端點偵測技術在強健語音參數擷取之研究
Study of the Voice Activity Detection Techniques for Robust Speech Feature Extraction

杜文祥 Wen-Hsiang Tu
暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
s94323537@ncnu.edu.tw

洪志偉 Jeih-weih Hung
暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
jwhung@ncnu.edu.tw

摘要

由於發展環境和應用環境兩者之間的不匹配，導致於語音辨識系統效能經常會下降，而引起這不匹配的主要原因之一是加成性雜訊，處理加成性雜訊的方法我們可以分成三類，語音強化法、強健性語音特徵參數、以及語音模型調適法，而本論文所討論的方法主要是屬於強健性語音特徵參數之技術。

在本論文中，我們主要的重點在於探討不同的信號特徵對於語音端點偵測的影響，所利用的特徵分別爲低頻帶頻譜強度、全頻帶頻譜強度、累積量化頻譜、以及高通對數能量等。利用以上這些不同的特徵進行語音之端點偵測，所得之純雜訊的位置資訊可以提供頻譜消去法與靜音對數能量正規化法中所需的雜訊頻譜或能量的估測。

在實驗環境上我們採用 Aurora2 語料庫，在八種背景雜訊以及訊雜比 0~20dB 下做實驗。在第五章中所呈現的實驗數據與分析可證明以上所述的各種特徵顯然可用以有效的鑑別出一段語音中純雜訊部分與語音部分，使之後所使用的頻譜消去法與靜音對數能量正規化法等強健性語音特徵技術，得以明顯提升在雜訊環境下語音辨識的精確度，增加語音辨識系統的強健性。

關鍵詞：端點偵測法，能量特徵，頻譜消去法，自動語音辨認

Abstract

The performance of a speech recognition system is often degraded due to the mismatch between the environments of development and application. One of the major sources that give rises to this mismatch is additive noise. The approaches for handling the problem of additive noise can be divided into three classes: speech enhancement, robust speech feature extraction, and compensation of speech models. In this thesis, we are focused on the second class, robust speech feature extraction.

The approaches of speech robust feature extraction are often together with the voice activity detection in order to estimate the noise characteristics. A voice activity detector (VAD) is used to discriminate the speech and noise-only portions within an utterance. This thesis

primarily investigates the effectiveness of various features for the VAD. These features include low-frequency spectral magnitude (LFSM), full-band spectral magnitude (FBSM), cumulative quantized spectrum (CQS) and high-pass log-energy (HPLE). The resulting VAD offers the noise information to two noise-robustness techniques, spectral subtraction (SS) and silence log-energy normalization (SLEN), in order to reduce the influence of additive noise in speech recognition.

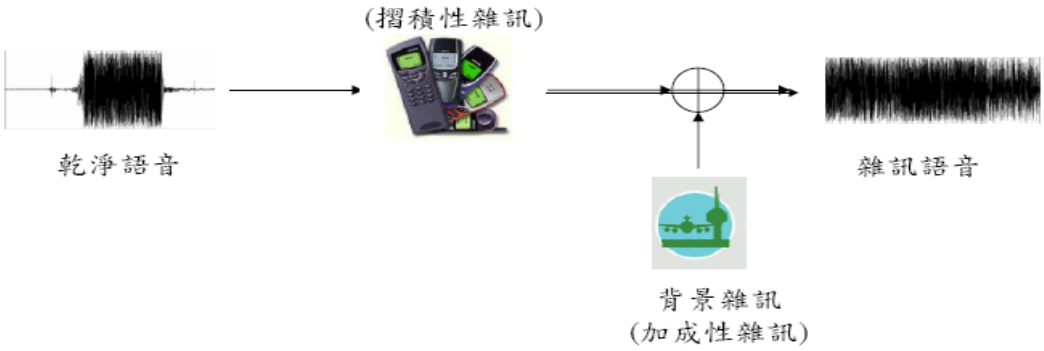
The recognition experiments are conducted on Aurora-2 database. Experimental results show that the proposed VAD is capable of providing accurate noise information, with which the following processes, SS and SLEN, significantly improve the speech recognition performance in various noise-corrupted environments. As a result, we confirm that an appropriate selection of features for VAD implicitly improves the noise robustness of a speech recognition system.

Keywords：voice activity detection, spectral magnitude, spectral subtraction, speech recognition

一、緒論

在我們的生活環境中，影響語音辨識結果的因素很多。其中一重要的因素爲語音辨識系統訓練與應用環境上的不匹配(environmental mismatch)，此不匹配的相關因素又包含了加成性雜訊(additive noise)、摺積性雜訊(convolutional noise)以及頻寬限制(bandwidth limitation) 等因素。其中加成性雜訊也可說是背景雜訊，這是因爲語音辨識系統所處在的環境，並非都像實驗室毫無其他干擾雜訊。也許系統是處於地鐵站中、餐廳、機場等這些具有其他干擾源的環境。甚至旁人呼吸喘息聲音都會混進語音裡面，造成辨識率的降低。摺積性雜訊也稱爲通道雜訊或是通道失真，主要是因爲麥克風的不同、傳輸線材的遮蔽效應不好而受外在電磁波影響所造成的。頻寬限制也是因爲收音通道的差異所帶來的影響。後面兩項的因素在電話語音辨識系統就非常的明顯。在有限頻寬的電話線，把通話者頻寬做限制以便利傳輸，這往往會造成語者的聲音變調，甚至通道失真的影響還會造成使用者兩端會發生吱吱的雜音，造成語音辨識很大的困擾。

爲了改善以上所述之環境上的不匹配，有眾多學者提出各種改進的方法，其中一類的方法爲強健性語音特徵參數技術，強健性語音特徵參數技術的主要目的是在抽取出不容易受到外在環境干擾而失真的語音特徵參數，進而突顯出語音變化的部分。在許多針對加成性雜訊所發展的強健性語音特徵參數技術裡，如何取得雜訊部分的資訊是很重要的，亦即在一段語音訊號中，我們通常必須偵測純雜訊所在的位置，以利於雜訊資訊的取得，其相關的技術則統稱語音活動偵測法(voice activity detection, VAD)，或簡稱爲端點偵測法(endpoint detection)。本論文的重點即在發展一系列使用端點偵測法的信號特徵，藉由這些特徵，使端點偵測的結果更精確，進而使之後的強健性語音特徵參數技術能達到更好的效果。在論文中。我們將介紹幾個用以語音偵測的信號特徵，分別爲低頻帶頻譜強度(low-frequency spectral magnitude, LFSM)、全頻帶頻譜強度(full-band spectral magnitude, FBSM)、累積量化頻譜(cumulative quantized spectrum, CQS)、以及高通對數能量(high-pass log-energy, HPLE)等。我們嘗試把這些端點偵測的方法與兩種強健式語音特徵參數擷取法結合，即頻譜消去法(spectral subtraction)與靜音對數能量正規化法(silence log-energy normalization, SLEN)等，發現皆有相當程度的提高辨識效率。



