強健性語音辨識中能量相關特徵之改良式正規化技術的研究

Study of the Improved Normalization Techniques of Energy-Related Features for Robust Speech Recognition

潘吉安 Chi-an Pan
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University, Taiwan
s95323544@ncnu.edu.tw

杜文祥 Wen-hsiang Tu
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University, Taiwan
aero3016@ms45.hinet.net

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

摘要

隨著科技的發展，自動語音辨識技術也逐漸成熟，而達實際應用的階段；但當一自動語音辨識系統使用於現實環境中時，往往會受到雜訊的干擾，而造成辨識率大幅的下降；因此，環境相關的語音強健技術就顯得格外重要。本論文是針對在加成性雜訊所造成的辨識系統之訓練與辨識環境不匹配作爲主要探討的課題，除了概述許多語音特徵之強健性處理技術外，主要重點在於介紹我們所新發展的能量相關特徵強健化演算法—靜音特徵正規化法。在此，我們以較嚴謹的數學分析，探討加成性雜訊對能量相關特徵造成的失真現象；接著根據這些現象，我們發展相對應的一套新技術，即靜音特徵正規化法，來降低這些失真。透過這一系列的辨識實驗，證實我們所提出的新技術能夠有效提升各種加成性雜訊環境下的語音辨識率，並與其它許多強健性技術有良好的加成性。

Abstract

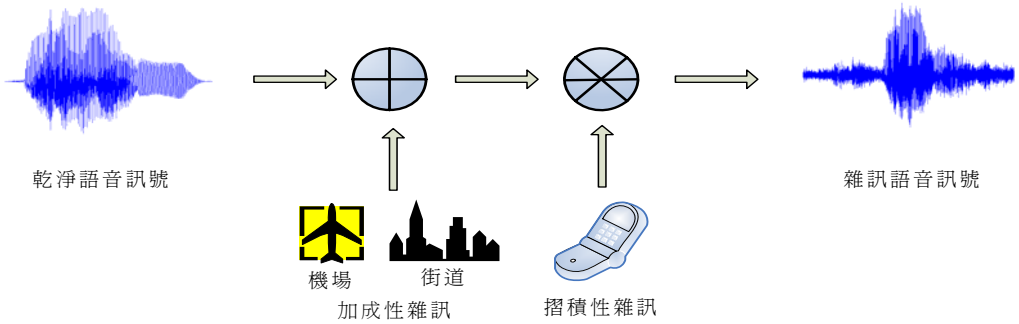
The rapid development of speech processing techniques has made themselves successfully applied in more and more applications, such as automatic dialing, voice-based information retrieval, and identity authentication. However, some unexpected variations in speech signals deteriorate the performance of a speech processing system, and thus relatively limit its application range. Among these variations, the environmental mismatch caused by the embedded noise in the speech signal is the major concern of this paper. In this paper, we provide a more rigorous mathematical analysis for the effects of the additive noise on two energy-related speech features, i.e. the logarithmic energy (log$E$) and the zeroth cepstral coefficient ($c0$). Then based on these effects, we propose a new feature compensation scheme, named silence feature normalization (SFN), in order to improve the noise robustness of the above two features for speech recognition. It is shown that, regardless of its simplicity in implementation, SFN brings about very significant improvement in noisy speech recognition, and it behaves better than many well-known feature normalization approaches. Furthermore,

SFN can be easily integrated with other noise robustness techniques to achieve an even better recognition accuracy.

關鍵詞：自動語音辨識、對數能量特徵、第零維倒頻譜特徵係數、強健性語音特徵
Keywords: speech recognition, logarithmic energy feature, the zeroth cepstral coefficient, robust speech features

一、緒論

近年來科技發展迅速，但是自動語音辨識仍然是一門相當具有挑戰性的課題。通常一自動語音辨識系統在不受外在雜訊干擾的研究室環境下，都可以獲得極高的辨識效能，但若是應用到實際的環境中，系統辨識效能則通常會大幅降低，這主要是被現實環境中許多的變異性(variation)所影響。而語音辨識的變異性種類繁多，例如訓練環境與測試環境間存在的環境不匹配(environmental mismatch)、語者變異(speaker variation)以及發音的變異(pronunciation variation)等。對於環境不匹配而言，其相關的變數可概略分爲下列幾項類型：加成性雜訊(additive noise)、摺積性雜訊(convolutional noise)以及頻寬的限制(bandwidth limitation)等。圖一爲乾淨語音訊號受到雜訊干擾之示意圖。



圖一、乾淨語音受雜訊干擾之示意圖

本論文是以上述所提及的環境不匹配中的加成性雜訊因素，作爲主要探討的主題，以期將加成性雜訊對語音辨識的影響降低。在特徵參數抽取步驟時，我們經常計算語音的能量值作爲特徵之一；根據過去的文獻指出[1][2]，語音訊號的能量特徵(energy feature)所包含的辨識資訊大過於其它特徵，且能量特徵的計算複雜度很低。所以根據上述能量特徵的優勢，在本論文中，我們特別對其強健性技術加以分析、討論與發展。

近年來，有許多成功的強健性對數能量特徵(logarithmic energy, log$E$)的技術相繼被提出，例如，對數能量動態範圍正規化法(log-energy dynamic range normalization, LEDRN)[3]其目標是使訓練與測試的語音資料其對數能量值之動態範圍一致化；對數能量尺度重刻法(log-energy rescaling normalization, LERN)[4]則是將對數能量特徵乘上一個介於 0 與 1 間的權重值，試圖重建出乾淨語音的對數能量特徵；而本實驗室先前所提出的靜音音框對數能量正規化法(silence energy normalization, SLEN)[5]，是將判別爲非語音音框(non-speech frame)的對數能量特徵設定爲一極小值的常數。上述的三種方法，皆傾向於將非語音部分的對數能量數值調低，並將語音部分的對數能量值保持不變；其主要的原因是一段語音特徵中，能量較低的部分通常會比能量較高的部分更容易受到雜訊的影響。本論文依據前人所發表的文獻加以改進，且針對語音訊號能量相關的特徵如何受到雜訊影響，以較嚴謹的數學理論加以分析，並提出一套新的強健技術，稱爲「靜音特徵正規化法」(silence feature normalization, SFN)，此方法可以有效地降低加成性雜訊對語音能量相關特徵的干擾，進而提高系統的辨識效能。

本論文其它章節概要如下：在第二章中，我們先主要將對能量相關特徵受雜訊影響的效應，做進一步的分析與探討，接著介紹本論文所新提出的之靜音特徵正規化法(SFN)；第三章包含了各種針對能量相關特徵之處理技術的語音辨識實驗數據及相關討論，其中除了介紹語音辨識實驗環境外，主要是評估靜音特徵正規化法的效能，並與其他方法作比較，藉此驗證我們所提出新方法能有效提升能量相關特徵在雜訊環境下的強健性。在第四章中，我們嘗試將所提的新方法結合其它的強健性特徵技術，對此類的結合作辨識實驗所得到的辨識率加以探討與分析，以驗證我們所提出的靜音特徵正規化法是否與其它技術有良好的加成性。第五章則爲本論文結論與未來展望。

二、靜音特徵正規化法

首先，我們在第一節中，針對語音能量相關特徵：對數能量(logarithmic energy, log*E*)與第零維倒頻譜係數(*c*0)受到環境雜訊干擾的變異現象做較深入的觀察分析與探討，接著在第二節中，我們根據這些結果，提出靜音特徵正規化法的新強健性技術。

（一）對數能量特徵及第零維倒頻譜特徵係數受加成性雜訊干擾之現象的探討

加成性雜訊對於能量相關特徵(log*E* 與 *c*0)造成的效應可由圖二看出端倪。圖二(a)、(b)與(c)分別表示一乾淨語音訊號(Aurora-2.0 資料庫中的"MAH_1390A"檔)的波形圖、對數能量(log*E*)曲線圖與第零維倒頻譜特徵係數(*c*0)曲線圖；而(b)與(c)中紅色實線、綠色虛線與藍色點線則分別爲乾淨語音、訊雜比 15dB 的語音及訊雜比 5dB 的語音所對應的曲線。由這三張圖中，可以很明顯地看出，在有語音存在的區域，log*E* 與 *c*0 特徵值較大，較不容易受到雜訊的影響而失真，而且隨時間上下振盪的情況較爲明顯；反之，在無語音存在的區段，其特徵值前後變化較平緩，且受到雜訊的干擾後，其值會很明顯地被改變許多。接下來，我們就以較嚴謹的數學理論，對以上兩種失真現象加以分析與探討。首先，我們探討加成性雜訊對於 log*E* 特徵的影響。假設一段受加成性雜訊干擾的語音(noisy speech)中，第 n 個音框的訊號 $x_n[m]$ 可表示爲：

$$x_n[m] = s_n[m] + d_n[m] \text{ ，} \qquad \text{式(2-1)}$$

其中 $s_n[m]$ 與 $d_n[m]$ 分別表示第 n 個音框之乾淨語音訊號(clean speech)以及雜訊(noise)，則此音框之 log*E* 特徵值可用下式表示：

