-
TECC+MS	89.14	89.33	89.66	89.32	16.50
TECC+MVN	90.71	90.34	89.96	90.41	25.02
LER-TECC	89.86	88.99	87.75	89.09	14.70
DEFR-TECC	90.23	89.26	88.21	89.44	17.44
LER-TECC+MVN ($c_1 \sim c_{12}$)	91.23	91.03	91.18	91.14	30.73
DEFR-TECC+MVN ($c_1 \sim c_{12}$)	91.16	90.98	91.02	91.06	30.10
AFE	91.79	90.76	89.11	90.84	-
AFE+MS	91.66	91.13	90.24	91.17	3.60
AFE+MVN	90.99	89.68	88.11	89.89	-10.37
LER-AFE	92.00	90.97	89.47	91.08	2.62
DEFR-AFE	91.85	90.85	89.29	90.93	0.98
LER-AFE+MVN ($c_1 \sim c_{12}$)	91.71	90.73	89.74	90.81	-0.33
DEFR-AFE+MVN ($c_1 \sim c_{12}$)	91.62	90.27	89.74	90.70	-1.53

表 4.5: Aurora 3.0 Spanish 之詞辨識率。

Feature	WM	MM	HM	Avg.	rimp
MFCC	86.90	73.74	42.23	71.13	-
MFCC+MS	90.27	83.43	66.50	81.93	37.41
MFCC+MVN	92.02	84.37	68.96	83.58	43.12
LER-MFCC	88.87	77.90	70.41	80.42	32.18
DEFR-MFCC	88.68	79.42	68.12	80.30	31.76
LER-MFCC+MVN ($c_1 \sim c_{12}$)	92.90	83.03	76.54	85.36	49.29
DEFR-MFCC+MVN ($c_1 \sim c_{12}$)	92.87	86.48	78.23	86.97	54.87
TECC	85.91	75.92	42.44	71.55	-
TECC+MS	89.44	82.68	59.25	79.53	28.05
TECC+MVN	91.04	87.23	76.69	86.12	51.21
LER-TECC	89.75	75.41	64.21	78.35	23.90
DEFR-TECC	88.95	80.92	64.33	79.98	29.63
LER-TECC+MVN ($c_1 \sim c_{12}$)	92.68	85.19	72.42	84.99	47.24
DEFR-TECC+MVN ($c_1 \sim c_{12}$)	92.92	88.03	74.17	86.52	52.62
AFE	92.17	84.02	82.65	86.94	-
AFE+MS	93.79	89.17	87.10	90.50	27.26
AFE+MVN	92.27	85.85	80.57	87.10	1.23
LER-AFE	93.62	86.22	86.29	89.20	17.30
DEFR-AFE	93.37	87.70	86.89	89.77	21.67
LER-AFE+MVN ($c_1 \sim c_{12}$)	94.48	88.36	84.39	89.82	22.05
DEFR-AFE+MVN ($c_1 \sim c_{12}$)	94.34	88.86	85.80	90.29	25.65

表 4.6: Aurora 3.0 Danish 之詞辨識率。

Feature	WM	MM	HM	Avg.	rimp
MFCC	79.64	49.01	33.19	57.31	-
MFCC+MS	84.35	63.14	39.30	65.66	19.56
MFCC+MVN	80.54	59.60	48.82	65.28	18.67
LER-MFCC	83.68	58.05	51.65	66.70	22.00
DEFR-MFCC	83.96	62.15	56.35	69.42	28.37
LER-MFCC+MVN ($c_1 \sim c_{12}$)	81.29	59.04	50.47	65.80	19.89
DEFR-MFCC+MVN ($c_1 \sim c_{12}$)	83.27	62.29	55.84	69.07	27.55
TECC	80.03	49.44	28.21	56.37	-
TECC+MS	83.64	64.83	33.58	64.54	18.73
TECC+MVN	81.66	58.62	52.27	66.25	22.64
LER-TECC	83.54	57.63	47.77	65.52	20.97
DEFR-TECC	84.64	58.90	52.98	67.72	26.01
LER-TECC+MVN ($c_1 \sim c_{12}$)	81.78	49.01	44.28	60.94	10.47
DEFR-TECC+MVN ($c_1 \sim c_{12}$)	83.84	51.98	51.14	64.51	18.66
AFE	85.84	62.15	64.77	72.28	-
AFE+MS	83.06	65.25	65.91	72.54	0.94
AFE+MVN	79.62	62.15	53.88	67.07	-18.80
LER-AFE	87.43	65.68	70.89	75.68	12.27
DEFR-AFE	87.31	66.67	68.77	75.45	11.44
LER-TECC+MVN ($c_1 \sim c_{12}$)	81.29	62.85	50.63	67.17	-18.43
DEFR-TECC+MVN ($c_1 \sim c_{12}$)	80.80	63.84	54.62	68.32	-14.29

