國 立 交 通 大 學

資訊工程學系

碩 士 論 文

語音強化技術在相加性雜訊環境下的語音辨識之研究

The Study of Speech Enhancement in Additive Noise Environment for Speech Recognition

研 究 生：沈 揚 智

指導教授：傅 心 家 教授

中華民國九十四年七月

語音強化技術在相加性雜訊環境下的語音辨識之研究

The Study of Speech Enhancement in Additive Noise Environment for Speech Recognition

研 究 生 ： 沈揚智　　Student: Yang-Chih Shen

指導教授 ： 傅心家　教授　　Advisor: Prof. Hsin-Chia Fu

A thesis Submitted to Institute of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Chiao Tung University

In Partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science and Information Engineering

July 2005

Hsinchu, Taiwan, Republic of China

中華民國九十四年七月

語音強化技術在相加性雜訊環境下的語音辨識之研究

研究生：沈揚智　　指導教授：傅心家 教授

國立交通大學資訊工程學系

摘 要

環境雜訊的干擾是導致目前語音辨識技術無法普遍應用在實際環境中的瓶頸。爲此，本論文針對了相加性雜訊環境下的語音辨識系統，提出了強化型 MMSE 語音強化法，以消除環境雜訊對語音的干擾。此方法是以最小平方誤差短時頻譜振幅估計法為基礎，並考慮語音訊號與雜訊訊號在某段時間中的變動程度，去調整濾波器的頻率響應，以達到強調語音訊號並壓抑雜訊訊號的目的。

我們根據 AURORA 提出的語音辨識架構進行實驗。實驗結果說明了：1. 透過根據時間變動程度的調整方式，強化型 MMSE 語音強化法確實能夠增加強化後語音特徵中差量參數的正確性；2.與其他的語音強化法進行比較，本方法也能夠在準確率上有所提升。我們並將此方法實作在一個分散式語音辨識系統上，經由多位使用者實際操作後，確實能有不錯的辨識效能。

The Study of Speech Enhancement in Additive Noise Environment for Speech Recognition

Student : Yang-Chih Shen Advisor : Prof. Hsin-China Fu

Institute of Computer Science and Information Engineering

National Chiao Tung University

Abstract

In practical environment, the speech recognition performance degrades drastically due to the background noise interference. For this reason, we propose the enhanced MMSE speech enhancement approach for the speech recognition in additive noise environment. This approach is based on Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator, and adjusts the filter frequency response according to the variation of speech and noise in the local time period, in order to boost the speech variance and suppress the noise variance.

The experiment follows the AURORA proposed architecture. The result shows this adjusting approach increases the correctness of delta-coefficient, and has better accuracy comparing to other speech enhancement method. Moreover, we apply the proposed method and implement a distributed speech recognition (DSR) system.

誌　謝

本論文能圓滿的完成，首先要謝謝我的指導老師，傅教授在這兩年來辛勤的指導。另外要特別感謝永煜博士，在論文完成的最後階段給予極大的協助，包括了演算公式推導、實驗設計及論文校稿等，陪著我一起忙碌了好長一段時間，還有士賢學長，教導我許多語音辨識的相關知識；此外清大的陸清達學長也不吝給我這個來自交大的研究生指教，幫我釐清研究的目標；還要感謝宜玲、宗儒一起相互加油打氣，實驗室的陳岳宏、曾政龍、賴柏伸學長在課業學習上的幫忙、以及學弟政邦、富評和建榮適時地加入，亦增添不少生活的樂趣。最後要謝謝我的家人，爺爺、奶奶、爸爸、媽媽還有哥哥，不斷地給我鼓勵，和那些曾經鼓勵過我的好朋友。

回首這兩年研究生涯中的點滴，實在是感觸良多。大學畢業後，我選擇進入交大資工，而沒有繼續留在台北科大，現在讓我重新選擇，我還是會選擇交大資工。因為在交大體驗到了另一個不同的學習環境，也認識了許多老師以及新朋友，這些點滴都是我人生中一輩子的寶藏。

目　　錄

圖　目　錄

表　目　錄

第一章 簡介

1.1 動機

隨著科技的進步，使用語音輸入代替手寫或打字已不再是夢想。然而在實際應用上，語音辨識所遇到的最大問題在於環境雜訊(background noise)的干擾，大幅提高了語音辨識的困難度。為了減少環境雜訊對語音的干擾，我們可以使用指向性麥克風或是頭戴式耳機麥克風來輸入語音，但是這卻不是最根本的解決辦法，因為這樣限制了語音辨識的應用環境，亦降低了使用語音輸入的便利性。導致辨識準確度降低的最根本原因就是在訓練語音模型時用的訓練語音與在實際環境中輸入的語音特性不一致。在訓練語音模型時所使用的是乾淨語音(clean speech)，但是應用在實際環境中，輸入的是含雜訊語音(noisy speech)，由於雜訊的干擾而使得語音特徵產生失真，因此語音辨識在面對含雜訊語音的辨識效能會大打折扣。

語音辨識系統在應用上可為單一麥克風或是多麥克風；由於多麥克風的系統需要兩個以上的麥克風去收集與噪音相關的額外資訊，並不常見於一般的語音辨識系統，所以本論文研究的目標是在單一麥克風的語音辨識系統下，減小環境雜訊的干擾，以增加語音的辨識率。

單一麥克風的語音辨識系統如下圖 1-1 所示：

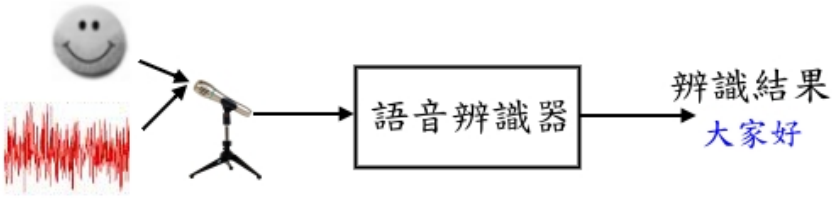

圖 1-1: 單一麥克風的語音辨識系統示意圖

由於系統只有一個麥克風，所以由麥克風接收進來的訊號是語音訊號與雜訊疊加後的結果，這類雜訊稱之為相加性雜訊(additive noise)。語音強化(speech enhancement)是一項能夠有效補償相加性雜訊造成語音失真的技術。此技術的處理方式是在抽取語音辨識所需的語音特徵參數前，先消除掉部分含雜訊語音中的雜訊成分，而能較接近原始語音的特性，這樣的做法能使辨識系統在有雜訊干擾的環境下仍能保有不錯的準確度。

語音強化的處理方式可分類為相減型(subtractive-type)語音強化法[1][2]和訊號子空間法(signal subspace) [3][4]。相減型語音強化法的做法是把含雜訊語音減去雜訊的估計值，還原出近似於乾淨語音的訊號。訊號子空間法的運作原理是將觀察到的含雜訊語音向量空間拆解成含雜訊訊號以及雜訊兩個子空間，在藉由移除雜訊的子空間以及從含雜訊訊號子空間中估測出較乾淨的語音訊號，而達到語音強化的目的。由於訊號子空間法需要進行向量空間的轉換，其運算複雜度比起相減型語音強化法高出許多，所以並不合適應用在一般的語音辨識系統，因此本論文中只針對相減型語音強化法進行討論。

大多數相減型語音強化方法依靠著短時間內(20~30ms)含雜訊語音中的穩定特性(stationary)去處理，而忽略了語音頻譜在長時間來看具有不穩定的特性(non-stationary)。本論文以最小平方誤差短時頻譜振幅估計法(Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator，簡稱 MMSE 估計法)為基礎，並考慮語音訊號與雜訊訊號在某段時間中的相對變動關係進行強化，而

提出了強化型 MMSE(enhanced MMSE)語音強化法。經過相加性雜訊環境下語音辨識的實驗，強化型 MMSE 語音強化法相較於其它的語音強化方法，在低訊雜比的環境下，確實能有效地提升語音辨識率。

1.2 章節組織

接下來本論文的組織如下：在第二章中，我們會先介紹一些與語音強化相關的背景知識和研究。本論文提出的強化型 MMSE 語音強化法將在第三章中描述，並在第四章中進行相加性雜訊環境下的實驗以及討論，然後以此方法架構出的分散式語音辨識系統將呈現在第五章。最後第六章為本論文的結論並探討未來的發展。

第二章 相關研究

本章將先介紹在相加性雜訊干擾下，語音強化系統的概念，並簡介一些如何從含雜訊語音中估計出雜訊的方法以及相減型語音強化法的處理方式。

2.1 相加性雜訊干擾下的語音強化系統

在時域中，我們將含雜訊語音$Y(n)$，看成是原來的乾淨語音$S(n)$，加上雜訊$N(n)$而組成的。由於雜訊是一個與語音訊號不相關(uncorrelated)的隨機程序(random process)，所以在功率頻譜密度(power spectral density)上，仍然維持著相加的(additive)特性。

$$P_Y = P_S + P_N \tag{2.1}$$

其中P_Y, P_S, P_N分別為含雜訊語音、乾淨語音以及雜訊的功率頻譜密度。由式(2.1)可知，假設我們能夠準確地知道雜訊的功率頻譜密度，就只要從含雜訊語音中將之減去，就能還原回乾淨語音了。由於語音與雜訊在短時間內具有穩定的特性，因此在進行語音強化時，可以先把含雜訊語音訊號切割成一串音框(frame)再進行處理。又因為功率頻譜密度可用訊號的離散傅利葉轉換計算出來，因此式(2.1)可改寫為

$$\frac{|Y(\lambda,k)|^2}{N^2} = \frac{|S(\lambda,k)|^2}{N^2} + \frac{|N(\lambda,k)|^2}{N^2} \tag{2.2}$$

其中

$$Y(\lambda,k)=\frac{1}{N}\sum_{n=0}^{N-1}y(\lambda N+n)e^{-j(2\pi/N)nk}=|Y(\lambda,k)|e^{j\phi_y(\lambda,k)}$$

N 代表一個音框的長度，λ,k 分別是音框索引(frame index)以及頻率槽索引(frequency bin index)。因為式(2.2)在等式左右兩邊都除上 N^2，所以可再把分母 N^2 消去，而改用能量頻譜取代功率頻譜就得到式(2.3)

$$|Y(\lambda,k)|^2=|S(\lambda,k)|^2+|N(\lambda,k)|^2 \tag{2.3}$$

由式(2.3)可知語音與雜訊的能量頻譜也是具有相加性的。語音強化的流程圖，如圖 2-1 所示：

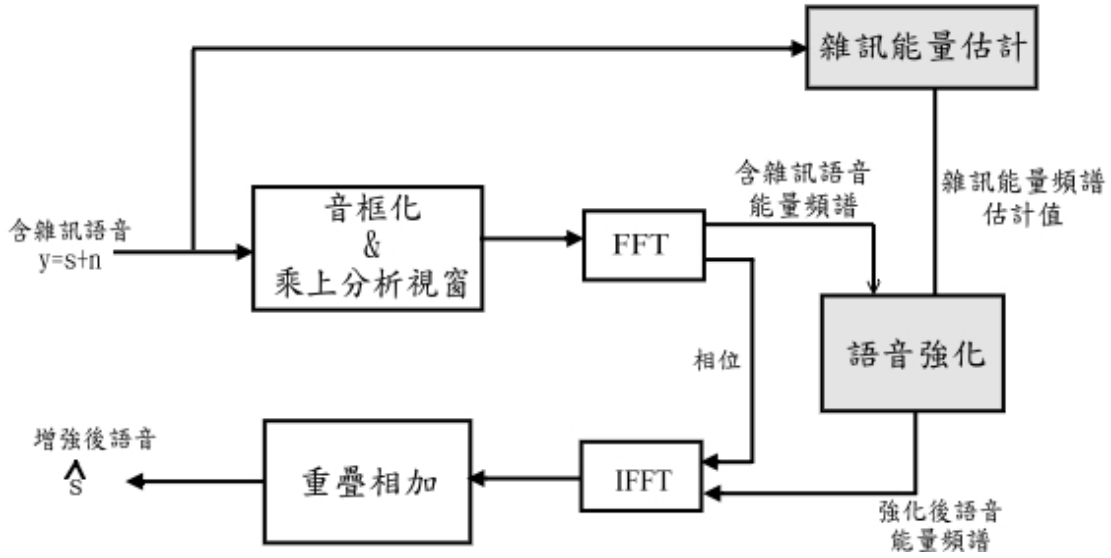

圖 2-1: 一般語音強化的系統方塊圖

在語音強化系統的處理流程中，首先含雜訊語音會被切成一個個音框，接下來會利用估計雜訊的方法，從含雜訊語音中估計出雜訊的能量頻譜。爲了降低音框內鄰近頻率的相互干擾，會先將音框乘上分析視窗後，再用 FFT 轉換到頻譜，得到其能量頻譜及相位。因為人類聽覺對相位的感覺較不靈敏，所以並不會對相位進行處理，而是只對能量頻譜進行強化。語音強化則利用先前估計得到的雜訊能量頻譜，把含雜訊語音中的雜訊部份消除掉，產生出近似於乾淨語音的能量頻譜，之後再與含雜訊語音的相位一起轉換回時域後疊加起來，即形成強化後的語