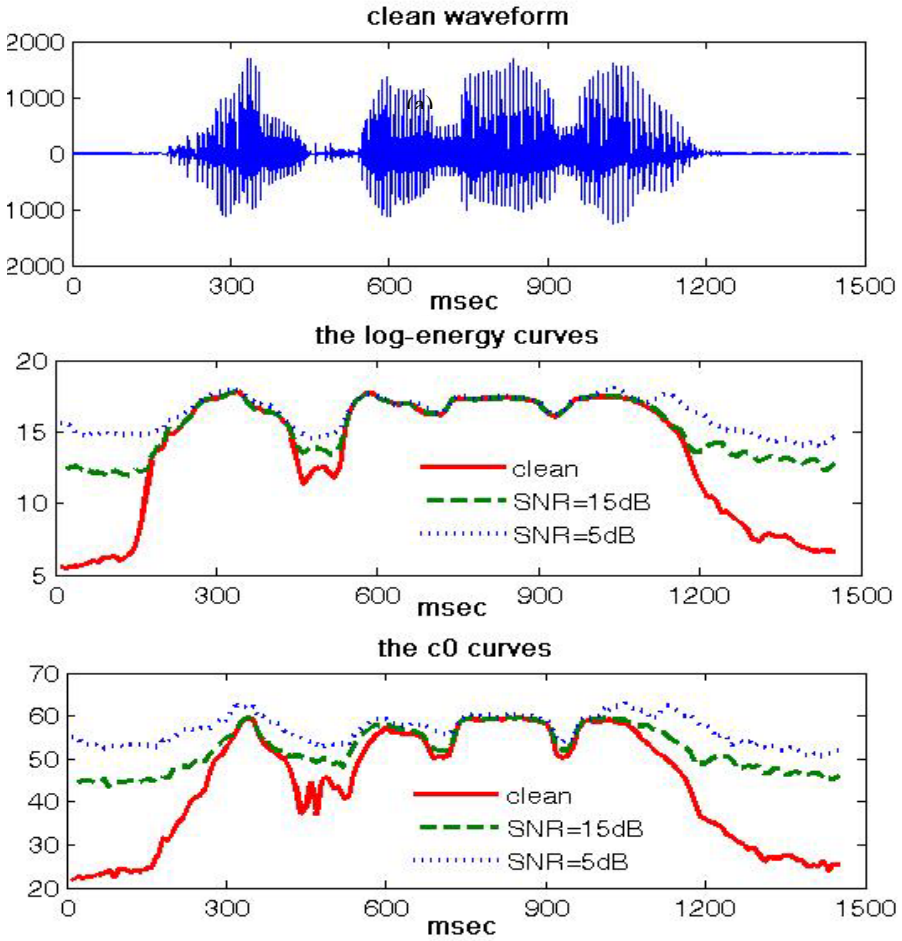
$$\begin{aligned} E^{(x)}[n] &= \log\left(\sum\nolimits_m x_n^2[m]\right) \approx \log\left(\sum\nolimits_m s_n^2[m] + \sum\nolimits_m d_n^2[m]\right) \\ &= \log\left(\exp\left(E^{(s)}[n]\right) + \exp\left(E^{(d)}[n]\right)\right) \text{ ，} \end{aligned} \qquad \text{式(2-2)}$$

其中 $E^{(x)}[n]$ 、 $E^{(s)}[n]$ 與 $E^{(d)}[n]$ 分別爲 $x_n[m]$ 、 $s_n[m]$ 以及 $d_n[m]$ 所對應之 log*E* 特徵值。因此，受到雜訊干擾所導致雜訊語音與乾淨語音訊號兩者間 log*E* 特徵的差異 $\Delta E[n]$ 可用下式表示：

$$\Delta E[n] = E^{(x)}[n] - E^{(s)}[n] \approx \log\left(1 + \exp\left(E^{(d)}[n] - E^{(s)}[n]\right)\right) \text{ 。} \qquad \text{式(2-3)}$$

由式(2-3)可觀察出，若在相同的雜訊能量($E^{(d)}[n]$)下，此差異值 $\Delta E[n]$ 與乾淨語音訊號之 $E^{(s)}[n]$ 兩者呈現負相關的關係，當 $E^{(s)}[n]$ 愈大時，$\Delta E[n]$ 愈小，反之則愈大。根據上述的推導，可以看出一雜訊語音訊號中，含有語音成份的音框($E^{(s)}[n]$ 較大)相較於純雜訊音框($E^{(s)}[n]$ 較小)而言，其 log*E* 特徵被雜訊影響的情況較小(即失真量 $\Delta E[n]$ 較小)。

接下來，我們探討加成性雜訊對於語音訊號的 log*E* 特徵序列於調變頻譜(modulation spectrum)上的影響。首先，我們將式(2-2)以泰勒級數(Taylor series)展開，其展開的中心點設定爲 $\left(E^{(s)}[n], E^{(d)}[n]\right) = (0,0)$ ，展開階層爲 2 階，如式(2-4)所示：



(a)

(b)

(c)

圖二、在不同 SNR 下，一語音訊號之波形圖及能量相關特徵時間序列圖，其中(a)爲乾淨語音波形、(b)爲 log*E* 特徵曲線、(c)爲 *c*0 特徵曲線

$$E^{(x)}[n] \approx \log\left(\exp\left(E^{(s)}[n]\right)+\exp\left(E^{(d)}[n]\right)\right)$$

$$\approx \log 2+\frac{1}{2}\left(E^{(s)}[n]+E^{(d)}[n]\right)+\frac{1}{8}\left(\left(E^{(s)}[n]\right)^2+\left(E^{(d)}[n]\right)^2-E^{(s)}[n]E^{(d)}[n]\right). \qquad \text{式(2-4)}$$

因此，若將上式(2-4)取傅立葉轉換，則此雜訊語音的對數能量序列$\left\{E^{(x)}[n]\right\}$的調變頻譜可用下式表示：

$$X(j\omega) \approx (2\pi\log 2)\delta(\omega)+\frac{1}{2}\left(S(j\omega)+D(j\omega)\right)$$

$$+\frac{1}{16\pi}\left(S(j\omega)*S(j\omega)+D(j\omega)*D(j\omega)-S(j\omega)*D(j\omega)\right), \qquad \text{式(2-5)}$$

式中$X\left(j\omega\right)$、$S\left(j\omega\right)$以及$D\left(j\omega\right)$分別爲雜訊語音之log*E*序列$\left\{E^{(x)}[n]\right\}$、乾淨語音之log*E*序列$\left\{E^{(s)}[n]\right\}$與雜訊之log*E*序列$\left\{E^{(d)}[n]\right\}$的調變頻譜。假設$\left\{E^{(s)}[n]\right\}$與$\left\{E^{(d)}[n]\right\}$兩序列皆爲低通(low-pass)訊號，且$B_s$與$B_d$爲其相對應之頻寬(bandwidth)，則式(2-5)中$D\left(j\omega\right)*D\left(j\omega\right)$與$S\left(j\omega\right)*D\left(j\omega\right)$兩項的頻寬分別爲$2B_d$與$B_s+B_d$；這意味著雜訊語音之log*E*序列$\left\{E^{(x)}[n]\right\}$相較於雜訊的log*E*序列$\left\{E^{(d)}[n]\right\}$將擁有更大的頻寬。換言之，對log*E*序列而言，雜訊語音比雜訊擁有較多高頻的調變頻譜成份；這便可以解釋爲何在一雜訊