表 4.7: Aurora 3.0 German 之詞辨識率。

Feature	WM	MM	HM	Avg.	rimp
MFCC	90.58	79.06	74.28	82.47	-
MFCC+MS	91.44	79.43	76.55	83.51	5.93
MFCC+MVN	90.98	81.04	77.20	84.06	9.07
LER-MFCC	91.10	81.55	75.39	83.83	7.76
DEFR-MFCC	91.42	80.53	75.21	83.56	6.22
LER-MFCC+MVN ($c_1 \sim c_{12}$)	93.23	84.63	83.63	87.82	30.52
DEFR-MFCC+MVN ($c_1 \sim c_{12}$)	92.41	81.26	81.91	85.88	19.45
TECC	90.84	78.62	68.73	81.04	-
TECC+MS	91.79	78.62	76.73	83.42	12.55
TECC+MVN	91.58	80.16	83.81	85.64	24.26
LER-TECC	92.03	81.99	77.94	84.99	20.83
DEFR-TECC	92.63	82.36	76.64	85.04	21.10
LER-TECC+MVN($c_1 \sim c_{12}$)	93.11	81.19	85.20	86.96	31.22
DEFR-TECC+MVN($c_1 \sim c_{12}$)	92.93	82.06	86.31	87.47	33.91
AFE	94.59	89.02	89.41	91.35	-
AFE+MS	94.53	87.12	90.06	90.82	-6.13
AFE+MVN	93.17	83.89	85.48	88.00	-38.73
LER-AFE	94.17	88.43	88.58	90.76	-6.82
DEFR-AFE	94.15	88.07	88.71	90.66	-7.98
LER-AFE+MVN ($c_1 \sim c_{12}$)	94.61	88.14	87.47	90.56	-9.13
DEFR-AFE+MVN ($c_1 \sim c_{12}$)	93.37	86.31	87.05	89.32	-23.47

表 4.8: Aurora 3.0 Finnish 之詞辨識率。

Feature	WM	MM	HM	Avg.	rimp
MFCC	90.39	72.37	31.06	69.25	-
MFCC+MS	93.76	86.11	44.91	78.87	31.28
MFCC+MVN	82.23	78.45	14.35	63.94	-17.27
LER-MFCC	93.66	75.85	40.85	74.22	16.16
DEFR-MFCC	94.12	76.13	41.66	74.71	17.76
LER-MFCC+MVN ($c_1 \sim c_{12}$)	86.13	68.54	39.12	68.22	-3.35
DEFR-MFCC+MVN ($c_1 \sim c_{12}$)	86.00	70.18	43.00	69.71	1.50
TECC	92.23	65.12	37.17	68.98	-
TECC+MS	93.31	83.45	45.02	77.79	28.40
TECC+MVN	81.07	77.50	60.78	74.75	18.60
LER-TECC	92.15	70.93	64.56	77.83	28.53
DEFR-TECC	93.04	72.78	51.48	75.56	21.21
LER-TECC+MVN ($c_1 \sim c_{12}$)	78.36	70.18	68.09	72.93	12.73
DEFR-TECC+MVN ($c_1 \sim c_{12}$)	80.55	74.28	70.92	75.95	22.47
AFE	95.19	77.70	68.66	82.44	-
AFE+MS	95.84	88.58	80.49	89.46	39.98
AFE+MVN	87.48	71.14	61.17	75.18	-41.34
LER-AFE	95.67	83.17	72.93	85.61	18.05
DEFR-AFE	95.79	82.83	79.65	87.22	27.22
LER-AFE+MVN ($c_1 \sim c_{12}$)	88.00	76.88	66.40	78.71	-21.24
DEFR-AFE+MVN ($c_1 \sim c_{12}$)	88.26	75.31	72.61	79.82	-14.92