音。

假設我們能夠準確地估計出雜訊的能量頻譜，則語音強化只需將含雜訊語音的能量頻譜減去雜訊的能量頻譜就能得到乾淨語音的能量頻譜，再配合含雜訊語音的相位就能得到乾淨語音的頻譜，如式(2.4)所示

$$S(\lambda,k)=\sqrt{|Y(\lambda,k)|^2-|N(\lambda,k)|^2}\,e^{j\phi_y(\lambda,k)}=|S(\lambda,k)|e^{j\phi_y(\lambda,k)} \tag{2.4}$$

然而，在使用單一麥克風的情況下，輸入的訊號是已經被雜訊污染過的語音(noise-corrupt speech)，所以無法得到真正的雜訊能量頻譜，而只能試著從含雜訊語音中估計出來。可是當估計的雜訊與真正的雜訊有誤差時，經過式(2.4)的計算後，得到的就不是真正的語音訊號了。爲了克服雜訊估計不準確所造成的問題，在估計出雜訊之後，還要經過語音強化步驟，將估計出的雜訊依照某種規則(rule)或是準則(criteria)進行調整，才能從含雜訊語音中還原出較為準確的乾淨語音。我們接下來將分別簡介估計雜訊估計以及語音強化的一些方法。

2.2 雜訊估計

在進行語音強化前，首先要從含雜訊語音中估計出雜訊。由於頻域上，語音和雜訊訊號是相加在一起的，所以必須對語音及雜訊的特性進行一些假設才能夠估計出雜訊[7]，這些假設包括了：

(1) 語音與雜訊是互相獨立的(statistically independent)

(2) 在輸入的含雜訊語音中，語音訊號只會出現在某段時間

(3) 雜訊比語音更穩定(stationary)

雜訊估計的概念就是透過上述假設找出含雜訊語音中非語音或是語音暫停(speech pause)的區域，因為在這些區域中只有雜訊存在，所以可用來估計雜訊的特性。在做法上可分為以音框為基準(frame-based)或是以子頻帶為基準

(subband-based)兩類。

以音框為基準的方式需要使用語音偵測器(Voice Activity Detector, VAD)[8]把含雜訊語音音框分類為語音音框(speech frame)或是非語音音框(non-speech frame)，再統計那些非語音音框的特性去估計雜訊。因為語音相較於雜訊有較高的能量和較低的過零率，因此大部分的語音偵測器[9]都是用單一音框內的能量與過零率(zero-crossing rate)作為該音框是否包含語音訊號的判斷標準。但是在較低訊雜比時，由於語音和雜訊的能量太過接近，導致語音偵測器誤判音框分類的可能性大增，所以估計出來的雜訊會很不準確。

以子頻帶為基準的雜訊估計方式就是把整個頻域分為數個子頻帶，雜訊訊號則是在各個頻帶內獨立地去估計。這樣的估計方式是因為語音並不會存在於所有的頻率內(母音大多分佈在低頻頻段、子音大多分佈在高頻頻段)，所以相較於使用語音偵測器決定單一音框都是語音或是雜訊的方式而言，子頻帶為基準的方式能夠更準確地估計出雜訊[10]。

Martin在 1994 年提出了以最小統計量(minima statistic)為基準的子頻帶雜訊估計方法[6]。其觀念為人在說話時，每個音素(phone)之間一定會有短暫的停頓，即表示此時語音是不存在的，$|S(\lambda,k)|^2=0$。故此時的含雜訊語音的能量頻譜就會很接近雜訊的能量頻譜，見式(2.5)。

$$|Y(\lambda,k)|^2=|N(\lambda,k)|^2 \quad ， \quad 當 |S(\lambda,k)|^2=0 \tag{2.5}$$

其方法是在含雜訊語音中，往前尋找一段時間內的最小能量頻譜作為雜訊能量頻譜的估計值；由於採用最小值可能會低估雜訊，因此需要乘上一個倍率為偏差補償(bias compensate)。假設搜尋的範圍為目前音框到到前$L-1$個音框，則雜訊能量頻譜估計值$|\hat{N}(\lambda,k)|^2$如下式(2.6)所示：

$$\left|\hat{N}(\lambda,k)\right|^2 = \alpha \cdot \min\left\{\left|Y(\lambda-d,k)\right|^2\right\} \tag{2.6}$$

其中 $d = 0,1,2,...,L-1$，L 是最小值搜尋範圍，α 為偏差補償倍率。

由於式(2.6)中採用的補償方式是乘一個固定的倍率，所以調整後的雜訊也不會很準確。因此Martin於 2001 年提出了較好的偏差補償方式[6]，他利用含雜訊語音與先前估計出雜訊的變異數比值來調整偏差補償函數，使得新的估計值能更接近真正的雜訊。

2.3 相減型語音強化法

語音強化的目的是藉由估計出的雜訊，將含雜訊語音中的雜訊成分去除，而還原出原始乾淨語音的部分。在本論文中只針對相減型語音強化法進行討論。

相減型語音強化法在處理的方式上可分成兩類，第一類是頻譜刪減法(spectral subtraction) [11][15]，第二類是以統計為基礎(statistical-based)的強化方法[12]。頻譜刪減法的處理方式是把含雜訊語音的能量頻譜減去雜訊的能量頻譜估計值後，剩下的部分即為語音的能量頻譜，如式(2.7)所示：

$$\left|\hat{S}(\lambda,k)\right|^2 = \left|Y(\lambda,k)\right|^2 - \left|\hat{N}(\lambda,k)\right|^2 \tag{2.7}$$

其中 $\left|\hat{N}(\lambda,k)\right|^2$ 是雜訊的能量頻譜估計值，$\left|\hat{S}(\lambda,k)\right|^2$ 是強化後的語音能量頻譜。

以統計為基礎的語音強化法則是把乾淨語音想像為含雜訊語音通過一個濾波器後的輸出。式(2.8)為此方法在頻譜上的表示式。

$$\left|\hat{S}(\lambda,k)\right| = \left|Y(\lambda,k)\right| H(\lambda,k) \tag{2.8}$$

其中 $H(\lambda,k)$ 是濾波器的頻率響應，且 $H(\lambda,k)$ 是基於一些統計和最佳化估計的方式推導而來的。其實這兩類語音強化法只是推導的觀念不同，在實作上都可以是含雜訊語音通過一個濾波器的結果，就像頻譜刪減的式(2.7)可改寫為式(2.8)，其

中用到濾波器的頻率響應即為 $H(\lambda,k)=\sqrt{1-\frac{\left|\hat{N}(\lambda,k)\right|^2}{\left|Y(\lambda,k)\right|^2}}$。我們將分別介紹這兩類語音強化的一些方法。

2.3.1 頻譜刪減法

頻譜刪減法就是在頻譜上把含雜訊語音減去估計出的雜訊，即可得到近似於乾淨語音的訊號，能夠這樣做的原因是語音與雜訊在頻譜上具有相加的特性。式(2.9)是頻譜刪減的基本概念，其中因為能量頻譜不能為負值，所以當相減後若小於一個臨界值時，就用該臨界值取代。

$$\left|\hat{S}(\lambda,k)\right|^2=\begin{cases}\left|Y(\lambda,k)\right|^2-\left|\hat{N}(\lambda,k)\right|^2 & ,\text{if } \left|Y(\lambda,k)\right|^2-\left|\hat{N}(\lambda,k)\right|^2>\beta\left|\hat{N}(\lambda,k)\right|^2\\ \beta\left|\hat{N}(\lambda,k)\right|^2 & ,\text{otherwise}\end{cases} \tag{2.9}$$

其中 $\beta\left|\hat{N}(\lambda,k)\right|^2$ 是臨界值，一般 β 會設定為一個遠小於 1 的值。

爲了方便表達，我們修改頻譜刪減的公式，使得強化後的語音可由含雜訊語音的頻譜乘上一個函數 $G(\lambda,k)$ 後得到，因此式(2.9)可改寫為式(2.10)

$$\left|\hat{S}(\lambda,k)\right|=G(\lambda,k)\left|Y(\lambda,k)\right| \tag{2.10}$$

其中，$G(\lambda,k)=\begin{cases}\left(1-\left[\frac{\left|\hat{N}(\lambda,k)\right|}{\left|Y(\lambda,k)\right|}\right]^2\right)^{1/2} & ,(1+\beta)\left[\frac{\left|\hat{N}(\lambda,k)\right|}{\left|Y(\lambda,k)\right|}\right]^2<1\\ \left(\beta\left[\frac{\left|\hat{N}(\lambda,k)\right|}{\left|Y(\lambda,k)\right|}\right]^2\right)^{1/2} & ,\text{otherwise}\end{cases}$

由於雜訊$\left|\hat{N}(\lambda,k)\right|$是從含雜訊語音中估計出來的，與真正的雜訊存在著誤差，尤其是當訊雜比愈低時，此誤差也愈大，因此估計出的雜訊就愈不準確。爲了補償雜訊估計值不準確的問題，一般在得到雜訊的估計值後，會根據某種規則乘上一個變動倍率$\alpha(\lambda,k)$去調整雜訊值，以期在相減後能得到較準確的乾淨語音。式(2.11)即為加上此概念後的頻譜刪減法，這也是頻譜刪減法的一般式

$$G(\lambda,k)=\begin{cases}\left(1-\alpha(\lambda,k)\left[\dfrac{\left|\hat{N}(\lambda,k)\right|}{\left|Y(\lambda,k)\right|}\right]^{\gamma}\right)^{\kappa} & ,(\alpha(\lambda,k)+\beta)\left[\dfrac{\left|\hat{N}(\lambda,k)\right|}{\left|Y(\lambda,k)\right|}\right]^{\gamma}<1\\ \left(\beta\left[\dfrac{\left|\hat{N}(\lambda,k)\right|}{\left|Y(\lambda,k)\right|}\right]^{\gamma}\right)^{\kappa} & ,\text{otherwise}\end{cases} \tag{2.11}$$

其中$\alpha(\lambda,k)$是一個會依不同音框λ和不同頻率槽k而變動的調整參數。不同設定的(γ,κ)分別對應為不同的刪減方法，當$(\gamma,\kappa)=(1,1)$是強度刪減(magnitude subtraction)、$(\gamma,\kappa)=(2,1/2)$是功率刪減(power subtraction)，而$(\gamma,\kappa)=(2,1)$是溫尼濾波器(Wiener filter)(見 2.3.2 節)，這三種的刪減的方式差異並不會太大[14]，最主要的差異在於如何調整$\alpha(\lambda,k)$。

Lockwood和Boudy提出了一種調整$\alpha(\lambda,k)$的方法，稱之為非線性頻譜刪減(Nonlinear Spectral Subtraction, NSS)[15]。$\alpha(\lambda,k)$可設定為任意函數，爲了能去除掉較多的雜訊，此函數的特性需為在高訊雜比時，有較小的$\alpha(\lambda,k)$，藉以調低雜訊的估計值，避免刪減後造成語音太多的失真；而當在低訊雜比時，有較大的$\alpha(\lambda,k)$，藉以調提高雜訊的估計值，期望刪減後能消除更多的雜訊。

2.3.2 以統計為基礎的語音強化法

以統計為基礎的語音強化法不同於頻譜刪減法的推導觀念。其觀念是想要根據某種準則去設計一個濾波器，並對此準則進行最佳化推導，使得含雜訊語音經

過該濾波器後就能得到近似於乾淨語音的結果。在這類方法中，我們將介紹溫尼濾波(Wiener filtering)[11]以及最小平方誤差短時頻譜振幅估計(Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator) [12]。