語音訊號中含有語音的區段，比起純雜訊的區段看起來振盪情形(fluctuating)更爲明顯。

接著我們探討加成性雜訊對於 *c*0 特徵的影響。假設雜訊語音中第 n 個音框的 *c*0 特徵值以 $c_0^{(x)}[n]$ 做表示，而 $c_0^{(s)}[n]$ 與 $c_0^{(d)}[n]$ 分別表示此音框之所含乾淨語音訊號及純雜訊的 *c*0 特徵值，則它們可被推導如下三式：

$$c_0^{(x)}[n]=\sum\nolimits_k \log\left(M^{(x)}[k,n]\right)\approx\sum\nolimits_k \log\left(M^{(s)}[k,n]+M^{(d)}[k,n]\right)\text{，} \qquad \text{式(2-6)}$$

$$c_0^{(s)}[n]=\sum\nolimits_k \log\left(M^{(s)}[k,n]\right)\text{，} \qquad \text{式(2-7)}$$

$$c_0^{(d)}[n]=\sum\nolimits_k \log\left(M^{(d)}[k,n]\right)\text{，} \qquad \text{式(2-8)}$$

其中，$M^{(x)}[k,n]$、$M^{(s)}[k,n]$與$M^{(d)}[k,n]$分別爲式(2-1)中雜訊語音訊號 $x_n[m]$、乾淨語音訊號 $s_n[m]$以及雜訊 $d_n[m]$於轉換成梅爾倒頻譜特徵時，第 k 個梅爾濾波器的輸出值。因此我們可推導出，由於加成性雜訊干擾所導致雜訊語音與乾淨語音訊號兩者之 *c*0 特徵值的差異 $\Delta c_0[n]$ 如下式所示：

$$\begin{aligned}\Delta c_0[n]&=c_0^{(x)}[n]-c_0^{(s)}[n]\approx\sum\nolimits_k \log\left(1+\frac{M^{(d)}[k,n]}{M^{(s)}[k,n]}\right)\\&=\sum\nolimits_k \log\left(1+\frac{1}{SNR[k,n]}\right)\text{，}\end{aligned} \qquad \text{式(2-9)}$$

式中 $SNR[k,n]$ 定義爲第 n 個音框中第 k 維梅爾頻帶的訊雜比，即

$$SNR[k,n]=\frac{M^{(s)}[k,n]}{M^{(d)}[k,n]}. \qquad \text{式(2-10)}$$

由式(2-9)可看出，若多數梅爾頻帶的訊雜比 $SNR[k,n]$ 都比較大時，差異值 $\Delta c_0[n]$ 也相對變小，因此這可約略解釋含語音之音框(*SNR* 較大)相對於純雜訊音框(*SNR* 較小)而言，*c*0 特徵值較不易受到影響的現象。

以下我們將探討加成性雜訊對於 *c*0 特徵序列之調變頻譜(modulation spectrum)上的影響。首先爲了推導起見，我們將式(2-6)、式(2-7)與式(2-8)改寫成下列三式：

$$c_0^{(x)}[n]=\sum\nolimits_k \tilde{M}^{(x)}[k,n]\approx\sum\nolimits_k \log\left(\exp\left(\tilde{M}^{(s)}[k,n]\right)+\exp\left(\tilde{M}^{(d)}[k,n]\right)\right), \qquad \text{式(2-11)}$$

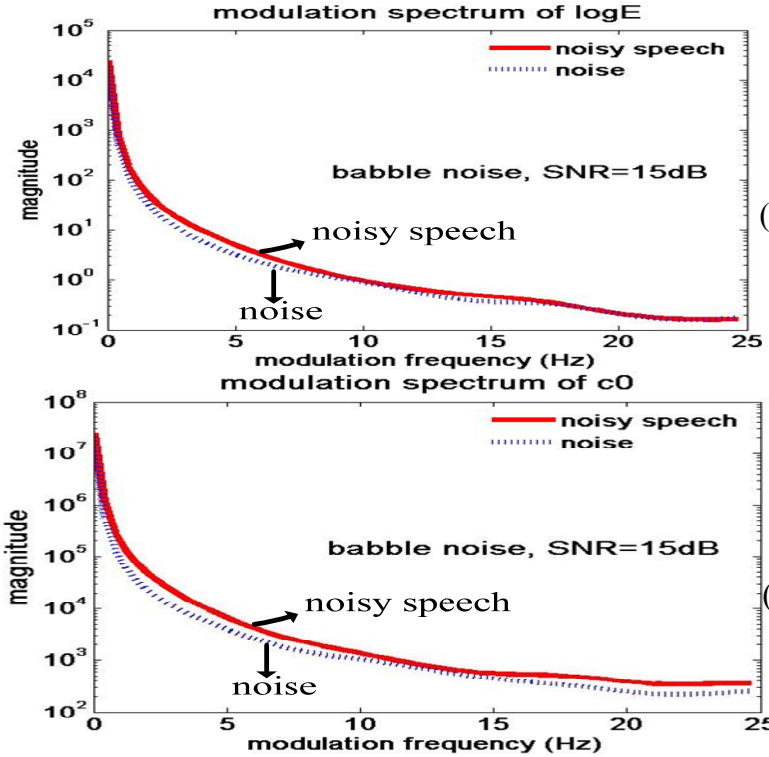
$$c_0^{(s)}[n]=\sum\nolimits_k \tilde{M}^{(s)}[k,n], \qquad \text{式(2-12)}$$

$$c_0^{(d)}[n]=\sum\nolimits_k \tilde{M}^{(d)}[k,n], \qquad \text{式(2-13)}$$

其中 $\tilde{M}^{(x)}[k,n]=\log\left(M^{(x)}[k,n]\right)$、$\tilde{M}^{(s)}[k,n]=\log\left(M^{(s)}[k,n]\right)$、$\tilde{M}^{(d)}[k,n]=\log\left(M^{(d)}[k,n]\right)$。類似將式(2-11)與式(2-2)作比較，可看出雜訊語音、乾淨語音與純雜訊三者的關係在 log*E* 與 *c*0 兩特徵中十分類似，因此藉由前面之式(2-4)與式(2-5)對於 log*E* 特徵序列之調變頻譜的推導，我們可以發現對每個梅爾濾波器輸出的對數值序列 $\left\{\tilde{M}^{(x)}[k,n]\right\}$ 而言，其頻寬仍是大於 $\left\{\tilde{M}^{(d)}[k,n]\right\}$，也就是說 $\left\{c_0^{(x)}[n]\right\}$ 比起 $\left\{c_0^{(d)}[n]\right\}$ 將擁有更大的頻寬，因此，類似 log*E* 特徵的結果，我們同樣歸納出雜訊語音之 *c*0 特徵序列比純雜訊之 *c*0 特徵序列擁有較多高頻的調變頻譜成份，亦即前者比後者有更明顯的上下振盪現象。

圖三(a)與圖三(b)分別爲一句語音訊號之 log*E* 特徵及 *c*0 特徵的功率頻譜密度(power spectral density, PSD)曲線圖，其中的語音訊號及雜訊爲 Aurora-2.0 資料庫中的 "FAC_5Z31ZZ4A"檔與人聲雜訊(babble noise)，訊雜比爲 15dB。由這兩圖我們可以很明

顯地看出，雜訊語音相對於純雜訊而言，其 log*E* 特徵序列與 *c*0 特徵序列都有較大的頻寬，此亦驗證了我們之前的推導。



圖三、能量相關特徵之功率頻譜密度圖，(a)爲 log*E* 特徵、(b)爲 *c*0 特徵

綜合上述的推導及圖例，我們驗證了一段雜訊語音中含有語音的音框其 log*E* 特徵與 *c*0 特徵相對於純雜訊音框而言，失真程度較小，且擁有較大的頻寬，亦即具有較明顯的上下振盪現象。基於上述觀察，我們將提出新的強健性語音特徵處理技術─靜音特徵正規化法(silence feature normalization, SFN)，其具有兩種模式，分述於之後的兩節中。

（二）靜音特徵正規化法 I (silence feature normalization I, SFN-I)

在本節中，我們介紹第一種模式的靜音特徵正規化法，稱之爲「靜音特徵正規化法 I」 (silence feature normalization I, SFN-I)；此方法是針對原靜音音框對數能量正規化法 (SLEN) [5]加以改良，目的是希望對 log*E* 與 *c*0 之能量相關特徵做處理，使一段訊號中非語音(non-speech)部份的特徵值做正規化，而含有語音之區域的特徵值則保持不變，以達到重建出乾淨語音訊號之能量相關特徵的效果。