圖一、雜訊干擾語音之示意圖

本論文其餘部分共分爲五章，其中第二章詳細介紹所提出之端點偵測的各種信號特徵，第三章介紹本論文所用的兩種強健語音特徵技術，第四章爲實驗環境的設定，第五章爲實驗結果與討論，最後，第六章則包含了簡要的結論。

二、端點偵測所使用之信號特徵

端點偵測法(endpoint detection)或稱爲語音活動偵測法(voice activity detection, VAD)是指可以將一段語音中雜訊與語音的位置偵測出來的演算法。藉由一個有效的端點偵測法，我們可以利用所求得的純雜訊音框，準確的估測雜訊的資訊，例如其頻譜能量等。進而促成各種強健技術的使用，以達到雜訊的抑制，降低雜訊對語音訊號的影響。在以下幾節，我們將介紹用以端點偵測的各種信號特徵。

（一）低頻帶頻譜強度(low-frequency spectral magnitude, LFSM)

我們觀察到無論任何種類的雜訊，在頻帶[0,50Hz]之間都有相當比例的能量。同時，語音在此低頻帶的能量也具有一定程度的比例。因此我們根據此頻譜上的特性，去計算每個音框在此低頻帶的頻譜強度。根據此強度值，判斷此語音音框是否爲純雜訊音框，或是包含語音的音框。

首先，我們假設 $\{x_m[n], 1 \le n \le N\}$ 是語音訊號的第 m 個音框，將其取 K $(K \ge N)$ 點的離散傅立葉轉換，我們將可得到此音框所對應之頻譜如下式(1)：

$$X^{(m)}(f_k) = X^{(m)}[k] = \sum_{n=0}^{N-1} x_m[n] e^{-j\frac{2\pi nk}{K}}, \quad 0 \le k \le K-1 \tag{1}$$

其中 f_k 爲頻率，其值如下式：

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

其中 F_s 爲取樣頻率，因此我們定義出頻帶 $[F_L, F_U]$ 之頻譜強度計算方式爲：

$$Y^{(m)}_{[F_L,F_U]} = \sum_{F_L \le f_k \le F_U} \left| X_m(f_k) \right| \tag{3}$$

根據式(3)，我們可以計算每一音框之低頻帶頻譜強度，即 0 至 50Hz 以內的低頻帶頻譜強度如下：

$$Y^{(m)}_{Low} = Y^{(m)}_{[0,50]} = \sum_{0 \le f_k \le 50} \left| X_m(f_k) \right| \tag{4}$$

接著我們以一段語音前 P 個音框之低頻帶頻譜強度的平均爲參考值，其計算如下：

$$\theta = \lambda\left(\frac{1}{P}\sum_{m=1}^{P} Y_{Low}^{(m)}\right) \tag{5}$$

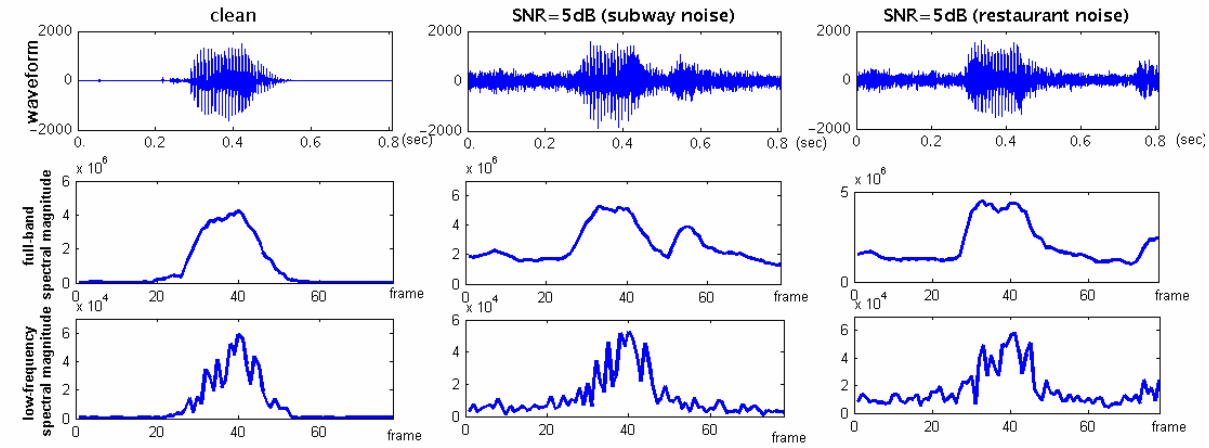
其中P表初步假設爲純雜訊之音框數。對於一段語音而言，前幾個音框通常是非語音的純雜訊音框，所以一開始我們假設前P個音框代表純雜訊。接著我們將每一個音框低頻帶內的頻譜強度$Y_{Low}^{(m)}$與門檻值θ做比較，若前者小於後者，則其歸類爲純雜訊音框，反之則爲語音音框。端點偵測的判斷式如下：

$$\text{第}m\text{個音框為}\begin{cases}\text{純雜訊音框，若 } Y_{Low}^{(m)} \le \theta \\ \text{語音音框 ，若 } Y_{Low}^{(m)} > \theta\end{cases} \tag{6}$$

（二）全頻帶頻譜強度(full-band spectral magnitude, FBSM)

利用全頻帶頻譜強度特徵進行端點偵測，其作法類似前一節，不同的是，我們並不針對特定頻帶來估測雜訊。假設取樣頻率爲 8kHz，我們是把每個音框的全部頻帶[0, 4kHz]的頻譜強度全部都考慮，如式(7)，我們計算每一音框全頻帶的頻譜強度：

$$Y_{Full}^{(m)} = Y_{[0,4000]}^{(m)} = \sum_{0 \le f_k \le 4000} \left|X_m(f_k)\right| \tag{7}$$



圖二、語音波形及頻譜強度分佈圖

圖二所示的三段語音，分別爲乾淨語音、受地下鐵雜訊干擾的語音(SNR=5dB)及受餐廳雜訊干擾的語音(SNR=5dB)。從語音波形及頻譜強度分佈圖看來，對於語音訊號我們可以很成功的區隔出語音的部分與非語音部分。但是在 SNR 很小的情況下，區隔語音和非語音部分變得比較困難。不同於[1]所提到，語音幾乎不分佈於[0,50Hz]的低頻帶，在第三列的低頻帶頻譜強度分佈圖中我們可看到，語音在此頻帶仍占有相當的比例，同時可看到此頻帶的頻譜強度可以有效用來區隔語音和非語音部分。觀察第二行的地鐵雜訊下的語音，大約在 0.5 到 0.6 秒之間所出現的較大能量之附加雜訊，但低頻帶頻譜強度幾乎不受其影響。而觀察第三行之餐廳雜訊下的語音，大約在 0.7 到 0.8 秒之間，也是出現頗像較大能量的附加雜訊，但我們發現，此附加雜訊在低頻帶頻譜強度的影響並不大。第二列的全頻帶頻譜強度可得正對純雜訊與語音音框也有很好的鑑別效果，爲其較容易受到較高頻譜強度影響而造成誤判。