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，然而特徵向量中最重要的參數爲對數能量與第零階倒頻譜參數。這兩種參數爲語音訊號在各個頻帶間的整體表現，因此極具重要性。在本論文中，我們針對這兩種能量特徵提出了資料驅動能量特徵重刻法，並以 MFCC 與 TECC 爲基準對其能量特徵做調整。此方法藉由 VAD 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雜訊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的差異性，故給予不同尺度的權重。其函數所使用的參數是由參數搜尋法自動決定。最後，我們採用 Aurora 2.0 與 Aurora 3.0 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本論文所提出之方法不論是在 Aurora 2.0 或 Aurora 3.0 語料庫中，皆有非常卓越的表現，這也證實了能量特徵的調整對於噪音強健性的影響非常大。

5.2 未來展望

資料驅動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 VAD 爲主要目標，主要原因爲本論文所使用之 VAD 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，很難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 VAD 對本論文的方法是相當重要的，若能提高準確度，辨識效能也會相對的有所突破。

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, pp. 1504–1516, August 2011.

[2] 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), Philadelphia*, vol. 1, pp. 245–249, March 2005.

[3] 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 (ISCSLP 2004), Hong Kong*, pp. 249 – 252, December 2004.

[4] 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(ICSLP 2004), Korea*, October 2004.

[5] 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), Hungary*, September 1999.

[6] 陳鴻彬, "On the Study of Energy-Based Speech Feature Normalization and Application to Voice Activity Detection," 國立臺灣師範大學資訊工程學系碩士論文, 2007.

[7] 杜文祥, ”Study on the Voice Activity Detection Techniques for Robust Speech Feature Extraction,” 國立暨南國際大學電機工程學系碩士論文, 2007.

[8] 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, pp. 1082 –1086, July 2010.

[9] 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.

[10] 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, pp. 882–892, March 2007.

[11] 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.

[12] 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, pp. 417–426, Feb 2012.

[13] R. Gomez, A. Lee, H. Saruwatari, and K. Shikano, “Robust speech recognition with spectral subtraction in low SNR,” in *proceedings of 8th International Conference on Spoken Language Processing (ICSLP 2004), Korea*, October 2004.

[14] D. Macho and Y. M. Cheng, “SNR-dependent waveform processing for improving the robustness of ASR front-end,” in *proceedings of 2001 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, vol. 1, pp. 305–308, 2001.

[15] C. Cerisara, S. Demange, and J. P. Haton, “On noise masking for automatic missing data speech recognition: A survey and discussion,” *Computer Speech & Language*, vol. 21, no. 3, pp. 443–457, 2007.

[16] 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(DSP 2011), Greece*, vol. 21, pp. 36–53, July 2011.

[17] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, pp. 133–147, August 1998.

[18] T. Claes, I. Dologlou, L. ten Bosch, and D. V. Compernolle, "A novel feature transformation for vocal tract length normalization in automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 549–557, November 1998.

[19] Y. Obuchi and R. M. Stern, "Normalization of time-derivative parameters using histogram equalization," in *proceedings of 8th European Conference on Speech Communication and Technology(EUROSPEECH 2003), Switzerland*, September 2003.

[20] 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), Philadelphia*, vol. 1, pp. 253–256, March 2005.

[21] 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), Phoenix*, vol. 2, pp. 777–780, March 1999.

[22] 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, April 1994.

[23] H. Veisi and H. Sameti, "An improved parallel model combination method for noisy speech recognition," in *proceedings of 2009 IEEE Workshop on Automatic Speech Recognition & Understanding(ASRU 2009), Italy*, pp. 237–242, December 2009.

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