首先，我們假設 $\{x[n]\}$ 爲一段雜訊語音訊號之 log*E* 特徵或 *c*0 特徵之序列；根據我們於上一小節所得到的結論，雜訊語音中含有語音的區段相較於純雜訊區段，其 log*E* 與 *c*0 特徵序列將擁有更高的調變頻譜成份；因此我們設計一高通無限脈衝響應濾波器 (high-pass infinite impulse response filter)來處理此段序列，其轉換函數如下：

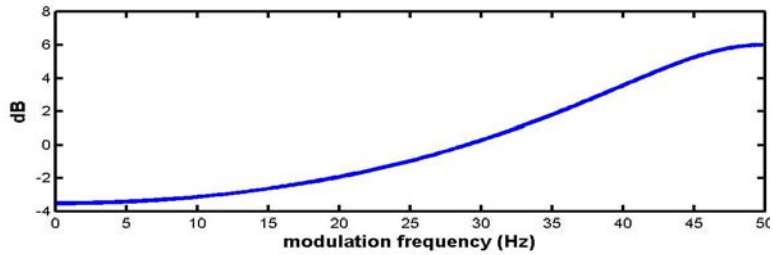
$$H(z)=\frac{1}{1+\alpha z^{-1}} \qquad 0<\alpha<1\text{。} \qquad \text{式(2-14)}$$

而此濾波器之輸入輸出關係式如下所示：

$$y[n]=-\alpha y[n-1]+x[n]\text{，} \qquad \text{式(2-15)}$$

式中 $y[n]$ 爲濾波器的輸出，且我們將其初始值設定爲 $y[0]=0$ 。式(2-14)之濾波器其強度響應(magnitude response)如圖四所示，由圖中可以發現，此濾波器能夠有效地降低特徵序列中接近直流(near-DC)的成份，並將較高頻率的部份加以強調，此較高頻率的成份

可突顯出語音與純雜訊的差異。因此經過此濾波器轉換後所得到的$\{y[n]\}$將比原始參數$\{x[n]\}$擁有更好的效能來判斷語音與非語音區段。



圖四、式(2-14)之高通濾波器的強度響應($\alpha = 0.5$)

根據式(2-15)所得之$y[n]$，我們可作一段訊號中語音與非語音音框的判別，並進而將其非語音的音框做正規化處理，此即爲靜音特徵正規化法 I (silence feature normalization I, SFN-I)，其式如下：

$$\text{SFN-I:}\quad \tilde{x}[n] = \begin{cases} x[n] & \text{if } y[n] > \theta \\ \log(\varepsilon) + \delta & \text{if } y[n] \le \theta \end{cases}, \qquad \text{式(2-16)}$$

其中θ、ε與δ分別爲門檻值、一極小的正數以及一平均值爲 0 且變異數很小的隨機變數，$\tilde{x}[n]$爲經過 SFN-I 處理後所得到的新特徵參數。其門檻值θ計算式如下：

$$\theta = \frac{1}{N}\sum_{n=1}^{N} y[n], \qquad \text{式(2-17)}$$

式中N爲此段語音的音框總數。因此，門檻值即爲整段語音所有$y[n]$的平均值，其計算十分簡便，且無需額外特別設計之處。

從式(2-16)看出，若$y[n]$大於門檻值θ，則將其所對應之音框判斷爲語音，且原特徵參數保持不變；反之則將其歸類爲非語音音框，並將原特徵參數正規化成一極小的隨機變數；相較於之前靜音音框對數能量正規化法(SLEN)[5]而言，靜音特徵正規化法 I 可避免將非語音部份的特徵正規化爲一定值，而可能導致之後所訓練的聲學模型中的變異數(variance)變爲 0 的錯誤現象產生。我們可以透過圖五來觀察 SFN-I 法的作用。圖五中，(a)與(b)分別爲原始的 log$E$ 特徵序列以及 $c0$ 特徵序列曲線；(c)與(d)分別爲經過靜音特徵正規化法 I 處理後所得到之 log$E$ 特徵序列以及 $c0$ 特徵序列曲線，其中紅色實線是對應至乾淨語音(Aurora-2.0 資料庫中的"FAK_3Z82A"檔)、綠色虛線與藍色點線則分別爲對應至訊雜比 15dB 與 5dB 的雜訊語音。由這些圖明顯地看出，SFN-I 法處理過後之能量相關特徵值可以較趨近於原始乾淨語音訊號之特徵值，達到降低失真的目的。

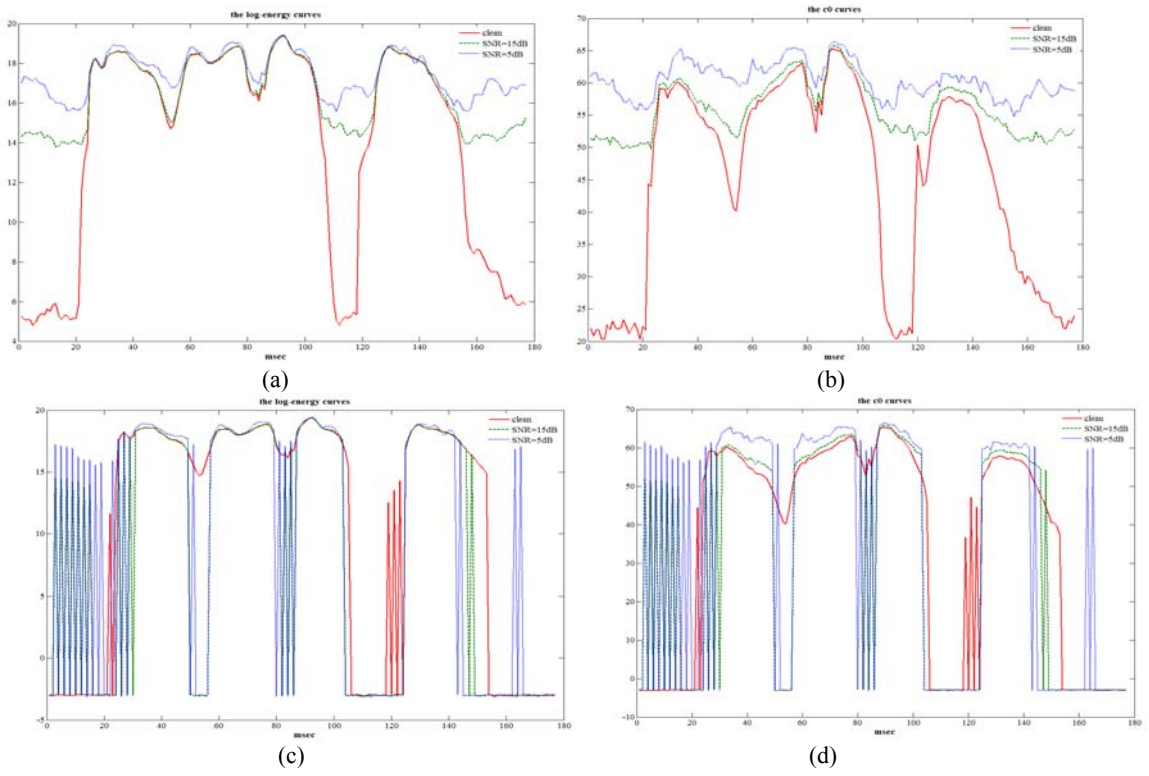
（三）靜音特徵正規化法 II (silence feature normalization II, SFN-II)

在本節中，我們將介紹第二種模式的靜音特徵正規化法，稱之爲「靜音特徵正規化法 II」(silence feature normalization II, SFN-II)，SFN-II 法與前一節之 SFN-I 法最大的差異在於，SFN-II 是將原能量相關特徵$\{x[n]\}$乘上一權重值(weight)，而得到新特徵值$\{\tilde{x}[n]\}$。SFN-II 的演算法如下式所示：

$$\text{SFN-II:}\quad \tilde{x}[n] = w[n]x[n], \qquad \text{式(2-18)}$$

其中，

$$w[n] = \begin{cases} 1/\left(1+\exp\left(-(y[n]-\theta)/\beta\sigma_1\right)\right) & \text{if } y[n] > \theta \\ 1/\left(1+\exp\left(-(y[n]-\theta)/\beta\sigma_2\right)\right) & \text{if } y[n] \le \theta \end{cases}, \qquad \text{式(2-19)}$$