（三）累積量化頻譜法(cumulative quantized spectrum, CQS)

在前兩小節裡，我們分別利用了低頻帶與全頻帶的頻譜強度來做爲端點偵測的標準，所用到的頻譜強度爲所用的頻帶內之每個頻率對應之強度總和。在這裡我們提出另一種利用頻譜性質來完成端點偵測的方法，所用的特徵稱爲累積量化頻譜(cumulative quantized spectrum, CQS)。其基本論點爲，將純雜訊與雜訊語音之離散頻譜比較，我們可以發現純雜訊音框所含較高強度之頻率個數，通常比雜訊語音音框之高強度的頻率個數少，因此我們可以藉由累積一音框中高強度之頻率的個數所得，來判斷此音框的種類。因爲這樣的做法，相當於將每個頻率的強度做量化(高者爲 1，低者爲 0)，再將這些量化後的強度值加總，因此我們稱所得之特徵值爲累積量化頻譜。我們假設每段語音的前 P 個音框爲純雜訊音框，利用此 P 個音框之強度頻譜的平均，我們定義每一頻率之強度的門檻值爲：

$$\theta(k)=\frac{1}{P}\sum_{m=1}^{P}\left|X_m(k)\right| \tag{8}$$

其中 $X_m(k)$ 表示第 m 個音框訊號取 N 點 DFT 後之頻譜。我們將每一個音框之離散頻譜強度 $\left\{\left|X_m(k)\right|,\ k=0,\ 1,\ 2,\ .....,\ \frac{N}{2}\right\}$ 與所得之門檻值 $\left\{\theta(k),\ k=0,\ 1,\ 2,.....,\ \frac{N}{2}\right\}$ 比較大小，所得到量化後的頻譜如下：

$$Y_m(k)=\begin{cases}1 & if\ \ \left|X_m(k)\right|>\theta(k)\\ 0 & if\ \ \left|X_m(k)\right|\le\theta(k)\end{cases}\quad k=0,\ 1,\ 2,.....,\ \frac{N}{2} \tag{9}$$

接著將每一音框的量化頻譜 $\left\{Y_m(k)\right\}$ 做累加，得：

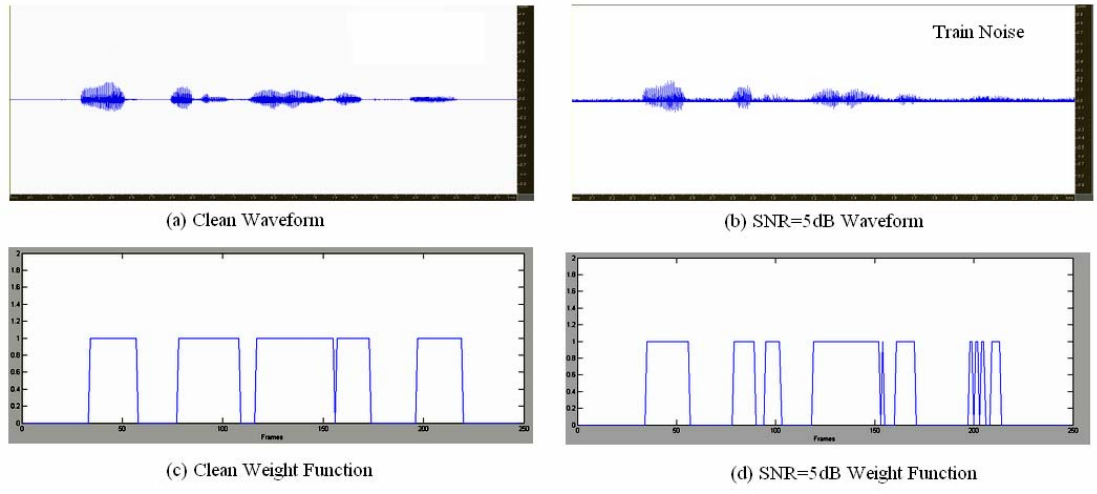
$$Z_m=\sum_{k=0}^{\frac{N}{2}}Y_m(k) \tag{10}$$

因此所得之累積量化頻譜 Z_m 即爲第 m 個音框中，強度大於門檻值的頻率個數，Z_m 越大，則意味著此音框有越大的機率是能量較大的音框(語音音框)；反之，則此音框爲一純雜訊音框。最後我們爲累積量化頻譜 Z_m 定一門檻值 $\frac{N}{4}$，即約爲頻率點數的一半，當 Z_m 小於 $\frac{N}{4}$ 時，則此音框歸類爲純雜訊音框；反之，則爲語音音框：

$$\text{第}m\text{個音框為}\begin{cases}\text{語音音框，若 } Z_m\ge\frac{N}{4}\\ \text{純雜訊音框，若 } Z_m<\frac{N}{4}\end{cases} \tag{11}$$

圖三所示的兩段語音，分別爲乾淨語音與受車站雜訊干擾的語音(SNR=5dB)。第一列兩圖爲語音的波形圖(圖三(a)與(b))，而第二列兩圖(圖三(c)與(d))意義爲音框判定的結果，若高強度之頻率的個數占多數，則爲雜訊語音音框，在圖中以 1 表示；反之，若高強度之頻率個數占少數，則此音框判定爲雜訊音框，圖中以 0 表示。在第一行的兩個組圖(圖七(a)與(c))，我們發現以累積量化頻譜分佈對於乾淨的語音中，對於語音音框與非語音音框，具有著很良好的辨別度。而在於 SNR 較小的語音環境(圖三(b)與(d))，在訊號較後段部分，其語音成分幾乎要被背景雜訊給覆蓋過去，但是經過累積量化頻譜做判定處理後，我們可以發現其對於鑑別出純雜訊部分以及雜訊語音部分比起完全乾淨的語音判定結