溫尼濾波是假設語音與雜訊的頻譜係數均為高斯分佈下進行推導的，其目的是使強化後的訊號 $\hat{S}(n)$ 與原始語音訊號 $S(n)$ 的平方誤差為最小，即 $\min\left\|\hat{S}(n)-S(n)\right\|^2$。其推導出來的濾波器 $H_w(\lambda,k)$ 為

$$H_w(\lambda,k)=\frac{\mathrm{E}[\left|S(\lambda,k)\right|^2]}{\mathrm{E}[\left|S(\lambda,k)\right|^2]+\mathrm{E}[\left|N(\lambda,k)\right|^2]} \tag{2.12}$$

其中 $\mathrm{E}\{\left|S(\lambda,k)\right|^2\}$ 與 $\mathrm{E}\{\left|N(\lambda,k)\right|^2\}$ 分別為語音以及雜訊的能量頻譜平均值。式(2.13)定義了事前訊雜比(*a priori* SNR) $\eta(\lambda,k)$

$$\eta(\lambda,k)=\frac{\mathrm{E}[\left|S(\lambda,k)\right|^2]}{\mathrm{E}[\left|N(\lambda,k)\right|^2]} \tag{2.13}$$

因此濾波器 $H_w(\lambda,k)$ 可由式(2.12)改寫為式(2.14)

$$H_w(\lambda,k)=\frac{\eta(\lambda,k)}{1+\eta(\lambda,k)} \tag{2.14}$$

由於我們無法從含雜訊語音中得知真正語音訊號及雜訊的平均能量頻譜，所以事前訊雜比通常使用直接決定(decision-directed)的方式[12]去估算而得到，如式(2.15)所示：

$$\eta(\lambda,k)=\alpha_\eta\frac{\left|\hat{S}(\lambda-1,k)\right|^2}{\mathrm{E}[\left|\hat{N}(\lambda-1,k)\right|^2]}+(1-\alpha_\eta)\max(\gamma(\lambda,k)-1,0) \tag{2.15}$$

其中 α_η 為內插參數，$\gamma(\lambda,k)$ 為瞬時事後訊雜比(instantaneous *a posteriori* SNR)，其定義如式(2.16)

$$\gamma(\lambda,k)=\frac{|Y(\lambda,k)|^2}{\mathrm{E}[\,|N(\lambda,k)|^2\,]} \tag{2.16}$$

由式(2.15)中可知，$\eta(\lambda,k)$的估算方式是利用前一個音框算出的語音能量$|\hat{S}(\lambda-1,k)|^2$以及前一個音框的平均雜訊能量$\mathrm{E}[\,|\hat{N}(\lambda-1,k)|^2\,]$重新計算前一個音框的事前訊雜比，然後再與現在音框的瞬時事後訊雜比減1後內插計算而成。若$\gamma(\lambda,k)-1$為負值，就用0取代。

Ephraim 和 Malah 提出了最小平方誤差短時頻譜振幅估計法(Minimum Mean-Square Error Shot-Time Spectral Amplitude Estimator)，簡稱MMSE估計法。此方法亦假設語音與雜訊的頻譜係數均為高斯分佈，並使強化後的語音頻譜振幅$|\hat{S}(n)|$與原始語音頻譜振幅$|S(n)|$的平方誤差為最小。其最佳的瞬時語音頻譜大小估計值$|\hat{S}(\lambda,k)|$可用式(2.17)的條件平均值計算

$$|\hat{S}(\lambda,k)|=\mathrm{E}[|S(\lambda,k)|\,|\,Y(\lambda,k)] \tag{2.17}$$

爲了方便表示，令$A=|S(\lambda,k)|$，於是式(2.17)可寫為式(2.18)

$$\mathrm{E}\{A\,|\,Y(\lambda,k)\}=\int_{-\infty}^{\infty}A p\big(A\,|\,Y(\lambda,k)\big)dA \tag{2.18}$$

其中$p\big(A\,|\,Y(\lambda,k)\big)=p\big(|S(\lambda,k)|\,|\,Y(\lambda,k)\big)$代表給定雜訊語音訊號後，語音訊號頻譜大小的機率分佈。透過貝式定理(Bayes theorem)，式(2.18)可再寫為式(2.19)

$$\mathrm{E}\{A\,|\,Y(\lambda,k)\}=\frac{1}{p\big(Y(\lambda,k)\big)}\int_{-\infty}^{\infty}A p(A)p\big(Y(\lambda,k)\,|\,A\big)dA \tag{2.19}$$

其中$p\big(Y(\lambda,k)\big)$代表雜訊語音的機率分佈。

由於式(2.19)需要較高的運算複雜度，因此Wolf和Godsill在2001年提出了低複雜度的近似版本[13]。他把原先估計語音訊號的頻譜強度A改為估計語音訊號

的頻譜能量 A^2 ，如式(2.20)

$$\left|\hat{S}(\lambda,k)\right|^2 = \mathrm{E}[A^2 \mid Y(\lambda,k)] \tag{2.20}$$

再假設 $p\left(A^2 \mid Y(\lambda,k)\right)$ 是 chi-square 的分佈，進而推導出較簡單的計算方式，如式(2.21)所示

$$\left|\hat{S}(\lambda,k)\right|^2 = \left[\frac{\eta(\lambda,k)}{1+\eta(\lambda,k)}\left(\frac{1}{\gamma(\lambda,k)}+\frac{\eta(\lambda,k)}{1+\eta(\lambda,k)}\right)\right]\left|Y(\lambda,k)\right|^2 \tag{2.21}$$

因此相對於式(2.8)，MMSE 估計法採用的濾波器 $H_{MMSE}(\lambda,k)$ 為式(2.22)

$$H_{MMSE}(\lambda,k) = \sqrt{\frac{\eta(\lambda,k)}{1+\eta(\lambda,k)}\left(\frac{1}{\gamma(\lambda,k)}+\frac{\eta(\lambda,k)}{1+\eta(\lambda,k)}\right)} \tag{2.22}$$

透過比較式(2.14)和式(2.22)，溫尼濾波與MMSE估計法的差別就是MMSE估計法會利用瞬間事後訊雜比的估計值 $\gamma(\lambda,k)$，來調整濾波器的頻率響應。Ephraim和Malah[12]證明了在高訊雜比時，MMSE的濾波器頻率響應會接近於溫尼濾波器，而Wolf和Godsill提出簡化的MMSE濾波器也具有相同的關係。如式(2.22)所示，當瞬間事後訊雜比 $\gamma(\lambda,k)$ 趨近無限大時，其濾波器型式就會近似於溫尼濾波；在低訊雜比時，則能減少語音訊號的衰減。MMSE估計法與NSS均是以訊雜比作為調整的參數，這也是大部分語音強化演算法使用的調整策略。

在簡介語音強化所需了解的背景知識以及一些方法後，我們會在接下來的第三章中說明本論文提出的語音強化法。

第三章 強化型 MMSE 語音強化法

大部分的語音強化技術，如 NSS 和 MMSE 估計法，都是估計出在短時間內訊雜比的大小，去調雜訊值或是濾波器的係數。但是在低訊雜比的情況下，由於雜訊值的估計不準確，連帶使得估計出訊雜比也有誤差，而導致這種調整方式對於提升辨識率的效果非常有限。

語音頻譜具有隨時間變化的現象，包含了一些語音特徵，像是基週軌跡(pitch contour)、共振峰特性(formant)等也是具有隨時間變化的特性，但是大多數的語音強化法並沒有考量到這個特性，所以經過強化後的語音，可能會失去一些頻譜隨時間變化的資訊，導致抽取出的語音特徵參數在時間軸上的資訊失真，而使得辨識率下降。因此本論文提出了一種考量了語音頻譜隨時間變化的語音強化方法，藉由突顯語音訊號並壓抑雜訊訊號，使強化後的語音仍能保留時間上的資訊，以減少語音特徵參數的失真，而提高辨識率。本論文提出了一個以 MMSE 估計法為基礎，並結合考量語音隨時間變化觀念的語音強化法，稱之為強化型 MMSE(enhanced MMSE, EMMSE)語音強化法。

3.1 基本概念

含雜訊語音中包含了語音訊號以及非語音訊號，而非語音訊號的部分因為不包含語音，而只包含雜訊，因此非語音訊號也就是雜訊訊號。透過一些雜訊的估計法，我們可以大約估計出哪些是語音訊號，哪些是雜訊訊號。語音強化的目的就是希望能夠突顯語音訊號，並壓抑雜訊訊號，以減少雜訊對語音的干擾。EMMSE 即根據此基本概念，對含語音訊號，加大其變異度(variance)，藉以突顯

語音訊號；對雜訊訊號，減小其變異度，藉以壓抑雜訊訊號。EMMSE 針對每個音框的每個頻率槽(frequency bin)進行調整，其調整方法如下

$$\hat{S} = \tilde{S} + f(x)(\tilde{S} - E[\tilde{S}_L]) \tag{3.1}$$

其中 $\hat{S}$ 是經由上述原理調整後的語音訊號，$\tilde{S}$ 是我們先前估計出的語音訊號，$E[\tilde{S}_L]$ 是前 $L-1$ 個到現在這個音框的語音估計的平均值(mean)。$f(x)$ 是調整函數，其式如下。

$$f(x) = \frac{1}{1+\exp\left(-\frac{(c-x)}{a}\right)} - \frac{1}{2} \tag{3.2}$$

$f(x)$ 的值限制在 1/2 到-1/2 之間，其中 a 是用來控制 $f(x)$ 隨 x 變化速度的參數，當 a 愈大，$f(x)$ 變化愈平緩。x 是用前 $L-1$ 個到現在這個音框所計算出來的估計雜訊變異度 $\sigma^2_{\hat{N}_L}$ 和含雜訊語音變異度 $\sigma^2_{Y_L}$ 的比值：

$$x = \frac{\sigma^2_{\hat{N}_L}}{\sigma^2_{Y_L}} \tag{3.3}$$

c 是一個用來判斷此頻率槽的變動是由語音主導(speech dominate)或是雜訊主導(noise dominate)的門檻值。當 $x < c$ 時，也就是此頻率槽的變動是由語音訊號主導，此時 $f(x) > 0$。因此經由式(3.1)調整後，語音訊號的變異度會加大；當 $x > c$ 時，也就是此頻率槽的變動是由雜訊訊號主導，此時 $f(x) < 0$。因此經由式(3.1)調整後，雜訊訊號的變異度會減小。

3.2 方法分析

EMMSE 語音強化法是以 MMSE 估計法為基礎，並考慮語音與雜訊在某段時間中的變動程度(變異度)對濾波器進行調整。式(3.1)中的 $\tilde{S}$ 是使用 MMSE 估計

法計算出的語音估計值。接下來將討論 EMMSE 與溫尼濾波和 MMSE 估計法三者之間的關係。爲了書寫方便，我們把頻率槽索引 k 與音框索引 λ 省略。因此溫尼濾波以及 MMSE 估計法可由式(2.12)與式(2.15)改寫為式(3.4)與式(3.5)

$$H_w = \frac{\eta}{1+\eta} \tag{3.4}$$

$$H_{MMSE} = \sqrt{\frac{\eta}{1+\eta}(\frac{1}{\gamma}+\frac{\eta}{1+\eta})} = \sqrt{H_w(\frac{1}{\gamma}+H_w)} \tag{3.5}$$

其中 η 和 γ 分別為事前訊雜比和瞬時事後訊雜比，其式如下：

$$\eta = \frac{\mathrm{E}[|S|^2]}{\mathrm{E}[|N|^2]} \quad , \quad \gamma = \frac{|Y|^2}{\mathrm{E}[|N|^2]} \tag{3.6}$$

式(3.1)經由推導後，可得 EMMSE 語音強化法的濾波器頻率響應如式(3.7)表示，詳細的推導過程，請見附錄說明。

$$H_{EMMSE} = \sqrt{H_w(\frac{1}{\gamma}+H_w+H_w f(x)(1-\frac{\mathrm{E}[Y_L]}{Y}))} \tag{3.7}$$

由式(3.7)與式(3.5)相比，EMMSE 語音強化法相較於 MMSE 估計法的差別在於 H_{EMMSE} 多了一項 $H_w f(x)(1-\frac{\mathrm{E}[Y_L]}{Y})$。我們依照 $f(x)$ 可能出現的三種情形來討論濾波器頻率響應 H_{EMMSE} 的調整方式，並與 MMSE 估計法的濾波器進行比較：