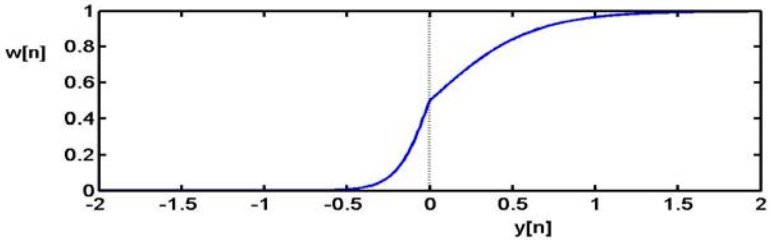
圖五、靜音特徵正規化法 I 處理前((a)與(b))與處理後((c)與(d))能量相關特徵序列曲線圖，其中(a)與(c)爲 log*E* 特徵序列曲線，(b)與(d)爲 *c*0 特徵序列曲線

其中 $y[n]$ 如前一節之式(2-15)所示，爲 $\{x[n]\}$ 通過一高通濾波器之輸出值，θ 爲門檻值、σ_1 與 σ_2 分別爲 $\{y[n]|y[n]>\theta\}$ (大於門檻值 θ 之所有的 $y[n]$)以及 $\{y[n]|y[n]\le\theta\}$ (小於或等於門檻值 θ 之所有的 $y[n]$)所對應之標準差、β 爲一常數。SFN-II 之門檻值 θ 跟 SFN-I 相同，計算式如下所示：

$$\theta=\left(1/N\right)\sum_{n=1}^{N}y[n], \qquad \text{式(2-20)}$$

式中 N 爲此段語音中音框總數。

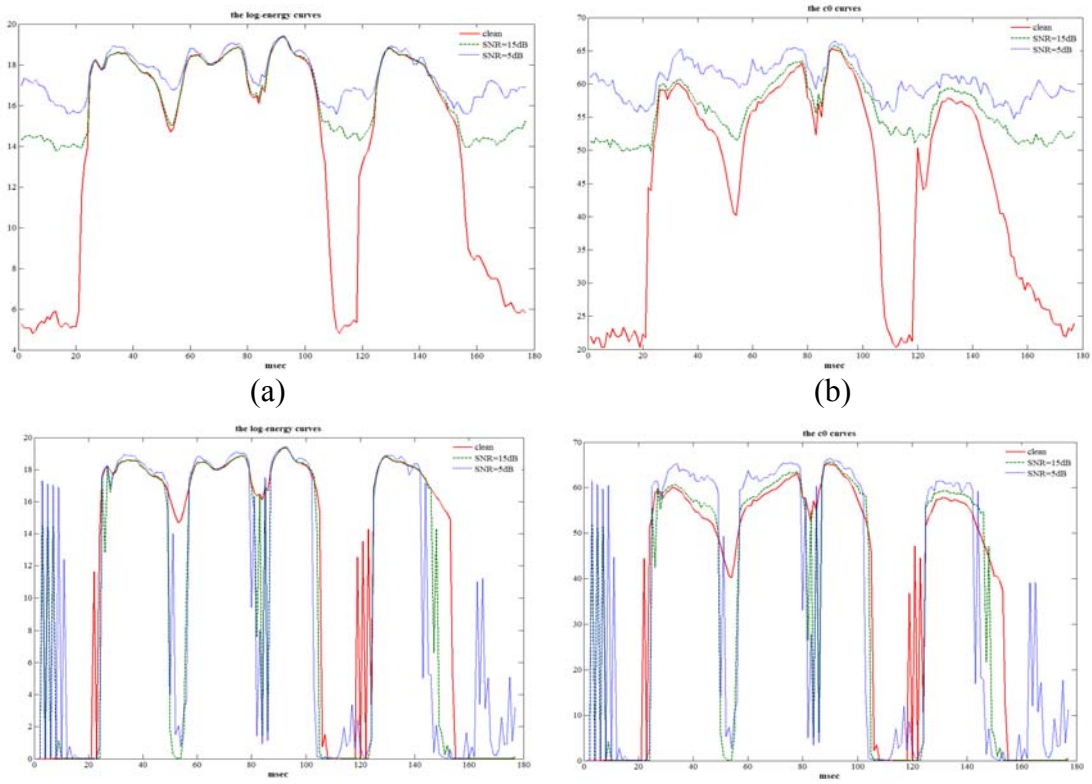
式(2-19)的權重值 $w[n]$ 如圖六所示，其中假設 $\theta=0$、$\sigma_1=1$、$\sigma_2=3$ 以及 $\beta=0.1$。由圖六可以發現，權重值函數 $w[n]$ 爲一個左右不對稱之遞增的 S 形曲線(sigmoid curve)，其值介於 0 和 1 之間。此權重值所代表的意義與 SFN-I 法相似，我們希望新得到的能量相關特徵 $\tilde{x}[n]$ 能在原始特徵值很大時，儘量維持不變；而原始值較小時，則使其變得更小。SFN-II 法和 SFN-I 法不同之處在於，SFN-II 法具有"軟式"的語音端點偵測決策(soft-decision VAD)，而 SFN-I 法則爲"硬式"的語音端點偵測決策(hard-decision VAD)；因此 SFN-II 法相較於 SFN-I 法而言，其 VAD 判定錯誤的影響可能相對來得比較小，效能也會比較好，這推想將會在之後的章節驗證。



圖六、權重值函數 $w[n]$ 曲線示意圖

圖七爲 SFN-II 法處理前與處理後能量相關特徵之曲線圖。與之前的圖三類似，(a)與(b)

分別爲原始的 log$E$ 特徵序列以及 c0 特徵序列曲線；(c)與(d)分別爲經過靜音特徵正規化法 II 處理後所得到之 log$E$ 序列以及 c0 序列曲線，其中紅色實線是對應至乾淨語音(Aurora-2.0 資料庫中的"FAK_3Z82A"檔)、綠色虛線與藍色點線則分別爲對應至訊雜比 15dB 與 5dB 的雜訊語音。很明顯地，經由 SFN-II 處理過後之雜訊語音的能量相關特徵，皆類似 SFN-I 法的效果，可以更趨近於原始乾淨語音之特徵，有效降低雜訊造成的失真。



圖七、靜音特徵正規化法 II 處理前((a)與(b))與處理後((c)與(d))能量相關特徵序列曲線圖，其中(a)與(c)爲 log$E$ 特徵序列曲線，(b)與(d)爲 c0 特徵序列曲線

三、能量相關特徵處理技術之實驗結果與討論

（一）、語音資料庫簡介

本論文中的語音辨識實驗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 Aurora-2.0 語料庫[7]。它是一套藉由以人工的方式錄製的連續英文數字字串，語者爲美國成年男女，加上八種加成性雜訊，分別爲地下鐵、人聲、汽車、展覽館、餐廳、街道、機場、火車站等，以及不同程度的訊雜比，分別爲 20dB、15dB、10dB、5dB、0dB 以及-5dB，附加上乾淨(clean)語料。

（二）、特徵參數的設定與辨識系統的訓練

本論文根據 Aurora-2.0 實驗語料庫標準設定[7]，語音特徵參數主要是使用梅爾倒頻譜係數(mel-frequency cepstral coefficients, MFCC)及能量相關特徵，附加上其一階差量與二階差量。爲了分析能量相關特徵的影響，於本論文中採用兩組不同的特徵參數；第一組是 12 維梅爾倒頻譜特徵值(c1～c12)加上 1 維的對數能量(log$E$)，另一組則是使用 12 維梅爾倒頻譜特徵值(c1～c12)加上第零維倒頻譜特徵係數(c0)；而每組皆會再加上一階與二階差量，故兩組皆用了 39 維的特徵參數。詳細的特徵參數設定，如表一所示。