果(圖三(c))只是略差一些，對於語音部分整體來說並沒有太大的遺漏或是誤判。



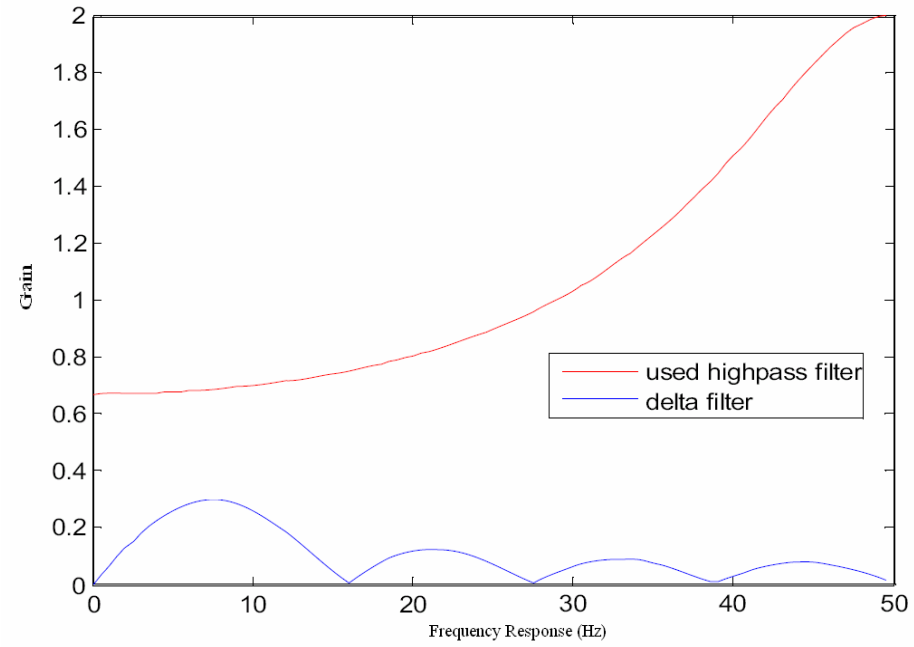
圖三、語音波形及累積量化頻譜法之權重圖

（四）高通對數能量(high-pass log-energy, HPLE)

能量大小一向是判定一音框是否爲雜訊或語音的重要指標，在[2-6]等諸多文獻裡提到音框能量或其他變化的形式(如對數能量，能量的差分等)可用作端點偵測的主要特徵。根據我們的觀察發現，將音框對數能量值通過一高通的時間序列濾波器，所得到之高通對數能量，可有效鑑別純雜訊與語音音框。相對於在一般使用之能量差分法中，差量濾波器(delta filter)捨棄能量之調變頻譜高頻成分，我們發現其實高頻成分也是包含著很多重要的語音資訊。因此我們所使用一高通時間序列濾波器來對能量做處理：我們做一無限脈衝響應(infinite impulse response, IIR)的高通時間序列濾波器來處理一連串的音框對數能量，其輸入與輸出的關係式爲：

$$E[n] = \frac{1}{2}\left(e[n+1] - E[n-1]\right) \qquad (12)$$

其中$E[n]$是每個音框更新後的對數能量值，而$e[n]$爲每個音框的原始對數能量值。此高通濾波器之頻率響應如圖三。我們從圖中可知道，此高通時間序列濾波器並沒特別抑制低頻部分，而在高頻部分卻有著放大的效果。



圖三、高通濾波器與差量濾波器振幅頻率響應圖

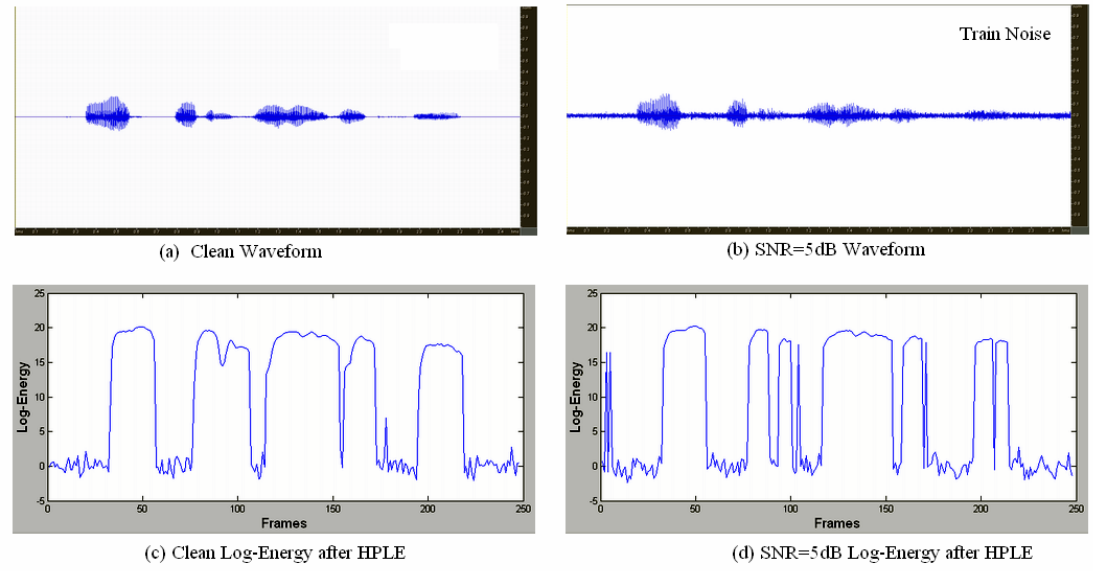
我們令整段語音音框高通對數能量的平均爲一門檻值，計算如下：

$$T = \frac{1}{N}\sum_{n=1}^{N} E[n] \tag{13}$$

其中N爲音框總數。語音音框與純雜訊音框判定的原則爲：

$$\text{第}n\text{個音框為一}\begin{cases}\text{純雜訊音框，若} & E[n] < T \\ \text{語音音框，若} & E[n] \geq T\end{cases} \tag{14}$$

圖四爲乾淨語音與雜訊語音之波形與高通對數能量圖，圖四(a)與(c)表示出，在乾淨情況下，高通對數能量法對於非語音與語音具有很不錯的鑑別力。而當我們在 SNR=5dB 時的雜訊環境，我們由圖四(b)與(d)發現，整段訊號末端的語音部分，幾乎被雜訊給覆蓋過去，然而經過高通對數能量法處理過後，這部分的語音有被鑑別出來，惟一誤判部分是在剛開始的一小部分。大體來說，整體語音部分幾乎都有被鑑別出來，因此可看出此方法的效能。



圖四、語音波形與其高通對數能量圖

三、強健式語音特徵參數擷取技術

本章節中，我們將介紹兩種強健性語音特徵擷取的方法，將語音中估測的雜訊消除，以還原乾淨的語音特徵。

（一）非線性頻譜消去法 (nonlinear spectral subtraction, NSS)

頻譜消去法[7-8]主要目的是估測出加成性雜訊的頻譜分佈，再將被附加雜訊干擾的語音訊號頻譜扣除所估測出的雜訊頻譜，以還原原始的乾淨語音頻譜。使用頻譜消去法中，有兩個基本假設：

（1）乾淨語音訊號與雜訊訊號在統計上是無關的(uncorrelated)，並且在時域 (time domain) 上是可線性加成的；

(2) 雜訊訊號相對乾淨語音訊號而言是變化較爲緩慢的。

根據這兩個假設，如果我們要得到乾淨語音訊號，通常必須從非語音的區域估測出雜訊的頻譜，再將受雜訊干擾的語音頻譜減去雜訊的頻譜，式(15)說明了雜訊語音訊號、乾

淨語音訊號和雜訊訊號的關係：

$$y_i(t) = x_i(t) + n_i(t) \xrightarrow{\text{Fourier Transform}} Y_i(f) \approx X_i(f) + N_i(f) \tag{15}$$