(1) $f(x)>0$ ，也就是 $\frac{\sigma^2_{\hat{N}_L}}{\sigma^2_{Y_L}} < c$ ，代表語音變動的程度大於雜訊變動的程度，即此頻率槽屬於語音主導。由於 $f(x)$ 是正值，所以當語音頻譜值(Y)大於其鄰近的平均值($\mathrm{E}[Y_L]$)時，就會加大語音頻譜值；而當語音頻譜值小於其平均值時，就會減低語音頻譜值。調整公式如式(3.8)：

$$\begin{cases} Y > \mathrm{E}[Y] \Rightarrow H_{EMMSE} > H_{MMSE} \Rightarrow \hat{S} > \tilde{S} \\ Y < \mathrm{E}[Y] \Rightarrow H_{EMMSE} < H_{MMSE} \Rightarrow \hat{S} < \tilde{S} \end{cases} \tag{3.8}$$

圖 3-2 為 MMSE 與 EMMSE 調整後訊號的比較圖。其中圖(a)中的 ‒·‒·‒ 為含雜訊語音，‒ ‒ ‒ 為所估計的雜訊，實線為 MMSE 估計出的語音訊號；圖(b)中的實線為 MMSE 估計出的語音訊號，‒ ‒ ‒ 為含雜訊語音的平均值，‒·‒·‒ 為 EMMSE 估計出的語音訊號

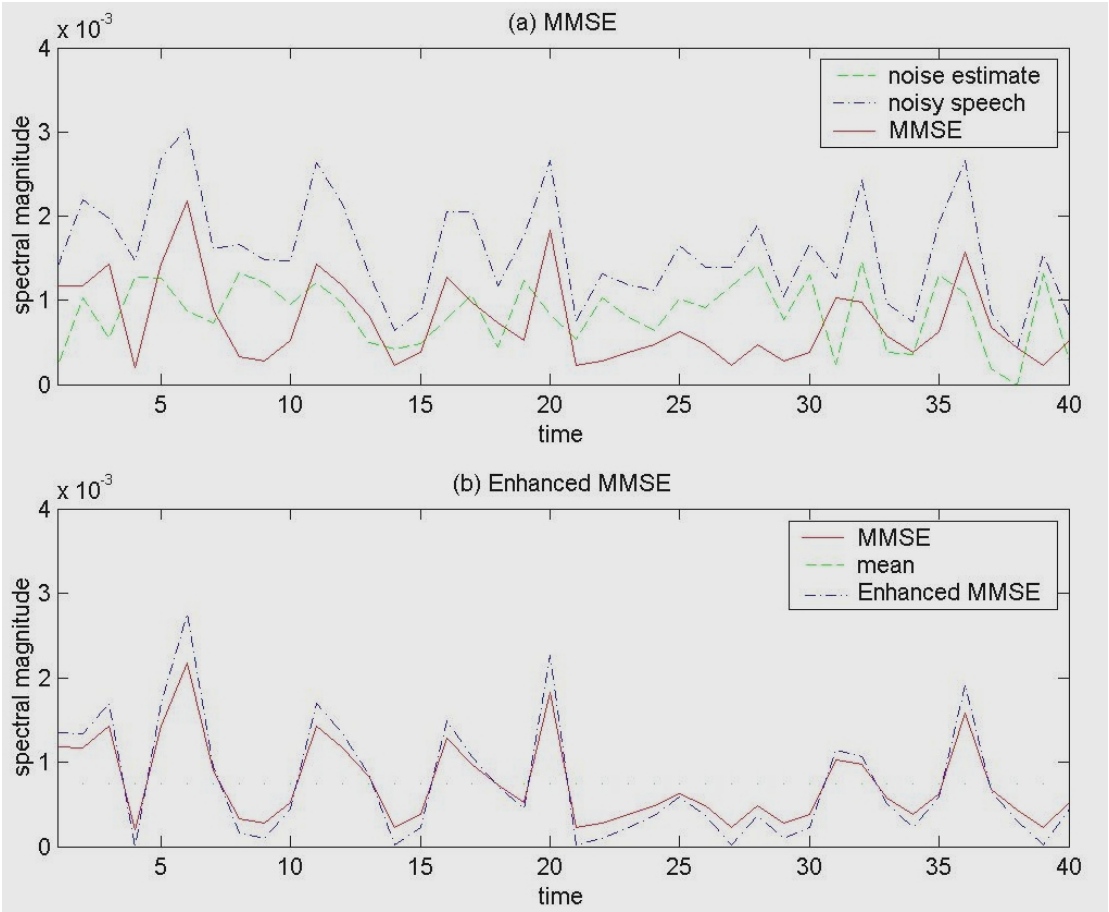

圖 3-1: 語音主導的調整概念圖。(a)MMSE 估計法的強化後結果；(b)Enhanced MMSE 強化後的結果，相比於 MMSE 估計法，會使訊號頻譜值遠離其平均值

由圖 3-1(b)中可看出 EMMSE 的調整方式確實能增加語音訊號的變異度。

(2) $f(x) < 0$ ，也就是 $\frac{\sigma^2_{\hat{N}_L}}{\sigma^2_{Y_L}} > c$ ，代表雜訊變動的程度大於語音變動的程度，即此頻率槽屬於雜訊主導。由於 $f(x)$ 是負值，所以當語音頻譜值(Y)小於其平均值

($\mathrm{E}[Y_L]$)時，就會減低其語音頻譜值；而當語音頻譜值小於其平均值時，就會加大語音頻譜值。調整公式如式(3.9)：

$$\begin{cases} Y > \mathrm{E}[Y] \Rightarrow H_{EMMSE} < H_{MMSE} \Rightarrow \hat{S} < \tilde{S} \\ Y < \mathrm{E}[Y] \Rightarrow H_{EMMSE} > H_{MMSE} \Rightarrow \hat{S} > \tilde{S} \end{cases} \tag{3.9}$$

圖 3-3 為 MMSE 與 EMMSE 調整後訊號的比較圖。其中圖(a)中的 為含雜訊語音， 為所估計的雜訊，實線為 MMSE 估計出的語音訊號；圖(b)中的實線為 MMSE 估計出的語音訊號， 為含雜訊語音的平均值， 為 EMMSE 估計出的語音訊號

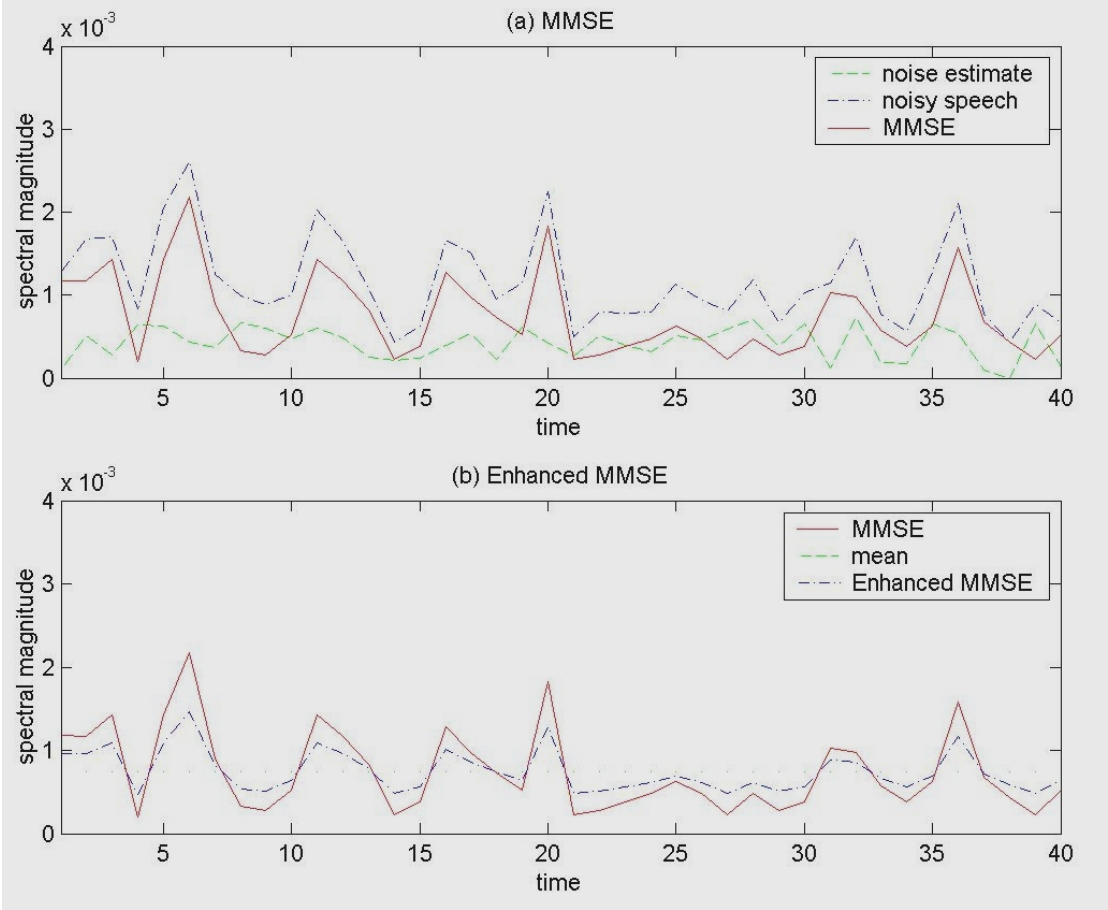

圖 3-2: 雜訊主導的調整概念圖。(a)MMSE 估計法的強化後結果；(b)Enhanced MMSE 強化後的結果，相比於 MMSE 估計法，會使訊號頻譜值靠近其平均值

由圖 3-2(b)中可看出 EMMSE 的調整方式確實能降低雜訊訊號的變異度。

(3) $f(x)=0$，也就是 $\frac{\sigma^2_{\hat{N}_L}}{\sigma^2_{Y_L}}=c$，代表雜訊變動與語音變動的程度是相似的，即此頻率槽是屬於語音與雜訊共同主導。在此情況下，本方法並不會做多餘的調整，也就是與 MMSE 估計法有相同的結果。

由上述三種情形可知，EMMSE 的演算法確實能依照頻率槽是語音主導或是雜訊主導的特性去調整濾波器的頻率響應，而給含雜訊語音不同加權。透過突顯語音訊號並壓抑雜訊訊號的方式，能消除強化後語音的特徵參數受到雜訊干擾而失真的影響，以達到提高辨識率的目的。在接下來的第四章中會進行一些實驗來驗證 EMMSE 語音強化法的效能。

第四章 實驗與結果討論

本論文模擬在相加性雜訊環境下，進行中文數字語音辨識的實驗。首先會驗證我們的方法在採用不同維度的語音參數下，對辨識率的影響。接著再與其他語音強化法比較進行效能比較。

4.1 實驗平台設計

本章實驗是依據AURORA提出的實驗架構[18]來設計的。AURORA工作小組是隸屬於ESTI委員會中的組織，其工作是定義並推動分散式語音辨識系統的標準。此實驗設計的目的是用來評估語音辨識系統的前端(front-end)模組在雜訊環境下的效能。整體的系統的概念圖如下：

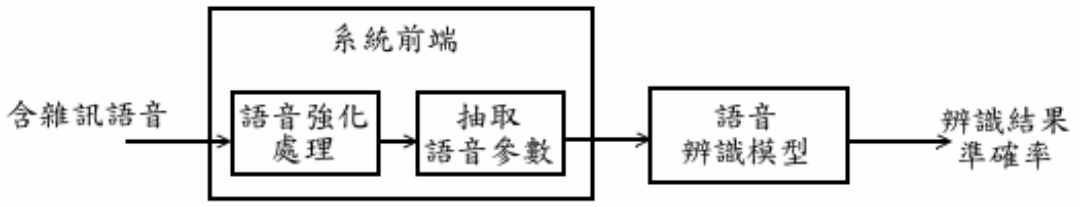

圖 4-1: 實驗系統概念圖

AURORA 實驗設計了兩種不同訓練模式的語音辨識模型，其測試語音為人工合成包含了不同種類雜訊以及不同訊雜比的含雜訊語音。輸入的含雜訊語音需要先經過語音強化的處理，去除掉雜訊對語音的影響之後，再抽取語音特徵參數，傳送到辨識模型進行辨識。

4.1.1 語音特徵參數抽取

在進行語音辨識時，抽取語音特徵參數的目的是要從語音訊號中抽取出足夠的資訊，用以分辨出不同音素之間的差異。由於語音訊號在時域上的變化是很快速的，可是其頻譜在一個短時間之內具有穩定的特性，因此語音特徵參數的抽取也是以音框為單位在其頻譜上進行處理。目前的研究發現[16]，用梅爾刻度倒頻譜係數(Mel-scaled cepstrum coefficient, MFCC)作為語音特徵參數，在一般的環境都能有不錯的辨識結果，所以本論文實驗所採用的語音特徵參數為梅爾刻度倒頻譜係數以及音框的對數能量。詳細的語音特徵參數的抽取流程[16]如下圖:

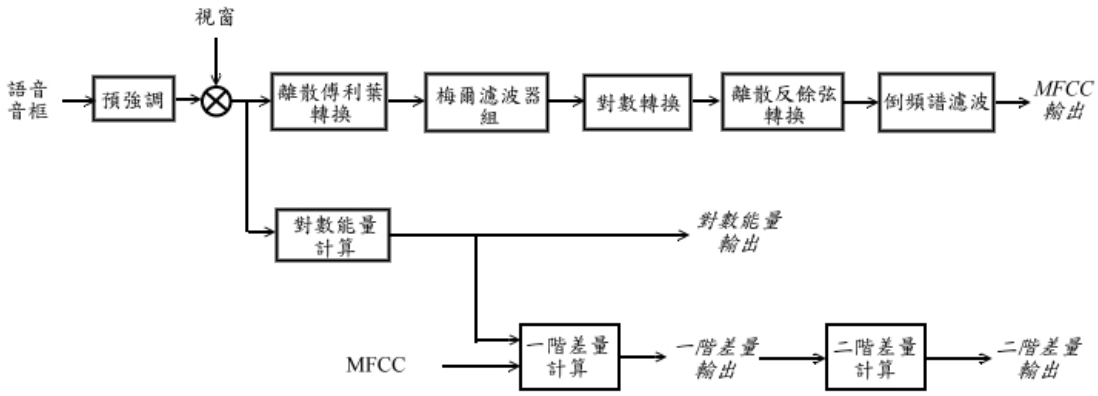

圖 4-2: 語音特徵參數的抽取流程圖

圖中斜體字的部分是語音特徵參數的輸出。詳細的抽取步驟說明如下：

(1) 預強調(pre-emphasis)

聲道(vocal tract)的特性就像一個低通濾波器，會使得語音高頻成分的振幅比低頻成分來的小。因此，爲了克服此效應，語音需要先經過預強調處理，使高低頻成分的大小差不多。預強調就是先讓語音訊號先通過一個高通濾波器，壓抑語音的低頻成分，並提升高低頻成分。通常使用的高通濾波器是一階的有限長度脈衝響應(Finite Impulse Response)，如式(4.2)。

$$S_i^{'}(n) = S_i(n) - aS_i(n-1) \tag{4.2}$$

其中 i 代表第 i 個音框，a 是預強調參數，，本論文中設定為 $a = 0.97$ 。

(2) 視窗處理(windowing)

爲了降低音框內鄰近頻率的相互干擾，會先將音框乘上分析視窗。在此用的是漢明(hamming)視窗，如式(4.3)所示

$$w(n) = \begin{cases} 0.54\text{-}0.46\cos\left(\dfrac{2\pi n}{N-1}\right), \text{n} = 0,1,..., \text{N -1} \\ 0 \qquad\qquad\qquad\qquad\quad , \text{otherwise} \end{cases} \tag{4.3}$$

其中 N 代表音框長度，乘上視窗後的訊號仍以 $S_i^{'}(n)$ 表示。

(3) 離散傅利葉轉換(Discrete Fourier Transform)

利用離散傅利葉轉換，求得語音音框在頻譜上的振幅值，如式(4.4)所示。

$$\left|S_i^{'}(e^{j2\pi k/N})\right| = \left|\sum_{n=0}^{N-1} S_i^{'}(n)e^{j2\pi k/N}\right| \tag{4.4}$$

(4) 濾波器組(filter bank)處理

爲了模擬人耳對不同頻率訊息的接收行為，再使用濾波器組將語音頻譜分成數組(bank)。在此用的濾波器組是梅爾三角濾波器組，其頻率軸的間隔是以梅爾刻度來計算，其與頻率軸間的轉換公式如式(4.5)。

$$Mel(f) = 1125\ln(1 + \frac{f}{700}) \tag{4.5}$$

(5) 對數轉換

人耳對頻譜上振幅的敏感度具有對數的效應。為了模擬此效應，將語音頻譜經過梅爾三角濾波器組後的輸出再取對數。我們以 $m_i\,(j)$ 表示第 j 個濾波器組的輸出對數值。

(6) 離散反餘弦轉換(Discrete Cosine Transform)

將濾波器組輸出的對數值，進行離散反餘弦轉換，就能獲得梅爾刻度倒頻譜係數，以 $c_i(k)$ 表示。轉換公式如式(4.6)

$$c_i(k)=\sum_{j=1}^{M} m_i(j)\cos\left(\frac{k\pi}{M}(j-0.5)\right),k=0,1,2,... \tag{4.6}$$

(7) 倒頻譜濾波(cepstrum liftering)

梅爾倒頻譜係數雖具有每維度之間是不相關的優點，但是卻存在著高維度係數的數值比較小的現象，這會導致由低維度到高維度係數數值的變異過大的缺點。爲了解決此問題，再透過一個倒頻譜濾波的處理。處理方式如式(4.7)所示

$$c_i^{'}(k)=(1+\frac{L}{2}\sin\frac{n\pi}{L})c_i(k) \tag{4.7}$$

其中 L 為倒頻譜濾波參數，本論文中設定為 $L=22$ 。

(8) 能量計算

因為不同音素會有不同的能量，所以額外加上一維的音框能量作為語音參數的一部分，即計算方式如式(4.8)

$$E_i=\sum_{n=0}^{N-1}\left(S_i^{'}(n)\right)^2 \tag{4.8}$$

(9) 一階差量計算

因為語音具有隨時間變化的性質，但是梅爾刻度倒頻譜係數以及音框能量所代表的只是單一音框內的資訊，所以再透過不同音框的係數進行差量計算，得到另外一組係數。此組係數稱之為一階差量的梅爾刻度倒頻譜係數和一階差量的能量係數，如式(4.9)和式(4.10)

$$\Delta c_i^{'}(k)=\sum_{p=1}^{P}\frac{\left(c_{i+p}^{'}(k)-c_{i-p}^{'}(k)\right)}{2\sum_{q=1}^{p}q^2} \tag{4.9}$$

$$\Delta E = \sum_{p=1}^{P} \frac{\left(E_{i+p} - E_{i-p}\right)}{2\sum_{q=1}^{p} q^2} \tag{4.10}$$

其中 P 為差量間距，本論文中設定為 $P = 2$ 。

(10) 二階差量計算

爲了得到更進一步語音隨時間變化的資訊，故進行二階差量計算。其計算方式也與一階差量的計算方式相同，如式(4.11) 和式(4.12)

$$\Delta\Delta c_i^{'}(k) = \sum_{p=1}^{P} \frac{\left(\Delta c_{i+p}^{'}(k) - \Delta c_{i-p}^{'}(k)\right)}{2\sum_{q=1}^{p} q^2} \tag{4.11}$$

$$\Delta\Delta E = \sum_{p=1}^{P} \frac{\left(\Delta E_{i+p} - \Delta E_{i-p}\right)}{2\sum_{q=1}^{p} q^2} \tag{4.12}$$

本實驗使用的語音特徵參數，如表 4-1 之設定。

表 4-1: 語音特徵參數抽取之設定

語音取樣頻率	8kHz
音框長度(frame length)	32ms
音框位移(frame shift)	16ms
使用分析視窗(window)	漢明(hamming)視窗
預強調(pre-emphasis)參數	0.97
能量正規化	有
語音特徵參數(13 維) 語音特徵參數(26 維) 語音特徵參數(39 維)	12 維梅爾刻度倒頻譜係數(MFCC)+1 維對數能量 13 維語音特徵參數及其一階差量(delta)參數 13 維語音特徵參數及其一階與二階差量參數
倒頻譜濾波器參數	22

4.1.2 語音聲學模型架構

實驗後端的語音辨識模型是採用HTK(Hidden Markov Model Tool Kit)3.2.1版[20]建構出以隱藏式馬可夫模型為基礎的模型[21]。模型的架構是透過語音特徵參數出現的機率以及狀態轉移的機率決定出通過哪個模型的機率最大，就最有可能是辨識的結果。一個擁有三個狀態的模型，其狀態轉移的概念如圖 4-2 所示：

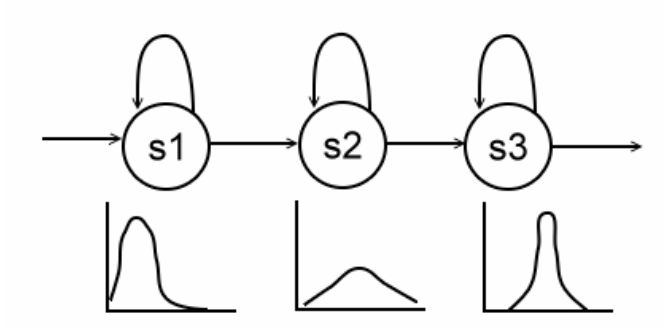

圖 4-3: 馬可夫模型-狀態轉移概念圖

我們選用中文數字的聲母與韻母作為模型的基本單位。每個聲母模型內有三個狀態，韻母模型則有四個狀態，每個狀態均是採用 32 個高斯分佈的混合高斯模型(Gaussian Mixture Model)。此外，還加上了一個四個狀態的靜音模型(sil)，用以過濾語音中靜音的部分。本實驗共使用了 18 個語音模型進行辨識，每個數字拼音及對照的模型如表 4-2 所示：

表 4-2: 數字拼音及語音模型對照表

數字拼音	語音模型	數字拼音	語音模型
(零)ㄌㄧㄥ	l_i ing	(五) ㄨ	sic_u u
(一) ㄧ	sic_i i	(六) ㄌㄧㄡ	l_i iou
(二) ㄦ	sic_e er	(七) ㄑㄧ	chi_i i
(三) ㄙㄢ	s_a an	(八) ㄅㄚ	b_a a
(四) ㄙ	s_empt empt2	(九) ㄐㄧㄡ	ji_i iou
(靜音) sil	Sil		

4.1.3 辨識效能之評估

實驗中比較了幾個不同的語音強化法。首先對輸入至系統的含雜訊語音進行強化後，再抽取語音特徵參數交由後端的語音辨識系統進行辨識，辨識的結果作為語音強化技術效能的評估標準。同時本論文中的實驗都只考慮不計聲調之音節的準確率(accuracy, acc)為主，而實驗數據中皆以百分比表示，其計算方式如下：

$$acc = \frac{N - E_S - E_D - E_I}{N} \times 100\% \tag{4.13}$$

其中 N 是輸入測試語音中所有音節(syllable)的個數，E_S,E_D 和 E_I 分別是辨識結果中替代型錯誤(substitution error)、刪除型錯誤(deletion error)和插入型錯誤(insertion error)的音節個數。

4.2 語音資料庫簡介

實驗中採用的語音資料庫是MAT2000 的子資料庫，MATDB-2[19]。此子資料庫錄製了近兩千人的連續數字語音，每人一句、一句七個數字，其取樣頻率為8kHz，我們用此作為乾淨語音。雜訊訊號是來自AURORA2 資料庫[18]，包含了八種在日常生活中經常會面臨到的雜訊，分別為地下鐵雜訊(subway)、人聲雜訊(babble)、車子雜訊(car)、展覽室雜訊(exhibition hall)、餐廳雜訊(restaurant)、街道雜訊(street)、機場雜訊(airport)和車站雜訊(train station)，取樣頻率也是 8kHz。

4.2.1 雜訊特性分析

同一種語音強化法在面對不同種類的雜訊時，會因為雜訊特性的不同，而有不同的強化效果。一般而言，具有不穩定的(non-stationary)的特性或是主要能量分佈在與語音相同頻帶上的雜訊是屬於較不易消除的。不穩定的特性代表了雜訊

的頻譜會隨時間呈現劇烈變化，而有忽隱忽現(impulsive)的情形，因此很難準確地估計出雜訊，易導致語音強化後的效果不明顯。當雜訊與語音分佈在相同頻帶上，也很難分辨出雜訊和語音的差異，亦增加了語音強化的困難度。由於從聲譜圖(spectrogram)能觀察出訊號頻譜隨時間變化的穩定程度，以及主要能量分佈的頻譜範圍，因此我們在圖 4-4 到圖 4-11 畫出實驗中所使用的八種雜訊之聲譜圖，並探討不同種類雜訊之特性。