我們利用 HTK 程式[8]來訓練聲學模型，產生了 11(oh, zero, one~nine)個數字模型以

及靜音模型，每個數字模型包含 16 個狀態(states)，而每個狀態是由 20 個高斯密度函數混合(Gaussian mixtures)所組成。

表一、本論文中所使用之語音特徵參數設定

取樣頻率	8kHz	
音框長度(Frame Size)	25ms, 200 點	
音框平移(frame Shift)	10ms, 80 點	
預強調濾波器	$1-0.97z^{-1}$	
視窗形式	漢明窗(Hamming window)	
傅立葉轉換點數	256 點	
濾波器組(filters)	梅爾刻度三角濾波器組， 共 23 個三角濾波器	
特徵向量 (feature vector)	第一組： $\{c_i \mid 1 \le i \le 12\}$, $\{\Delta c_i \mid 1 \le i \le 12\}$, $\{\Delta^2 c_i \mid 1 \le i \le 12\}$, $\log E, \Delta \log E, \Delta^2 \log E$ 共計 39 維	第二組： $\{c_i \mid 1 \le i \le 12\}$, $\{\Delta c_i \mid 1 \le i \le 12\}$, $\{\Delta^2 c_i \mid 1 \le i \le 12\}$, $c0, \Delta c0, \Delta^2 c0$ 共計 39 維

（三）語音辨識實驗結果

在這一節中，我們將執行各種針對能量相關特徵之強健性技術的語音辨識，並比較其效能。除了我們所新提出的靜音特徵正規化法（SFN-I與SFN-II）外，我們同時實驗了平均與變異數正規化法(mean and variance normalization, MVN)[9]、平均與變異數正規化附加ARMA濾波器法(MVN plus ARMA filtering, MVA)[10]、統計圖等化法(histogram equalization, HEQ)[11]、對數能量動態範圍正規化法(log-energy dynamic range normalization, LEDRN)[3]、對數能量尺度重刻法 (log-energy rescaling normalization, LERN)[4]與靜音對數能量正規化法(silence log-energy normalization, SLEN)[5]，值得注意的是，原始之MVN、MVA與HEQ三方法雖是設計於所有種類的特徵上，我們爲了評估其在能量相關特徵的效能，在這裡只將它們運用於log*E*與*c*0特徵的正規化上，另外，LEDRN法有分線性與非線性兩種，在這裡我們分別以LEDRN-I與LEDRN-II表示，而LERN亦有兩種版本，我們分別以LERN-I與LERN-II表示。

1、針對對數能量特徵(log*E*)之強健式語音技術綜合分析

此小節之實驗所用到語音特徵爲前述之第一組的特徵參數，即 12 維梅爾倒頻譜特徵值(*c*1～*c*12)加上 1 維的對數能量(log*E*)，附加其一階與二階差量，共 39 維。而這裡所用到的十種特徵強健性方法，皆是單純處理 log*E* 特徵，不考慮其它 12 維的梅爾倒頻譜係數，表二列出了基礎實驗及這十種方法所得之平均辨識率（20dB、15dB、10dB、5dB 與 0dB 五種訊雜比下的辨識率平均），其中 AR 與 RR 分別爲相較基礎結果之絕對錯誤降低率(absolute error rate reduction)和相對錯誤降低率(relative error rate reduction)。從表二的數據，我們可觀察到下列幾點現象：

①原始作用於所有種類特徵之 MVN、MVA 與 HEQ 法單純作用於 log*E* 特徵時，其提供的改進效果也十分明顯，相較於基礎實驗結果，分別具有 10.18%、11.70%與 14.97%的辨識率提升。相對於 MVN 而言，由於 MVA 多使用了一個 ARMA 低通濾波器以強調語音的成分，而 HEQ 額外對語音特徵的高階動差(higher-order moments)作正規化，所以兩者效果皆比 MVN 還來得好。

②以往文獻所提出之針對 log*E* 特徵作補償的各種方法：LEDRN-I、LEDRN-II、LERN-I、LERN-II 與 SLEN，都能帶來十分顯著的辨識率提升，其中線性 LEDRN(LEDRN-I)明顯優於非線性 LEDRN(LEDRN-II)，其平均辨識率相差了大約 4%，兩種版本的 LERN(LERN-I 與 LERN-II)，效果則十分接近，且表現優於 LEDRN。而本實驗室過去所提出的 SLEN 法，相對於基礎實驗的平均辨識率而言，有 15.19%的提升，明顯優於之前所提之 LEDRN 與 LERN 等方法。

③ 本論文所提出的兩種靜音特徵正規化法，SFN-I 與 SFN-II，相對於基礎實驗結果而言，平均辨識率分別提升了 15.38%與 16.11%，相對錯誤降低率都在 50%以上，相較於之前所提的各種方法，SFN-I 與 SFN-II 都有更優異的表現，此驗證了我們所提的兩個新方法，都能有效地提昇 log*E* 特徵在加成性雜訊環境下的強健性，且優於目前許多著名的 log*E* 特徵正規化技術。此外，我們也發現，SFN-II 所得之辨識率比 SFN-I 更好，此可能原因如之前所述，由於 SFN-II 在語音偵測(voice activity detection)的決策機制與 SFN-I 並不相同，語音偵測之錯誤在 SFN-II 中相對影響較小，而使其相對表現較佳。

表二、針對 log*E* 特徵之強健式語音技術之辨識率的綜合比較表(%)

	Method	Set A	Set B	average	AR	RR
(1)	Baseline	71.98	67.79	69.89	—	—
(2)	MVN	79.04	81.08	80.06	10.18	33.79
(3)	MVA	80.53	82.64	81.59	11.70	38.85
(4)	HEQ	83.91	85.79	84.85	14.97	49.69
(5)	LEDRN-I	82.01	79.70	80.86	10.97	36.43
(6)	LEDRN-II	77.21	75.53	76.37	6.49	21.53
(7)	LERN-I	83.64	83.35	83.50	13.61	45.19
(8)	LERN-II	82.71	81.94	82.33	12.44	41.31
(9)	SLEN	84.87	85.27	85.07	15.19	50.42
(10)	**SFN-I**	**85.02**	**85.50**	**85.26**	**15.38**	**51.05**
(11)	**SFN-II**	**85.67**	**86.32**	**86.00**	**16.11**	**53.49**

2、針對第零維倒頻譜特徵係數(*c*0)之強健式語音技術綜合分析

此小節之實驗所用到語音特徵爲前述之第二組的特徵參數，即 12 維梅爾倒頻譜特徵值(*c*1～*c*12)加上第零維倒頻譜特徵係數(*c*0)，附加其一階與二階差量，共 39 維。類似前一小節，我們將原始針對 log*E* 特徵的十種特徵強健性方法，作用於 *c*0 特徵上，其它 12 維的梅爾倒頻譜係數則維持不變。雖然目前處理的是 *c*0 特徵，但爲了簡明起見，這裡我們不將原本各種技術的名稱作修改，例如 LEDRN 法，我們並不特別將其改名爲 *c*0-DRN 法，而仍沿襲其名，其他方法名稱依此類推。

表三列出了基礎實驗及這十種方法所得之平均辨識率（20dB、15dB、10dB、5dB 與 0dB 五種訊雜比下的辨識率平均），而其中的 AR 與 RR 分別爲相較於基礎實驗結果之絕對錯誤降低率和相對錯誤降低率。從表三的數據，我們可觀察到下列幾點現象：

①類似之前的表二之結果，各種方法作用於 *c*0 特徵時，都能帶來提昇辨識率的效果，其中，LEDRN-I 與 LEDRN-II 的表現比其他方法稍差，尤其是 LEDRN-II，只有 3.57% 之絕對錯誤降低率(AR)，其可能原因在於，LEDRN 原本是針對 log*E* 特徵所設計，若我們直接將其套用於 *c*0 特徵處理上，其所使用的參數並非是最佳化而得，導致效果不彰。

②三種原本作用於所有種類特徵之方法：MVN、MVA 與 HEQ 法，單純作用於 *c*0 特徵

時，仍然以 HEQ 表現最好，MVA 法次之，MVN 法較差，但彼此表現的差距並未如之前在表二來得明顯。此外，LERN-I、LERN-II 與 SLEN 都有十分顯著的改進效果，唯與表二的數據不同之處，在於三種方法的效能十分接近，而 LERN-I 略優於 SLEN。

③本論文所提出的兩種靜音特徵正規化法，SFN-I 與 SFN-II，相對於基礎實驗結果而言，平均辨識率分別提升了 13.79%與 14.13%，相對錯誤降低率約爲 46%，類似表二的結果，SFN-II 仍然優於 SFN-I，且這兩種方法之表現仍優於其他所有的方法。此結果驗證了我們所提的兩個新方法，能有效地提昇 *c*0 特徵在加成性雜訊環境下的強健性。