其中$y_i(t)$、$x_i(t)$與$n_i(t)$分別代表第i個音框的雜訊語音訊號、乾淨語音訊號以及雜訊訊號，而$Y_i(f)$、$X_i(f)$與$N_i(f)$則是$y_i(t)$、$x_i(t)$與$n_i(t)$的強度頻譜(magnitude spectrum)值。

因此，理想上，我們若能精確得到雜訊之強度頻譜$N_i(f)$，則可從$Y_i(f)$直接扣除$N_i(f)$而得到乾淨語音強度頻譜$X_i(f)$。實際上，$N_i(f)$通常無法十分精確的估測，這會導致一個問題：若強度頻譜$Y_i(f)$比估測之雜訊訊號強度頻譜$N_i(f)$還小時，相減的結果會得到一個負值，而乾淨語音強度頻譜是不應該出現負值的，因此解決這問題的方法之一就是當估計到的乾淨語音強度頻譜值爲負時，我們就以一個極小的值代替，這樣的方法我們稱爲非線性頻譜消去法[8]，如式(16)所示：

$$X_i(f) = \begin{cases} Y_i(f) - \alpha N(f) & if \;\; Y_i(f) > N(f) \\ \beta Y_i(f) & otherwise \end{cases} \tag{16}$$

其中$N(f)$是估測而得的雜訊強度頻譜，α是過度估測因子(over-estimation factor)，用來控制估測雜訊功率頻譜被減去的程度，而β是底限因子(flooring factor)。在本論文中，$N(f)$是利用前一章之端點偵測法所得之純雜訊音框強度頻譜的平均，公式如下：

$$N(f) = \frac{1}{M}\sum_{j=1}^{M} N_j(f) \tag{17}$$

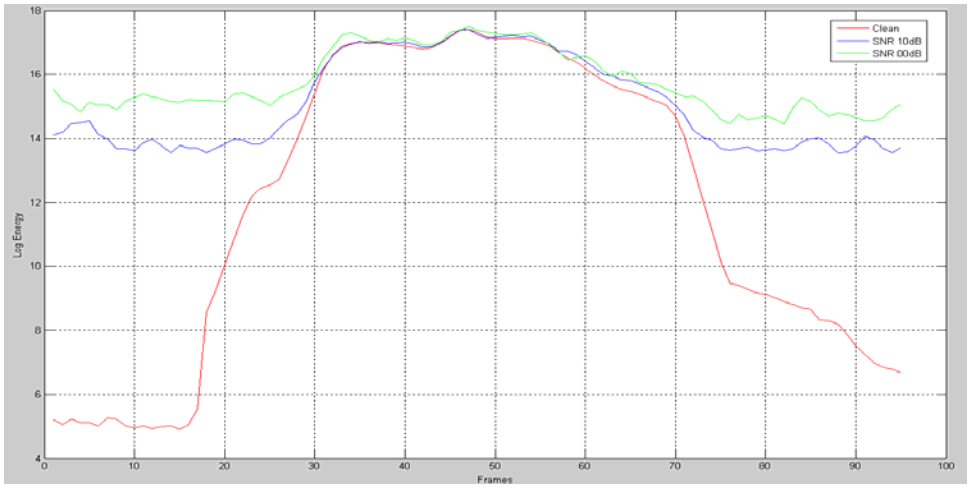
其中$N_j(f)$爲第j個被標示爲純雜訊音框之強度頻譜，M爲純雜訊音框總數。

（二）靜音對數能量正規化法(silence log-energy normalization, SLEN)

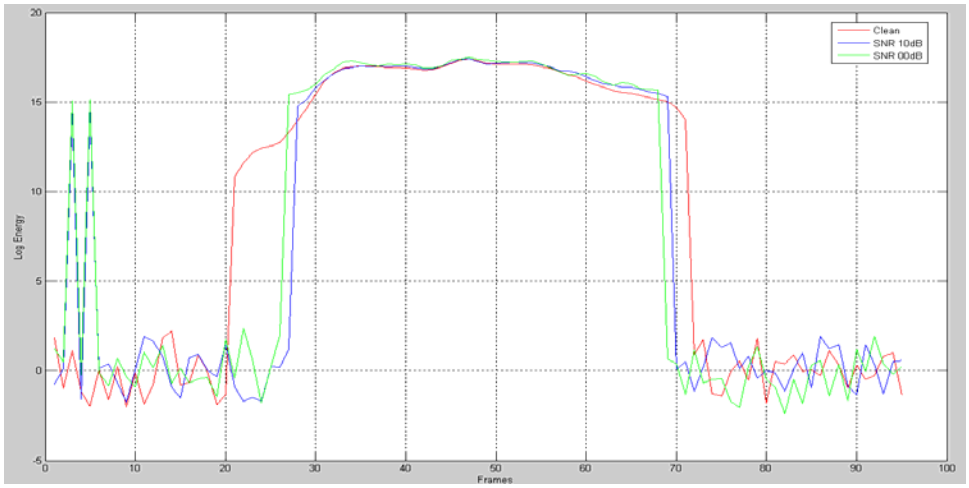
靜音對數能量正規化法[9]的原理在於觀察發現，在雜訊語音的能量曲線上，受雜訊干擾較爲嚴重部分爲能量小的波谷，而能量大的波峰部分比較不受影響。而且許多實驗發現，只保留能量波峰部分的音框，而捨棄能量小的音框，經由辨識依然可以得到不錯的辨識率。此外我們假設，對能量特徵而言，最重要的是原始語音能量的變化曲線之波形，是否被完整保留。也就是說一段語音整體的能量波形，比單一音框之能量的降低失真距離重要。若是保留原始整體能量波形的完整，即使原始語音能量曲線波形有小幅度的位移，最後的辨識效果也不會相差太多。依照上述的原理，我們找出能量曲線波形中非語音的部分，並且把它正規化處理過後，乾淨的語音訊號能量曲線波形將會與受雜訊影響的語音能量曲線波形十分的相似。根據四之一節的各種端點偵測法，我們將判定爲純雜訊音框之對數能量正規化爲一極小值，而語音音框的對數能量則維持不變：

$$\hat{E}[n] = \begin{cases} E[n] & \text{若第}n\text{個音框為語音音框} \\ \varepsilon & \text{若第}n\text{個音框為純雜訊音框} \end{cases} \tag{18}$$

其中ε爲一極小值。我們利用圖五來觀察靜音音框對數能量正規化法的作用。



圖五(a) 一段語音之原始對數能量軌跡圖



圖五(b) 一段語音經靜音音框對數能量正規化法後之對數能量特徵軌跡圖

由圖五可以很明顯發現，雜訊語音(SNR=10dB，SNR=0dB)之兩條對數能量曲線在較低值處與乾淨能量的曲線有很大的不匹配(圖五(a))，經由靜音對數能量正規化之後，此不匹配的現象改善了許多(圖五(b))。因此，靜音音框對數能量正規化法能有效降低受雜訊污染的語音與乾淨語音在能量上之不匹配，減少雜訊對語音能量的干擾。