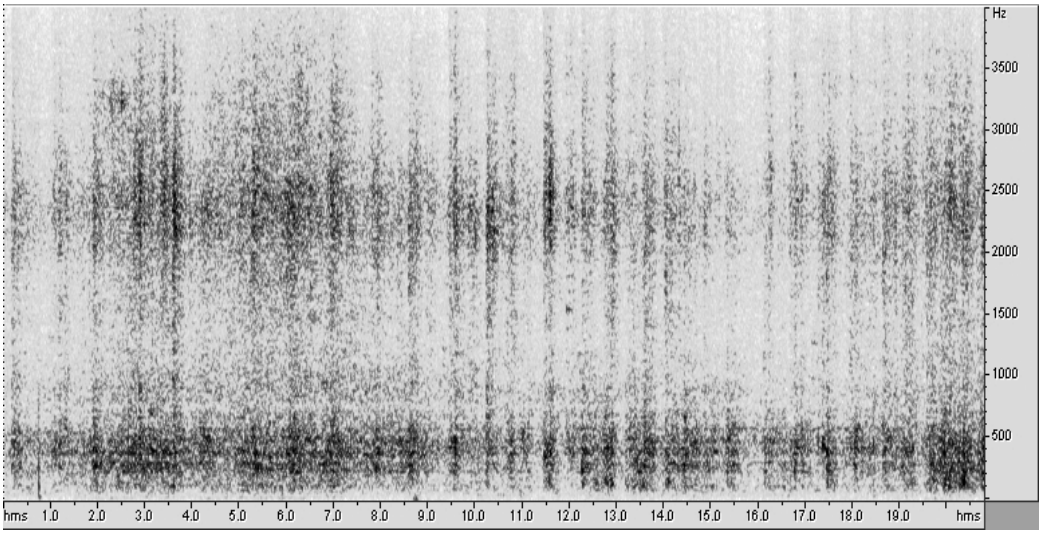

圖 4-4：地下鐵雜訊的聲譜圖

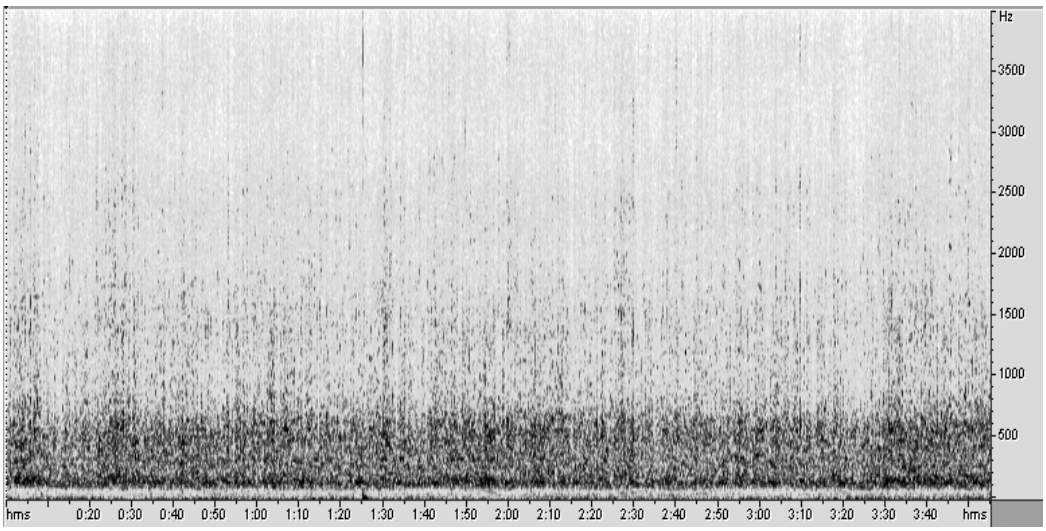

圖 4-5：人聲雜訊的聲譜圖

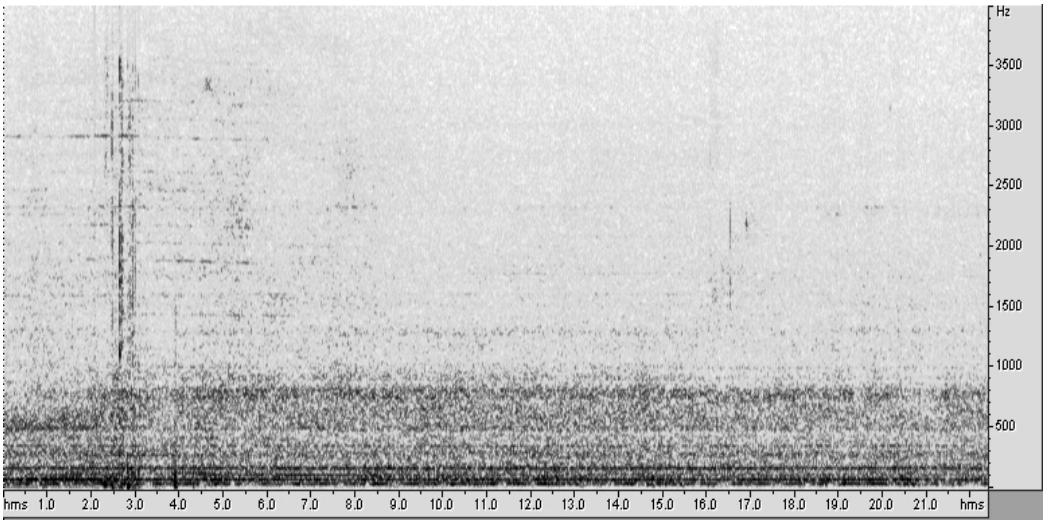

圖 4-6：車子雜訊的聲譜圖

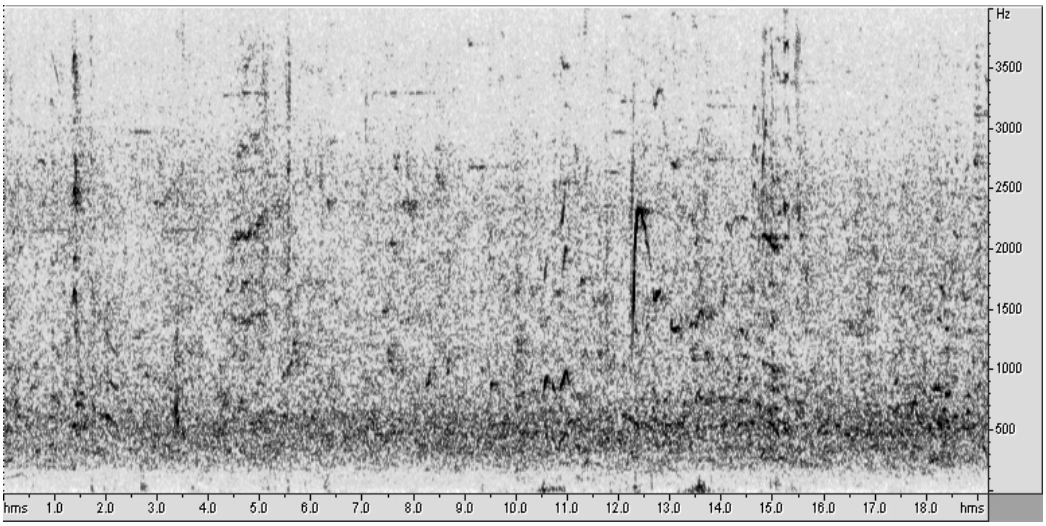

圖 4-7：展覽室雜訊的聲譜圖

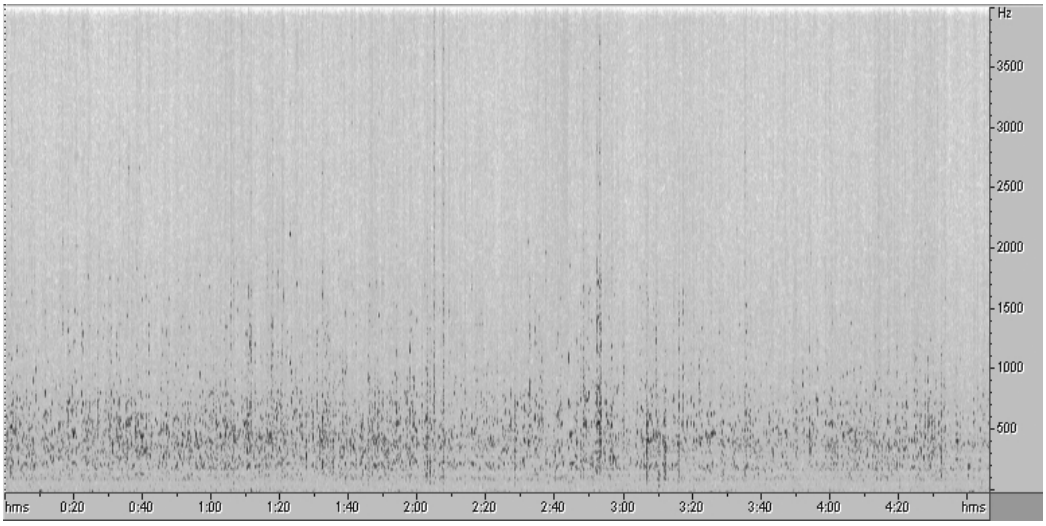

圖 4-8：餐廳雜訊的聲譜圖

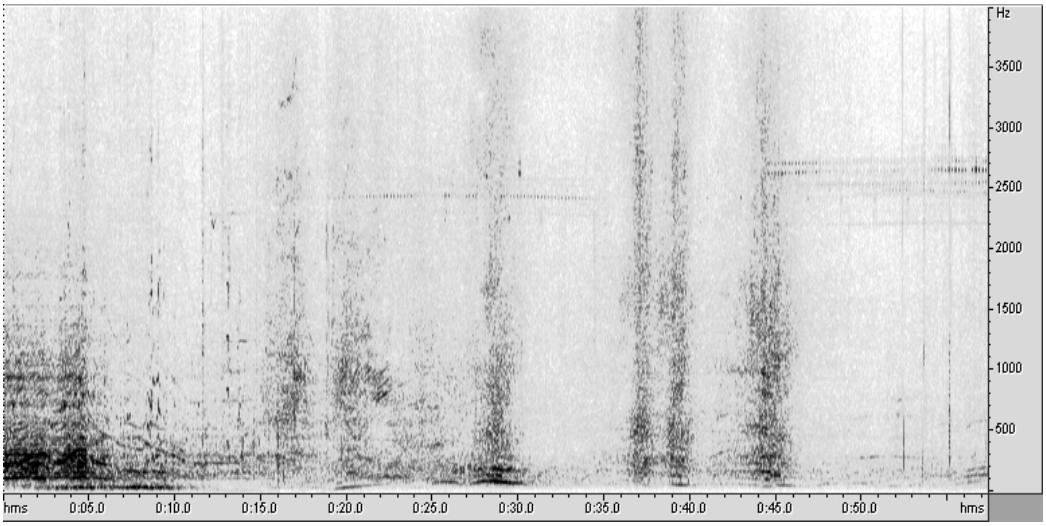

圖 4-9：街道雜訊的聲譜圖

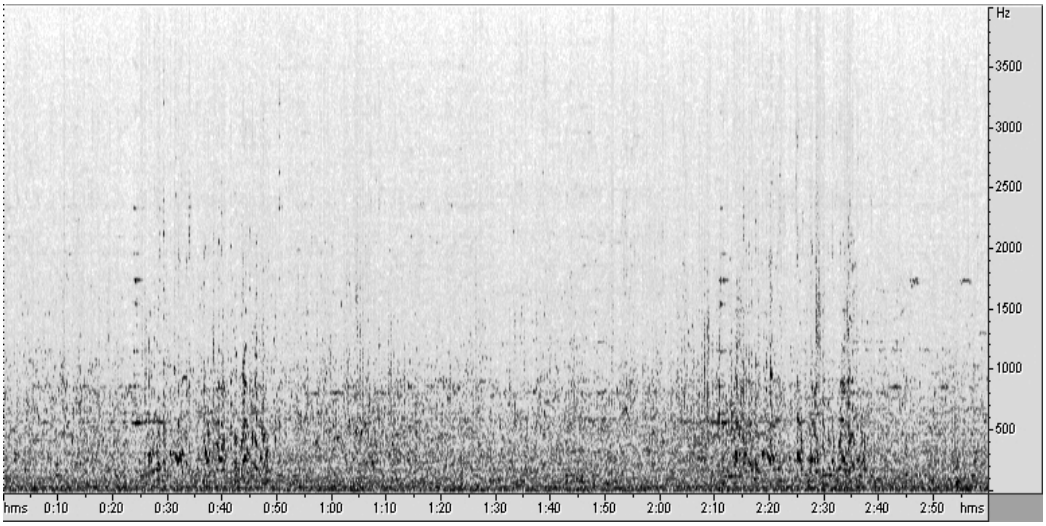

圖 4-10：機場雜訊的聲譜圖

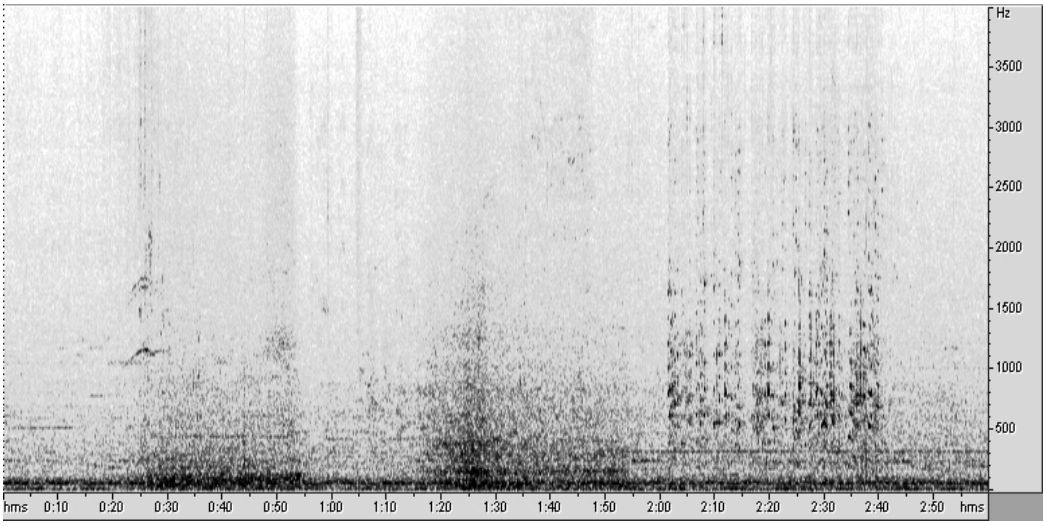

圖 4-11：車站雜訊的聲譜圖

我們依雜訊的穩定性以及主要能量分佈的頻率範圍，對雜訊的特性進行說明：

(1) 依照穩定性可分成不穩定雜訊或是穩定雜訊：

- 不穩定雜訊：最不穩定的雜訊是地下鐵雜訊和餐廳雜訊，接下來是人聲雜訊，其次是機場雜訊，而街道雜訊與車站雜訊的不穩定特性是其中較不明顯的。
- 穩定雜訊：最穩定的是車子雜訊，再來是展覽室雜訊。

(2) 依照主要能量分佈的頻率範圍可分成類人聲雜訊與非類人聲雜訊：

- 類人聲雜訊：人聲雜訊與餐廳雜訊均屬此類。
- 非類人聲雜訊：主要能量分佈在中高頻(500~2500Hz)的有展覽室雜訊，而分佈在中低頻 (50~1000Hz) 的有機場雜訊和街道雜訊；分佈在最低頻(150Hz 以下)的有車子雜訊和車站雜訊。而地下鐵雜訊則是分佈在高頻與低頻的範圍內。

4.2.2 語音訓練模型及測試語音

本實驗依照 4.1.3 節的語音聲學模型設計了兩種模型的訓練方式：

(1) 乾淨語音訓練(Clean Speech Training, CST)模型 – 此模型只使用乾淨語音作為訓練語音。

(2) 多環境訓練(Multi-Condition Training, MCT)模型 – 此模型一併使用乾淨和含雜訊的語音為訓練語音。

CST 模型從 MATDB-2 中抽取了 1726 句的乾淨語音進行訓練(其中男生 753 人、女生 973 人)。MCT 模型也是使用與 CST 模型相同的 1726 句語音，但是需先把它們等分為 20 個子集合(每個子集合約 86 句)，每個子集合分別加上不同種類的雜訊，組成不同訊雜比後，才能作為 MCT 模型的訓練語音。這 20 個子集

合各自代表了 4 種不同種類的雜訊以及 5 種不同的訊雜比，這四種雜訊分別為地下鐵雜訊、人聲雜訊、車子雜訊和展覽室雜訊；而 5 種訊雜比為 20dB、15dB、10dB、5dB 和未加雜訊。

含雜訊的測試語音是從 MATDB-2 中選擇與訓練語音不同的 320 句(其中男女各 160 人)，把它們等分成 4 個子集合(每個子集合 80 句，其中男女各 40 人)。每個子集合再與一種雜訊依照六種不同的訊雜比(20dB、15dB、10dB、5dB、0dB、-5dB)組成含雜訊語音。依照使用雜訊的種類，可把測試語音分成 Set A 與 Set B；Set A 所用的雜訊與訓練 MCT 模型時所使用的一樣，分別是地下鐵雜訊、人聲雜訊、車子雜訊和展覽室雜訊；而在 Set B 所用的雜訊是餐廳雜訊、街道雜訊、機場雜訊和車站雜訊。

而含雜訊語音的產生方式，是先隨機由該種類雜訊中取一段與語音相同長度的雜訊訊號，並乘上一個能夠滿足設定的訊雜比的倍率後，再與乾淨語音相加而成。其中訊雜比的計算是依照分段式訊雜比(segmental SNR)的量度方式，如式(4.1)所示：

$$SNR_{dB} = \frac{1}{M}\sum_{m=0}^{M-1} 10\log_{10}\frac{\sum_{m\in\{\mathrm{I}\}} S_m^2(n)}{\sum_{m\in\{\mathrm{I}\}} N_m^2(n)} \tag{4.1}$$

其中 $\{\mathrm{I}\}$ 代表語音訊號中含有語音的音框所成的集合，M 代表所有語音音框的總數，而 $S_m^2(n)$ 和 $N_m^2(n)$ 分別為第 m 個音框的語音和雜訊的能量。這種計算方式只考慮含有語音的音框去計算訊雜比，因此能夠較準確地合成出符合設定訊雜比的含雜訊語音。

4.3 語音特徵參數的影響

強化型 MMSE 語音強化法的方法是考慮到語音頻譜隨時間變化的情形，去

調整濾波器的頻率響應，而隨時間變化的特性能夠表現在語音特徵中的差量參數中，因為差量參數就是抓取鄰近音框參數間差異的資訊。為了驗證語音頻譜隨時間變化與差量參數的相關性，以及對語音辨識的影響，我們分別使用 13 維的語音特徵參數(12 MFCC+1 維對數能量)、26 維的語音特徵參數(13 維語音特徵參數以及其一階差量參數)以及 39 維的語音特徵參數(13 維語音特徵參數以及其一階與二階差量參數)進行語音辨識，並比較本論文提出的方法與其他語音強化法在辨識準確率的差異。