表三、針對 *c*0 特徵之強健式語音技術之辨識率的綜合比較表(%)

	Method	Set A	Set B	Average	AR	RR
(1)	Baseline	71.95	68.22	70.09	—	—
(2)	MVN	80.80	82.95	81.88	11.79	39.41
(3)	MVA	81.76	84.04	82.90	12.82	42.84
(4)	HEQ	82.89	84.59	83.74	13.66	45.65
(5)	LEDRN-I	79.04	77.36	78.20	8.11	27.13
(6)	LEDRN-II	74.08	73.22	73.65	3.57	11.92
(7)	LERN-I	83.81	83.65	83.73	13.65	45.61
(8)	LERN-II	83.03	82.53	82.78	12.70	42.44
(9)	SLEN	82.94	84.28	83.61	13.53	45.21
(10)	**SFN-I**	**83.04**	**84.70**	**83.87**	**13.79**	**46.08**
(11)	**SFN-II**	**83.29**	**85.14**	**84.22**	**14.13**	**47.23**

雖然 SFN 法有效地降低雜訊對 *c*0 造成的失真，進而提昇辨識率，但當我們比較表二與表三時，發現無論是 SFN-I 或 SFN-II，作用於 log*E* 特徵可得到的辨識率會高於作用在 *c*0 特徵所得之辨識率；由此，我們推斷由 log*E* 特徵所得之 SFN-I 法與 SFN-II 法其中的語音端點偵測(VAD)結果，可能會比由 *c*0 所得結果來的好。根據此推想，我們將原來針對 *c*0 特徵的兩種 SFN 法稍作修改。於 SFN-I 中，我們先利用 log*E* 對音框做語音/非語音的分類，再將此判別結果套用於 *c*0 上，對非語音音框的 *c*0 做如式(2-16)之正規化處理；而 SFN-II 也是利用相同的方式，先利用 log*E* 對音框做語音/非語音的分類，再將其結果轉換至 *c*0 上，並對語音與非語音音框的 *c*0 特徵序列求取其式(2-19)所用的標準差 σ_1 與 σ_2，然後作式(2-18)之正規化處理。我們將以上的修正作法分別稱作針對 *c*0 特徵之修正式 SFN-I 法(modified SFN-I)與修正式 SFN-II 法(modified SFN-II)。

針對 *c*0 特徵之修正式 SFN-I 法與修正式 SFN-II 法，其所得之平均辨識率如表四所示，如我們所預期的，修正式 SFN 法相對於原始 SFN 法，能有更進一步的改進效果，對 SFN-I 而言，前者相較於後者額外提昇了 1.29%的平均辨識率，而對 SFN-II 而言，前者相較於後者額外提昇了的 1.33%平均辨識率。此結果部分驗證了我們的推想，即利用 log*E* 特徵來執行語音端點偵測(VAD)，其效果會比 *c*0 特徵來的好。

表四、針對 *c*0 特徵之原始 SFN 法與修正式 SFN 法之辨識率比較表(%)

Method	Set A	Set B	Average	AR	RR
Baseline	71.95	68.22	70.09	—	—
SFN-I	83.04	84.70	83.87	13.79	46.08
modified SFN-I	**84.54**	**85.79**	**85.17**	**15.08**	**50.41**
SFN-II	83.29	85.14	84.22	14.13	47.23
modified SFN-II	**85.03**	**86.06**	**85.55**	**15.46**	**51.68**

四、靜音特徵正規化法與其它特徵強健法結合之實驗結果與討論

前一章之一系列的實驗，主要是探討各種能量相關特徵處理技術效能，進而突顯出我們所新提出之靜音特徵正規化(SFN)法的優異表現，這些實驗中，只有 logE 與 $c0$ 兩種能量相關特徵被處理，剩餘的梅爾倒頻譜特徵係數($c1$~$c12$)則維持不變。在這一章中，我們嘗試將作用於 log$E$ 與 $c0$ 特徵的 SFN 法與作用於 $c1$~$c12$ 之梅爾倒頻譜特徵係數的強健性技術加以結合，藉以觀察兩者之間是否有加成性，能進一步改進語音辨識率。

在這裡，我們選擇之前所提之 MVN[9]、MVA[10]以及 HEQ[11]三種強健性技術，分別作用於 $c1$~$c12$ 之梅爾倒頻譜特徵係數上，而將我們所提之 SFN-I 或 SFN-II 法作用於能量相關特徵(log$E$ 或 $c0$)上，我們將其上述所有的實驗結果分別彙整成表五與表六。

針對第一組特徵(log$E$, $c1$~$c12$)處理之表五的數據中，列(2)~(4)是利用單一強健技術(MVN, MVA 或 HEQ)處理全部特徵參數之結果，而列(5)~(10)則分別爲靜音特徵正規化法(SFN)結合其它方法之結果。當我們將列(2)、列(5)與列(8)的結果相比較、列(3)、列(6)與列(9)的結果相比較，及列(4)、列(7)與列(10)的結果相比較，都可以看出將 SFN-I 或 SFN-II 使用於 log$E$ 特徵，並用其他方法使用在 $c1$～$c12$ 特徵上，所得到的辨識率比單獨使用一種方法處理全部特徵的辨識結果高出許多，例如列(9)之『SFN-II (log$E$) + MVA ($c1$~$c12$)』法，其平均辨識率高達 89.97%，超越了列(4)之『HEQ (log$E$, $c1$~$c12$)』法所得之 87.44%的平均辨識率。同時，我們也看出 SFN-II 的效能普遍優於 SFN-I，此結果跟前一章的結論是一致的。而當我們將表五與表二的數據相比較時，也可以看出，使用 SFN 處理 log$E$ 特徵結合使用 MVN、MVA 或 HEQ 法額外處理 $c1$～$c12$ 特徵，可以比單獨使用 SFN 處理 log$E$ 特徵得到更佳的辨識效果，此結果驗證了 SFN 法與 MVN、MVA 或 HEQ 法的確具有加成性。

表五、SFN 法作用在 log$E$ 特徵結合其它語音強健技術作用於 $c1$～$c12$ 特徵參數之平均辨識率的綜合比較表(%)

	Method	Set A	Set B	average	AR	RR
(1)	Baseline	71.98	67.79	69.89	—	—
(2)	MVN (log$E$, $c1$~$c12$)	83.55	83.75	83.65	13.77	45.71
(3)	MVA (log$E$, $c1$~$c12$)	86.69	86.89	86.79	16.91	56.13
(4)	HEQ (log$E$, $c1$~$c12$)	87.15	87.72	87.44	17.55	58.28
(5)	SFN-I (log$E$) + MVN ($c1$~$c12$)	87.33	87.81	87.57	17.69	58.72
(6)	SFN-I (log$E$) + MVA ($c1$~$c12$)	88.40	88.84	88.62	18.74	62.21
(7)	SFN-I (log$E$) + HEQ ($c1$~$c12$)	87.93	88.04	87.99	18.10	60.10
(8)	SFN-II (log$E$) + MVN ($c1$~$c12$)	88.45	88.88	88.67	18.78	62.36
(9)	SFN-II (log$E$) + MVA ($c1$~$c12$)	89.82	90.12	89.97	20.09	66.69
(10)	SFN-II (log$E$) + HEQ ($c1$~$c12$)	89.29	89.33	89.31	19.43	64.50

針對第二組特徵($c0$, $c1$~$c12$)處理之表六的數據中，列(2)~(4)是利用單一強健技術(MVN, MVA 或 HEQ)處理全部特徵參數之結果，而列(5)~(16)則分別爲靜音特徵正規化法(SFN)結合其它方法之結果。類似表五中列(1)~(10)所呈現的結果，從表六中之列(1)~(10)與表三的數據相較，使用 SFN 處理 $c0$ 特徵結合使用 MVN、MVA 或 HEQ 法額外處理 $c1$～$c12$ 特徵，可以比單獨使用 SFN 處理 $c0$ 特徵得到更佳的效能，然而我們發現，將 SFN-I 或 SFN-II 使用於 $c0$ 特徵，並用其他方法使用在 $c1$～$c12$ 特徵時，所得到的辨識率並非總是優於單獨使用一種方法處理全部特徵的辨識結果（這些較差的數據在表中以*號加以註記），例如列(6)之『SFN-I ($c0$) + MVA ($c1$~$c12$)』法，其平均辨識率爲