五、實驗設定

(一) 語音資料庫簡介

本論文所使用的語音資料庫爲歐洲電信標準協會(European Telecommunication Standard Institute，ETSI)發行的 AURORA2 語音資料庫[10]，它是一套連續的英文數字字串，內容是以美國成年男女所錄製的乾淨環境連續數字，再加上雜訊與通道效應。加成性雜訊共有八種，分別爲地下鐵、人聲、汽車、展覽館、餐廳、街道、機場、火車站等，前四種歸類爲 Set A，後四種歸類爲 Set B。其訊雜比(signal-to-noise ratio, SNR)則有七種，分別爲 20dB, 15dB, 10dB, 5dB, 0dB. -5dB 與完全乾淨狀態。

(二) 特徵參數的設定與辨識系統的訓練

在本論文的語音辨識實驗中，我們使用的特徵參數包含了12維梅爾倒頻譜參數與1維對數能量，附加上其一階與二階差量（在部分實驗中，我們會省卻對數能量部分）。

特徵參數抽取之詳細設定，如表一所示。對於每個欲辨識的數字模型而言，本論文使用隱藏式馬可夫模型工具(hidden Markov model toolkit，HTK)來訓練，包含11個數字模型(0~9以及oh 11個數字模型)以及靜音模型，每個數字模型包含10個狀態，各狀態包含4個高斯密度混合。隱藏式馬可夫模型是一種運用統計理論推導出來的模型，用來描述語音產生的過程，相當適合用在連續語音的辨認。HMM有很多種類型，本論文採用由左到右的形式，也就是每個狀態在下一個時間只能跳到此刻狀態或下一個鄰近的狀態，隨著時間的增加，狀態由左至右依序轉移。另外，模型中的狀態觀測機率函數是選用連續式的高斯混合機率密度函數(Gaussian Mixture probability density function，簡稱GM)，因此我們也稱此模型爲連續密度隱藏式馬可夫模型(continuous density HMM，簡稱CDHMM)。

表一 本論文實驗所使用特徵參數抽取之設定

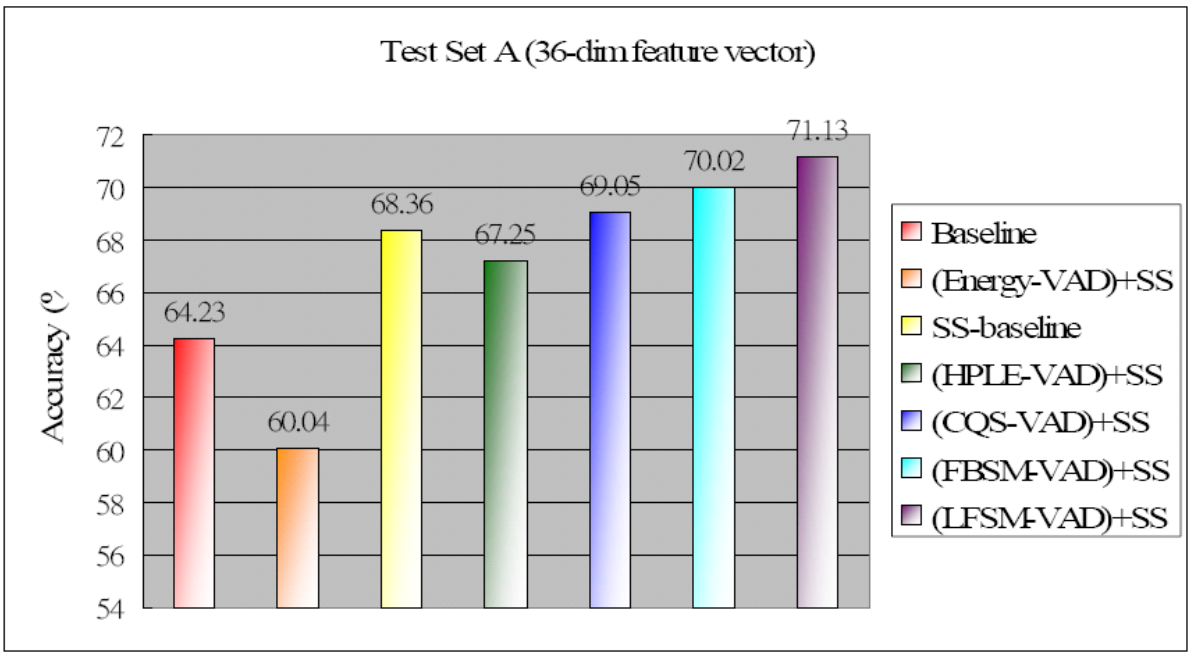
取樣頻率	8000 Hz
音框長度(frame size)	25 ms
音框平移(frame shift)	10 ms
預強調濾波器	1-(0.97)z^{-1}
視窗形式	漢明窗(Hamming window)
快速傅立葉轉換點數	256 點
濾波器組	梅爾刻度三角濾波器組(Mel-scaled triangular filter bank)，共 23 個濾波器
語音特徵參數	13 維 MFCCs(含對數能量)+Δ13 維 MFCCs +ΔΔ13 維 MFCCs，共 39 維

五、 端點偵測法配合強健性語音技術之實驗結果

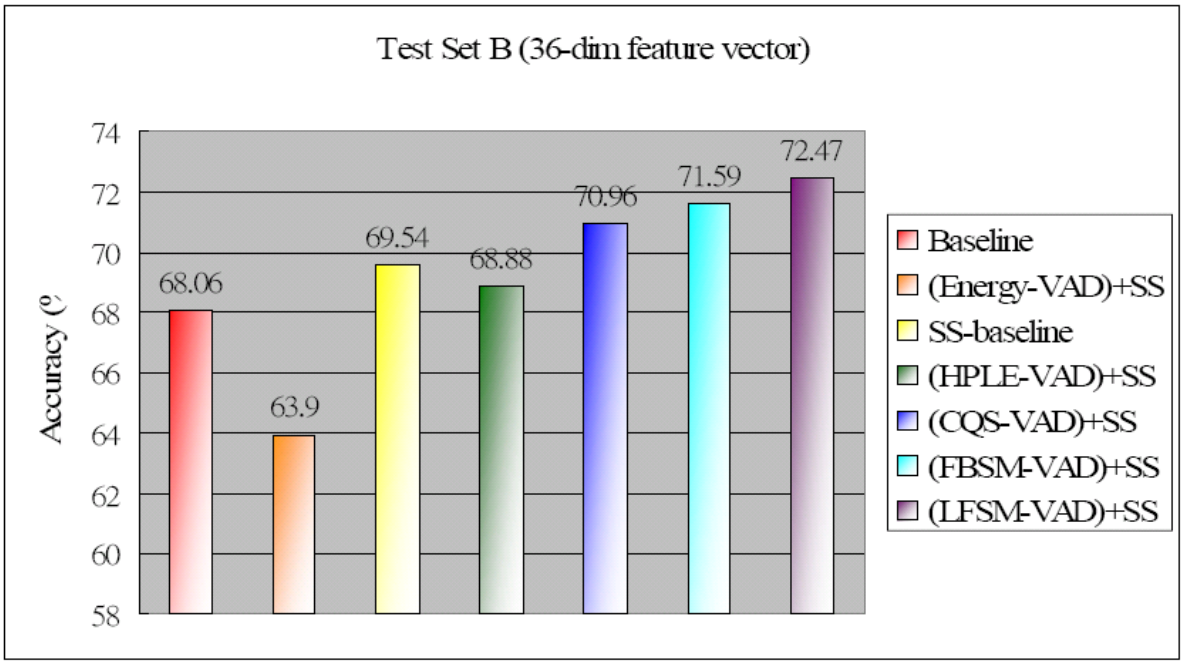
本章將會介紹第四章所有端點偵測法配合強健式語音的辨識實驗結果。我們將實驗區分爲「省略能量維特徵參數」以及「加入能量維特徵參數」這兩種，藉此分析能量維對於語音辨識上的影響。實驗結果可以證明本論文所提到之方法幾乎都可提升雜訊環境下語音辨識率，降低雜訊對語音的干擾。