實驗中除了使用強化型MMSE語音強化法之外，還會與其他 2 種方法比較，分別是固定減去一倍雜訊的語音刪減法以及MMSE估計法，實驗結果的表格中亦會列出不做語音強化就進行辨識的結果。表格中方法的名稱代號分別是NNR、SS、MMSE和EMMSE，其簡述如表 4-3 所列。為了讓各個方法都能在公平的基準上進行比較，所以我們一致採用Martin[6]提出的雜訊估計法。此方法是屬子頻帶的估計方式，它利用前 1.5 秒的最小值作為雜訊估計的基礎值，再乘上一個倍率進行調整。

表 4-3: 實驗採用的語音強化方法代號及描述

方法代號	方法描述
NNR	未經過語音強化處理
SS	固定減去一倍的雜訊估計
MMSE	依照 SNR 值調整濾波器頻率響應
EMMSE	以 MMSE 為基礎，再依據訊號區域的變動特性進行調整

此實驗使用的語音模型為乾淨語音訓練(CST)模型，並以 SetA 中四種雜訊合成的含雜訊語音作為測試語音。表 4-4、表 4-5 及表 4-6 分別為使用 13 維、26 維和 39 維語音參數的辨識結果。表中所列之 H-SNR 與 L-SNR 是所有的雜訊種類，分別針對訊雜比是 20dB~15dB 以及訊雜比是 10dB~-5dB 情況下，取辨識結

果平均後的數據。

表 4-4: CST 模型，測試語音為 SetA，使用 13 維語音參數的結果

辨識準確率	NNR	SS	MMSE	EMMSE
H-SNR	68.88	74.73	74.59	73.24
L-SNR	28.82	43.19	42.70	43.65
Improvement to NNR at L-SNR	0	14.37	13.88	14.83

表 4-5: CST 模型，測試語音為 SetA，使用 26 維語音參數的結果

辨識準確率	NNR	SS	MMSE	EMMSE
H-SNR	87.34	85.51	86.16	86.03
L-SNR	34.37	47.28	48.40	50.74
Improvement to NNR at L-SNR	0	12.91	14.03	16.37

表 4-6: CST 模型，測試語音為 SetA，使用 39 維語音參數的結果

辨識準確率	NNR	SS	MMSE	EMMSE
H-SNR	89.42	85.87	87.21	86.54
L-SNR	30.32	39.76	43.47	46.84
Improvement to NNR at L-SNR	0	9.44	13.15	16.52

由表 4-4 和表 4-5 相比，不論是在高訊雜比或是在低訊雜比的情況下，使用 26 維的語音參數，相較於 13 維的語音參數，都能提升辨識率，因此差量參數能在雜訊環境下，提供抵抗雜訊的能力。但是由表 4-5 和表 4-6 相比，使用 39 維的語音參數相較於使用 26 維的語音參數，在高訊雜比時，能夠有較好的辨識率，但在低訊雜比時，辨識率反而變差了。這是因為在低訊雜比時，雜訊的干擾是很嚴重的，因此會對 39 維中的二階差量參數造成更多失真，導致辨識率反而下降。

在高訊雜比的情況下，不管是經 SS、MMSE 或是 EMMSE 強化後的語音辨

識率，都會比直接拿含雜訊語音進行辨識還差一點。因為在此情況下，雜訊值相對於語音是很小的，所以很容易錯估雜訊值，反倒使得強化後對語音產生更多的失真，而使辨識率更差。

在低訊雜比的情況下，使用 13 維的語音參數，經過 SS、MMSE 和 EMMSE 強化後，相對於 NNR 在辨識率上的提升都是差不多的(約 14.37%)，也就代表了 EMMSE 針對單一個頻率槽隨時間變動的調整方式並不會對前 13 維的語音特徵參數造成太多的影響；但是在使用 26 維的語音參數下，由於差量參數已經具備了部分抵抗雜訊的能力，使得 26 維的 NNR 相比於 13 維的 NNR 已有大幅提升。所以就辨識率的提升而言，MMSE 估計法在使用 26 維語音參數與 13 維語音參數下，辨識率的提升幾乎是差不多的(13.87%與 14.02%)。而 SS 在使用 26 維語音參數下，辨識率的提升甚至比使用 13 維語音參數時少了 1.5%，但是 EMMSE 在使用 26 維語音參數下，卻仍有將近 1.5%的提升。這個情形在使用 39 維的語音參數下更為明顯，SS、MMSE 和 EMMSE 相對於 NNR 的提升分別為 9.44%、13.14%和 16.52%。由此可知，經由 EMMSE 語音強化法強化後的語音訊號，確實能夠加強差量參數受雜訊干擾的能力。這是因為 MMSE 語音強化法利用了語音及雜訊隨時間變動的資訊去調整濾波器的頻率響應，而差量參數是抽取相鄰音框間語音特徵的差異，亦代表了語音訊號隨時間變化的資訊。在使用 26 維的語音特徵參數下，EMMSE 的辨識率與使用 39 維時差不多，在低訊雜比時甚至比 39 維高，因此在接下來的實驗中，我們使用的是 26 維的語音特徵參數，包含 12 維的 MFCC 和 1 維的對數能量，以及這 13 維的一階差量參數。

4.4 乾淨語音訓練模式

在本節中，我們使用乾淨語音訓練(CST)模式訓練語音模型，並將與其他語音強化法比較在不同雜訊以及使用不同語音模型的情況下，對辨識率的影響。表 4-7 與表 4-8 分別為使用含雜訊語音 SetA 與 SetB 作為測試語音的辨識結果。