87.77%，相較於列(3)之『MVA (*c*0, *c*1~*c*12)』法所得之 88.46%來得差。此現象的可能原因，在前一章已經提到，即利用 *c*0 特徵執行 SFN 法中的語音端點偵測(VAD)會比較不精確，進而降低 SFN 的效能。因此，類似前一章，在這裡我們使用針對 *c*0 特徵之修正式的 SFN 法，來與 MVN、MVA 或 HEQ 法作結合，這些結果列於表六的列(11)~(16)中。

表六、SFN 法作用在 *c*0 特徵結合其它語音強健技術作用於 *c*1～*c*12 特徵參數之平均辨識率綜合比較表(%)

	Method	Set A	Set B	Average	AR	RR
(1)	Baseline	71.95	68.22	70.09	—	—
(2)	MVN (*c*0, *c*1~*c*12)	85.03	85.54	85.29	15.20	50.81
(3)	MVA (*c*0, *c*1~*c*12)	88.11	88.81	88.46	18.38	61.42
(4)	HEQ (*c*0, *c*1~*c*12)	86.99	88.13	87.56	17.48	58.42
(5)	SFN-I (*c*0) + MVN (*c*1~*c*12)	85.62	86.62	86.12	16.04	53.60
(6)	SFN-I (*c*0) + MVA (*c*1~*c*12)	87.38*	88.16*	87.77*	17.69	59.12
(7)	SFN-I (*c*0) + HEQ (*c*1~*c*12)	85.95*	86.53*	86.24*	16.16	54.00
(8)	SFN-II (*c*0) + MVN (*c*1~*c*12)	86.92	87.69	87.31	17.22	57.56
(9)	SFN-II (*c*0) + MVA (*c*1~*c*12)	89.04	89.61	89.33	19.24	64.32
(10)	SFN-II (*c*0) + HEQ (*c*1~*c*12)	87.43	87.88*	87.66	17.57	58.73
(11)	modified SFN-I (*c*0) + MVN (*c*1~*c*12)	87.49	87.89	87.69	17.61	58.85
(12)	modified SFN-I (*c*0) + MVA (*c*1~*c*12)	89.30	89.54	89.42	19.34	64.63
(13)	modified SFN-I (*c*0) + HEQ (*c*1~*c*12)	88.10	88.39	88.25	18.16	60.71
(14)	modified SFN-II (*c*0) + MVN (*c*1~*c*12)	88.25	88.33	88.29	18.21	60.86
(15)	modified SFN-II (*c*0) + MVA (*c*1~*c*12)	89.87	89.98	89.93	19.84	66.32
(16)	modified SFN-II (*c*0) + HEQ (*c*1~*c*12)	89.25	89.46	89.36	19.27	64.42

將表六之列(11)~(16)的數據與列(1)~(10)相比較，我們可以明顯看出針對 *c*0 特徵之修正式 SFN 法(modified SFN-I 與 modified SFN-II)，比原始 SFN 法的效能高出許多，且與 MVN、MVA 或 HEQ 一併使用後，其結果必然優於 MVN、MVA 或 HEQ 處理所有特徵的結果，其中以列(15)之『modified SFN-II (*c*0) + MVA (*c*1~*c*12)』法所得到的平均辨識率最高，爲 89.93%，與之前表五中最佳辨識率 89.97%（列(9)的『SFN-II (log*E*) + MVA (*c*1~*c*12)』法）十分接近，此結果明顯驗證了修正式 SFN 法確實更進一步改進了 *c*0 特徵在加成性雜訊環境下的強健性。

由第三章與第四章之全部的實驗數據中，我們可以充分驗證所提出的兩種靜音特徵正規化法(SFN-I 與 SFN-II)對於能量相關特徵具有良好的強健化效果，而 SFN-II 所得到的辨識率總是比 SFN-I 高，其可能原因如第二章所陳述，因爲 SFN-II 法具有"軟式"決策之語音端點偵測(soft-decision voice activity detection)的機制，相較於 SFN-I 法"硬式"決策之語音端點偵測(hard-decision voice activity detection)的機制，前者的語音/非語音判別錯誤所造成的影響相對較小。然而，總括而言，SFN-I 法 SFN-II 法的共同優點在於執行上十分簡易（即複雜度極低）且效果很優異，因此極具實用的價值。

五、結論

在本論文中，我們提出一個新的語音強健技術 —「靜音特徵正規化法」(silence feature normalization, SFN)，此方法執行上十分簡易且效果優異。它是針對能量相關特

徵(log*E* 與 *c*0)因加成性雜訊造成的失真現象作適當的補償。SFN 法利用了一個高通濾波器去處理原始能量相關特徵序列，並將通過此高通濾波器所的之輸出特徵序列拿來作語音/非語音的分類，並應用簡單且有效的方法來處理非語音部份的特徵，將雜訊對語音特徵的干擾降低，以期提升訓練與測試環境匹配度，進而提升雜訊環境下的語音辨識率。

由實驗數據中可發現，就處理能量相關特徵而言，SFN 法比基本實驗以及許多強健式語音技術得到更好的辨識率；由此可知針對能量相關特徵做適當的補償，在穩定以及非穩定雜訊環境下皆得到十分顯著的辨識率提升，顯示了能量相關特徵所含的語音鑑別資訊是影響辨識率的一個重要指標。此外，當我們將 SFN 法與其它強健式語音技術做結合，發現其辨識率比單獨使用一種強健式語音技術所得到的辨識率更高，其中又以 SFN-II 法結合 MVA 法得到的辨識率最高，可達到將近 90%的平均辨識率。

能量相關特徵雖然具高度語音鑑別力，但是雜訊對其干擾程度也相對很大，因此能量相關特徵處理的好壞，將會很直接地影響到系統的辨識效能，由此可知能量相關特徵的強健化處理在未來仍是值得探討的一大課題；我們希望未來可以將所發展的技術，擴展測試至其它較大字彙量的語音辨識系統上，探討這類技術在不同複雜度之語音辨識系統的效能。另外，未來我們仍可朝向消除加成性雜訊的方向繼續深入研究，也可以針對消除通道性雜訊的方法去作相關的探討，並嘗試將兩者結合，使得語音辨認系統能更有效地降低各類雜訊的干擾，而擁有令人滿意之辨識率。

參考文獻

[1] Bocchieri , E. L., and Wilpon, J. G., "Discriminative Analysis for Feature Reduction in Automatic Speech Recognition", 1992 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1992).

[2] Julien Epps and Eric H.C. Choi, "An Energy Search Approach to Variable Frame Rate Front-End Processing for Robust ASR", 2005 European Conference on Speech Communication and Technology (Interspeech 2005—Eurospeech).

[3] Weizhong Zhu and Douglas O'Shaughnessy, "Log-Energy Dynamic Range Normalization for Robust Speech Recognition", 2005 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005).

[4] Hung-Bin Chen, "On the Study of Energy-Based Speech Feature Normalization and Application to Voice Activity Detection", M.S. thesis, National Taiwan Normal University, Taiwan, 2007.

[5] C-F. Tai and J-W. Hung, "Silence Energy Normalization for Robust Speech Recognition in Additive Noise Environments", 2006 International Conference on Spoken Language Processing (Interspeech 2006—ICSLP).

[6] Tai-Hwei Hwang and Sen-Chia Chang, "Energy Contour Enhancement for Noisy Speech Recognition", 2004 International Symposium on Chinese Spoken Language Processing (ISCSLP 2004).

[7] H. G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," Proceedings of ISCA IIWR ASR2000, Paris, France, 2000

[8] http://htk.eng.cam.ac.uk/

[9] S. Tiberewala and H. Hermansky, "Multiband and Adaptation Approaches to Robust Speech Recognition", 1997 European Conference on Speech Communication and Technology (Eurospeech 1997)

[10] C-P. Chen and J-A. Bilmes, "MVA Processing of Speech Features", IEEE Trans. on Audio, Speech, and Language Processing, 2006

[11] A. Torre, J. Segura, C. Benitez, A. M. Peinado, and A. J. Rubio, "Non-Linear Transformations of the Feature Space for Robust Speech Recognition", 2002 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2002)