（一）梅爾倒頻譜特徵參數之實驗結果

本章節所有實驗所使用的特徵參數爲 12 維梅爾倒頻譜參數及其一階和二階差量，總共爲 36 維特徵參數。圖六與圖七分別爲 A 組環境與 B 組環境下各種方法所得之平均辨識率。其中「baseline」是指沒有處理過的原始特徵參數、「SS-baseline」是指利用每段語句前五個音框作爲純雜訊音框所作的頻譜消去法，「(Energy-VAD)+SS」、「(HPLE-VAD)+SS」、「(CQS-VAD)+SS」、「(FBSM-VAD)+SS」與「(LFLE-VAD)+SS」則分別爲以音框能量、高通對數能量、累積量化頻譜、全頻帶頻譜強度與低頻帶頻譜強度作爲端點偵測特徵，執行端點偵測並與頻譜消去法作結合。



圖六、A 組環境下平均辨識率(%)比較圖



圖七、B 組環境下平均辨識率(%)比較圖

從圖六與圖七的辨識結果，我們有以下幾點的發現：

1. 以傳統的能量特徵作爲端點偵測的信號特徵，其效果不盡理想，其得到的端點偵測之結果配合頻譜消去法，所得到的辨識率甚至比基礎實驗還差。
2. 未作端點偵測而純粹以每段語句前 5 個音框爲純雜訊所作之頻譜消去法，相較於基礎實驗約可得到 2-4%的平均進步率。
3. 本論文所提出的四種端點偵測的特徵(HPLE, CQS, FBSM 與 LFSM)應用於端點偵測，配合頻譜消去法之下，都能得到明顯的進步。其中以高通對數能量(HPLE)表現稍差，但仍比傳統之能量特徵來的好。而這四種特徵又以低頻帶頻譜強度(LFSM)表現最

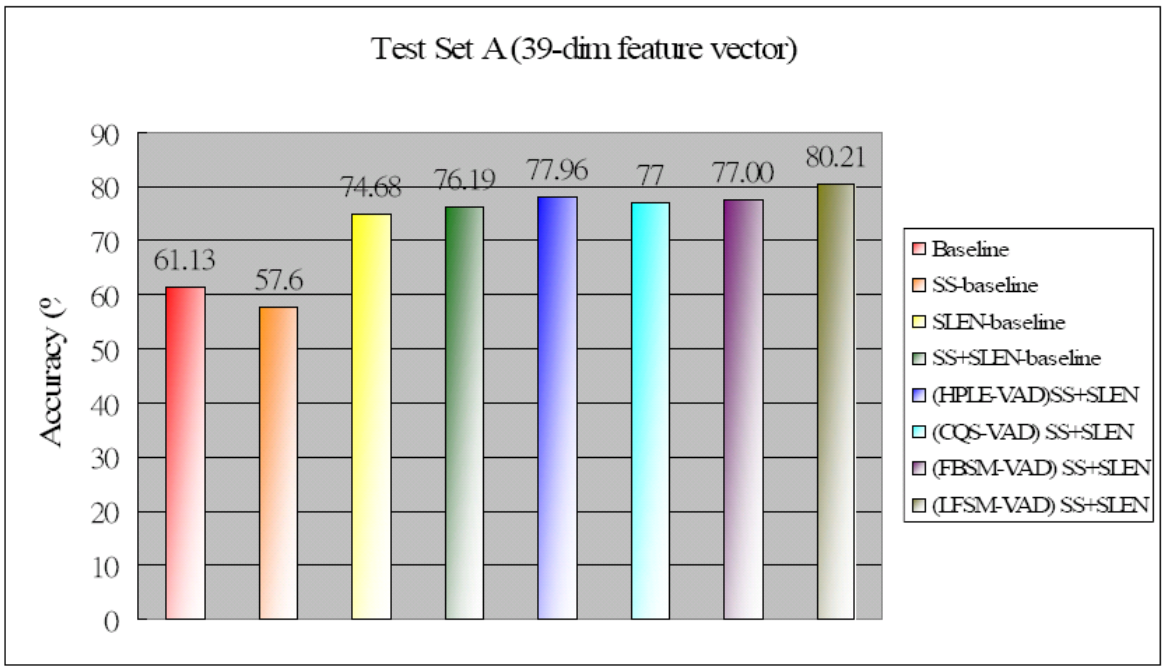
好，相較於基礎實驗約可得到 4-7%的平均進步率。

（二）梅爾倒頻譜系數與對數能量之特徵參數的實驗結果

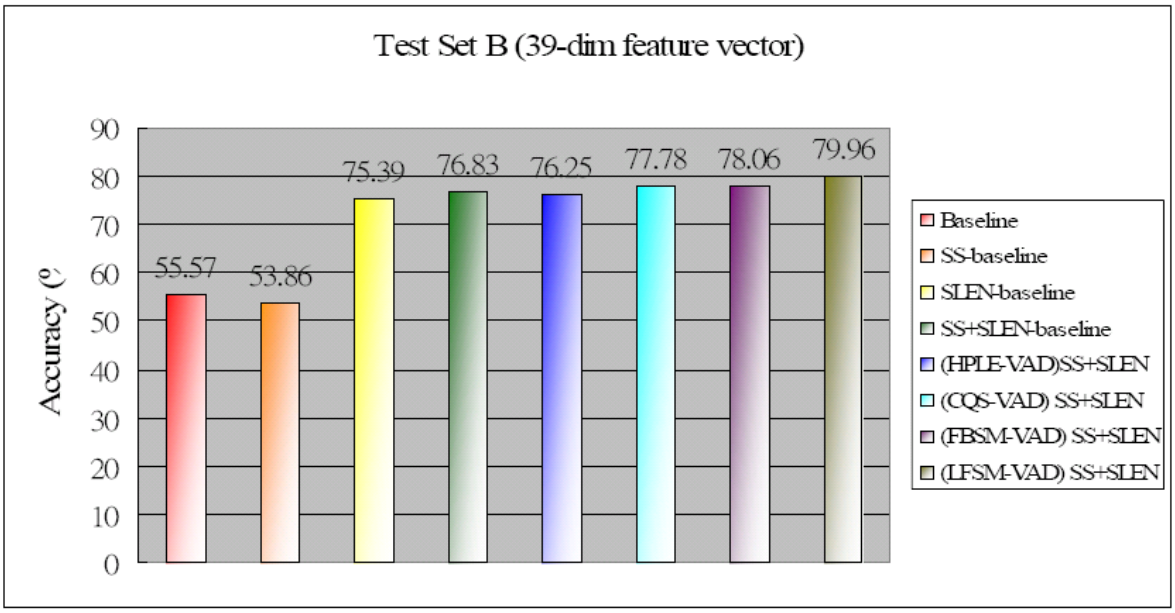
本章節所有實驗所使用的特徵參數爲 12 維梅爾倒頻譜參數與 1 維對數能量，附加其一階和二階差量，總共爲 39 維特徵參數。圖八與圖九分別爲 A 組環境與 B 組環境下各種方法所得之平均辨識率。其中「baseline」是指沒有處理過的原始特徵參數、「SS-baseline」、「SLEN-baseline」與「SS+SLEN baseline」是指利用每段語句前五個音框作爲純雜訊音框分別作頻譜消去法(SS)、靜音音框對數能量正規化法(SLEN)及 SS 和 SLEN 的結合，「(HPLE-VAD) SS+SLEN」、「(CQS-VAD) SS+SLEN」、「(FBSM-VAD) SS+SLEN」與「(LFLE-VAD) SS+SLEN」則分別爲以音框能量、高通對數能量、累積量化頻譜、全頻帶頻譜強度與低頻帶頻譜強度作爲端點偵測特徵，執行端點偵測，再使用頻譜消去法與靜音音框對數能量正規化法。

從圖八與圖九的辨識結果，我們有以下幾點的發現：

1. 相較於前一節的基本實驗，本節的特徵參數額外引進了對數能量及其一階與二階差量，然而其基本實驗辨識率反而較未引進這三個能量參數的結果還來的差，辨識率降低了 3%至 7.8%。這可能是因爲，雖然能量特徵它包含了很多語音的資訊，但是相對的它也深受雜訊的影響，反而不利於系統辨識。
2. 未作端點偵測而純粹以每段語句前 5 個音框爲純雜訊所作之頻譜消去法，相較於基礎實驗結果反而退步了約 3%，其可能原因如同前述，即能量特徵的失真所帶來的嚴重影響。
3. 未作端點偵測而純粹以每段語句前 5 個音框爲純雜訊所作之靜音音框對數能量正規化法，在辨識率上有十分顯著的提升，相較於基礎實驗結果提升了 13%至 20%之多。這代表了在雜訊環境下，能量特徵強健性處理的重要性，也凸顯了靜音音框對數能量正規化法的明顯效能。
4. 未作端點偵測而純粹以每段語句前 5 個音框爲純雜訊，同時執行頻譜消去法與靜音音框對數能量正規化法，其辨識率比單純使用靜音音框對數能量正規化法可以再進步 1%-2%左右。
5. 本論文所提出的四種端點偵測的特徵(HPLE, CQS, FBSM 與 LFSM)應用於端點偵測，配合頻譜消去法與靜音音框對數能量正規化法，都能得到明顯的進步。其中除了高通對數能量(HPLE)在 Set B 環境下稍微退步外，其他情形下皆比未做端點偵測的結果進步 1%以上。類似前一節，這四種特徵又以低頻帶頻譜強度(LFSM)表現最好，相較於未做端點偵測的結果，可得到 3%-4%的平均進步率。



圖八、A 組環境下平均辨識率(%)比較圖



圖九、B 組環境下平均辨識率(%)比較圖

六、結論

在本論文中，我們提出了幾種端點偵測所用的信號特徵，包括低頻帶頻譜強度、全頻帶頻譜強度、根據頻譜強度分佈的累積量化頻譜、及根據能量調變頻譜特性的高通對數能量。其目的是偵測出純雜訊音框，進而結合頻譜消去法與靜音對數能量正規化法，達到強健語音消除雜訊的功能。其中低頻域頻譜強度之端點偵測法配合頻譜消去法與靜音對數能量正規化法，可以最有效地提升雜訊環境下的辨識率。而其他三種信號特徵也有不錯的端點偵測效果。由比較中可得到，這幾個方法比起基本實驗，都有著顯著的進步。

此外，我們發現能量維特徵蘊藏著很多語音鑑別資訊，但相對而言，雜訊對其影響也很大。但是經過靜音對數能量正規化法處理，可以明顯提升辨識效果。此外，低頻帶頻譜強度之端點偵測法配合頻譜消去法與靜音對數能量正規化法可達到平均 80%的優異表現。由實驗結果也發現，八種雜訊環境在雜訊比 0~20dB 的條件下，每個辨識率都很接近，即表示此方法不受雜訊型態的影響。在穩定與非穩定雜訊都有很好的辨識率。

參考文獻

[1] K. Yamashita,; T. Shimamura; " Nonstationary noise estimation using low-frequency regions for spectral subtraction", Signal Processing Letters, IEEE Volume 12, Issue 6, June 2005

[2] Tai-Hwei Hwang, “Energy Contour Extraction for In-Car Speech Recognition”, 9th European Conference on Speech Communication and Technology (Eurospeech 2003)

[3] Weizhong Zhu and Douglas O’Shaughnessy “Log-energy Dynamic Range Normalization for Robust Speech Recognition”, 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)

[4] Tai-Hwei Hwang and Sen-Chin Chang, “Energy Contour Enhancement for Noisy Speech Recognition”, 2004 International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)

[5] M. Ahadi, H. Sheikhzadeh, R. Brennan and G. Freeman, “An Energy Normalization Scheme for Improved Robustness in Speech Recognition” 8th International Conference on Spoken Language Processing (ICSLP 2004)

[6] R. Chengalvarayan, “Robust Energy Normalization Using Speech/nonspeech Discriminator for German Connected Digit Recognition”, 6th European Conference on Speech Communication and Technology (Eurospeech 1999)

[7] S.F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction” IEEE Trans. on Acoustics, speech, and Processing, VOL. ASSP-27, NO. 2, April 1979

[8] P. Lockwood and J. Boudy, "Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the Projection, for Robust Speech Recognition in Cars", Eurospeech 1991

[9] C.F. Tai, J.W. Hung, "Silence Energy Normalization for Robust Speech Recognition in Additive Noise Environments", INTERSPEECH 2006 – ICSLP

[10] H.-G Hirsch, D. Pearce, “The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” ISCA ITRW ASR 2000, Paris, France, September 18-20, 2000