表 4-7: CST 模型，測試語音為 Set A 的實驗結果

地下鐵	NNR	SS	MMSE	EMMSE
20dB	87.5	84.11	85	84.2
15dB	74.46	78.57	78.75	81.25
10dB	53.21	62.32	63.75	68.75
5dB	29.46	38.39	39.82	45.89
0dB	6.61	13.39	15.18	19.46
-5dB	-7.5	-4.29	-2.68	-0.71
H-SNR	80.98	81.34	81.88	82.73
L-SNR	20.45	27.45	29.02	33.35
人　聲	NNR	SS	MMSE	EMMSE
20dB	93.21	88.75	89.46	87.5
15dB	85.18	80.71	83.75	82.86
10dB	65.36	68.75	70.71	68.75
5dB	43.93	49.82	52.86	50.71
0dB	8.75	25.89	27.86	25.36
-5dB	-30.89	11.07	11.07	10.89
H-SNR	89.2	84.73	86.61	85.18
L-SNR	21.79	38.88	40.63	38.93
車　子	NNR	SS	MMSE	EMMSE
20dB	91.07	88.75	88.39	88.31
15dB	89.46	86.96	86.43	86.25
10dB	84.11	85.36	84.29	86.25
5dB	68.04	79.11	80.18	81.07
0dB	44.29	67.68	68.04	72.14
-5dB	28.04	43.04	44.82	51.61
H-SNR	90.27	87.86	87.41	87.28
L-SNR	56.12	68.8	69.33	72.77
展覽室	NNR	SS	MMSE	EMMSE
20dB	90.89	89.46	90	90.36
15dB	86.96	86.79	87.5	87.5
10dB	72.86	82.5	81.96	82.14
5dB	52.14	69.11	67.86	72.32
0dB	24.11	47.14	48.21	54.11
-5dB	7.5	17.32	20.54	23.04
H-SNR	88.93	88.13	88.75	88.93
L-SNR	39.15	54.02	54.64	57.9

表 4-8: CST 模型，測試語音為 Set B 的實驗結果

餐　廳	NNR	SS	MMSE	EMMSE
20dB	88.21	83.75	83.93	81.25
15dB	76.07	71.96	73.93	71.61
10dB	46.79	57.14	58.75	56.07
5dB	4.46	36.79	37.5	36.43
0dB	-42.68	7.68	7.68	7.14
-5dB	-70.54	-8.21	-9.11	-13.39
H-SNR	82.14	77.86	78.93	76.43
L-SNR	-15.49	23.35	23.71	21.56
街　道	NNR	SS	MMSE	EMMSE
20dB	91.25	88.93	89.29	87.14
15dB	88.57	84.64	85	83.93
10dB	73.04	74.29	75.18	74.64
5dB	55.36	61.96	65.18	70
0dB	31.61	42.86	45.89	47.32
-5dB	19.82	31.61	31.79	33.21
H-SNR	89.91	86.79	87.15	85.54
L-SNR	44.96	52.68	54.51	56.29
機　場	NNR	SS	MMSE	EMMSE
20dB	86.79	85.36	84.82	84.11
15dB	82.32	81.61	82.32	80.36
10dB	68.04	73.21	74.11	73.93
5dB	45.54	60.36	61.25	62.32
0dB	21.25	37.86	41.96	41.25
-5dB	-2.68	20.89	20.54	19.11
H-SNR	84.56	83.49	83.57	82.24
L-SNR	33.04	48.08	49.47	49.15
車　站	NNR	SS	MMSE	EMMSE
20dB	93.39	90	90.36	89.11
15dB	89.46	88.75	89.46	88.75
10dB	78.39	84.11	85.36	83.75
5dB	54.64	78.93	79.64	81.07
0dB	26.07	56.07	58.93	61.43
-5dB	13.39	38.57	39.11	43.39
H-SNR	91.43	89.38	89.91	88.93
L-SNR	43.12	64.42	65.76	67.41

由表 4-7 與表 4-8 的實驗結果看出，在高訊雜比時，幾乎所有強化後的語音都會使辨識率降低一些，包括了本論文提出的 EMMSE 語音強化法，也只能達到與 MMSE 差不多的辨識率。而爲了容易討論各方法在低訊雜比時的效能，我們將表 4-7 與 4-8 中，各語音強化法在低訊雜比時的平均辨識準確率列為下表 4-9。

表 4-9: CST 模型，測試語音為 SetA 和 SetB，在低訊雜比時的平均辨識率

辨識準確率	NNR	SS	MMSE	EMMSE
地下鐵	20.45	27.45	29.02	33.35
人　聲	21.79	38.88	40.63	38.93
車　子	56.12	68.80	69.33	72.77
展覽室	39.15	54.02	54.64	57.90
餐　廳	-15.49	23.35	23.71	21.56
街　道	44.96	52.68	54.51	56.29
機　場	33.04	48.08	49.47	49.15
車　站	43.12	64.42	65.76	67.41
非人聲平均	3.15	31.12	32.17	30.25
類人聲平均	39.47	52.57	53.79	56.15

從表 4-9 可知，所有的語音強化法在面對較不穩定的雜訊時，像是地下鐵、餐廳、人聲以及機場等雜訊，辨識率都會比較低一些。這是由於雜訊快速的變動，導致雜訊的估計會比較不精確，連帶地使語音強化的效能大打折扣。而像是車子、車站和街道這些比較穩定的雜訊，語音強化能夠較有效消除雜訊造成的干擾而使辨識率有明顯地提升。類人聲雜訊也是屬於比較難強化的雜訊，像是人聲和餐廳雜訊，強化後所能提升的辨識率都很有限。

EMMSE 語音強化法相比於他種的語音強化法，在大多數的雜訊環境，即使是不穩定的雜訊下，辨識率都能夠有些許提升，唯獨在處理類人聲雜訊時，此方法的調整方式會使辨識率下降一點點。這是因為這類雜訊主要的能量分佈在與語

音能量的相同頻段，所以我們在判斷該頻率是語音主導或是雜訊主導時，可能會產生誤判，導致強化後的語音產生更多的失真，而使辨識率下降。

4.5 多環境語音訓練模式

接下來我們使用多環境訓練(MCT)模型作為語音模型，並使用含雜訊語音 SetA 與 SetB 作為測試語音。下表 4-10 與 4-11 為實驗的結果。

由表 4-7 相比於表 4-10 的結果看出，表 4-10 的辨識率比 4-7 高出許多。這是因為表 4-7 所使用的語音模型是 CST 模型，其訓練語音是乾淨語音；而表 4-8 所使用的模型是 MCT 模型，其訓練語音是使用與 SetA 相同的雜訊合成的含雜訊語音，因此 MCT 模型能夠包含語音被這四種雜訊干擾後的特性。

相對的，測試語音為 SetB 而言，表 4-8 的辨識率比 4-11 高出許多。因為 SetB 所加入的雜訊與 MCT 模型使用的雜訊是不同種類的，所以當不做語音強化時，辨識率反而會降低。其中只有車站雜訊會有些幅提升，這可能是因為車站雜訊是與車子雜訊較為相似，因此使用含有車子雜訊語音訓練出來的模型，而輸入的是含有車站雜訊的語音時，模型能夠包含部分含車站雜訊語音的失真，故辨識率會些幅提升。

所有的語音強化法在面對地下鐵以及類人聲雜訊時，提升辨識率是比較困難的，因為這兩類雜訊是較難估計的，導致強化後反而會使語音產生額外的失真。但是當雜訊是車子及展覽室雜時，由於這兩類雜訊是屬於較穩定且能量分佈與人聲在不同的頻帶上，因此在進行強化後均能使辨識率提升，而 EMMSE 語音強化法相對於另外兩種方法都能夠有再多一些的提升。

表 4-10: MCT 模型，測試語音為 Set A 的實驗結果

地下鐵	NNR	SS	MMSE	EMMSE
20dB	90.71	86.25	86.25	83.93
15dB	86.43	81.25	82.14	81.96
10dB	80.18	66.79	68.39	68.04
5dB	68.93	48.04	50	50
0dB	53.57	21.07	26.79	25.18
-5dB	33.93	-2.32	1.79	2.32
H-SNR	88.57	83.75	84.19	82.95
L-SNR	59.15	33.39	36.74	36.39
人聲	NNR	SS	MMSE	EMMSE
20dB	94.11	88.57	89.29	87.5
15dB	91.25	86.25	86.07	85.89
10dB	86.43	76.61	78.04	77.68
5dB	78.93	60.54	61.96	62.5
0dB	54.46	33.21	35.18	35.18
-5dB	32.32	11.43	13.57	7.14
H-SNR	92.68	87.41	87.68	86.7
L-SNR	63.03	45.44	47.18	45.63
車子	NNR	SS	MMSE	EMMSE
20dB	90.89	88.39	88.75	85.89
15dB	88.04	87.14	87.32	86.79
10dB	86.61	84.64	84.82	86.07
5dB	77.5	79.46	79.46	81.25
0dB	64.82	72.14	72.86	73.93
-5dB	38.21	51.25	53.57	58.04
H-SNR	89.46	87.76	88.03	86.34
L-SNR	66.78	71.87	72.67	74.82
展覽室	NNR	SS	MMSE	EMMSE
20dB	91.25	88.93	88.21	87.86
15dB	90.71	86.96	88.39	88.21
10dB	80.18	81.43	81.79	82.14
5dB	72.32	73.75	74.11	76.61
0dB	52.32	54.11	54.11	57.5
-5dB	31.61	31.79	33.04	37.86
H-SNR	90.98	87.94	88.3	88.03
L-SNR	59.1	60.27	60.76	63.53

表 4-11: MCT 模型，測試語音為 Set B 的實驗結果

餐廳	NNR	SS	MMSE	EMMSE
20dB	85.89	83.57	85.36	81.25
15dB	74.11	76.96	77.32	73.21
10dB	41.43	58.21	59.82	56.43
5dB	-8.04	34.29	35.71	32.5
0dB	-70.89	1.43	2.14	0.89
-5dB	-114.11	-14.46	-15.71	-18.04
H-SNR	80	80.26	81.34	77.23
L-SNR	-37.9	19.86	20.49	17.95
街道	NNR	SS	MMSE	EMMSE
20dB	91.25	89.64	90.18	85.54
15dB	84.11	83.93	83.04	82.5
10dB	68.04	75	75.36	73.93
5dB	55.71	66.43	67.5	68.93
0dB	22.86	46.43	47.14	50.71
-5dB	12.5	34.29	32.86	32.86
H-SNR	87.68	86.785	86.61	84.02
L-SNR	39.77	55.53	55.71	56.61
機場	NNR	SS	MMSE	EMMSE
20dB	84.64	85.71	85.36	83.57
15dB	74.29	81.61	81.61	80.18
10dB	63.21	75	75.18	75.18
5dB	44.82	67.5	68.75	67.07
0dB	8.75	43.39	43.57	44.14
-5dB	-22.14	19.11	18.57	19.96
H-SNR	79.46	83.66	83.48	81.88
L-SNR	23.66	51.25	51.51	51.59
車站	NNR	SS	MMSE	EMMSE
20dB	92.86	89.82	91.07	89.11
15dB	86.61	87.5	88.04	87.68
10dB	76.43	83.75	84.11	83.93
5dB	69.64	77.86	78.39	79.46
0dB	41.79	57.68	58.39	62.5
-5dB	25.71	41.43	41.25	42.5
H-SNR	89.73	88.66	89.55	88.39
L-SNR	53.39	65.18	65.53	67.1

第五章 應用於分散式語音辨識系統

本論文將提出的語音強化法應用在分散式語音辨識系統上，在系統中扮演著強健式前端(robust front-end)的角色。透過實際環境上的使用，本系統確實能夠有不錯的辨識效能。

5.1 系統架構

隨著科技進步，個人的無線手持式設備已經相當普及了，然而想要透過這些手持式設備輸入指令或是記事時，難免需要手寫輸入，這是相當地不便利的，而改用語音輸入是很好的解決方法。但是大多數的手持式設備都存在著運算效能不足的瓶頸，所以想在手持設備上完成大字彙的語音辨識並不容易達成，因此分散式語音辨識的架構是一個很好的想法[17]。分散式語音辨識系統透過將辨識語音所需的運算量適當地分散在使用者端與伺服器端，並經由無線網路傳送資料，以達到在移動環境中完成語音辨識的工作，而不需要手寫輸入。有鑑於此，本實驗室建立了一個分散式語音辨識系統，並應用了本論文提出的語音強化法，在系統前端對含雜訊語音進行語音強化的處理。

本系統使用的平台，在伺服器端是一般的桌上型電腦，其處理器為 Intel Pentium4 3.0G 並搭配 1GB 的記憶體，作業系統為 Windows XP；在使用者端的手持式設備為 HP iPAQ 5550 個人數位助理(PDA)，其處理器為 Intel PXA255，作業系統是 Packet PC 2003。伺服器與手持設備間是透過無線網路進行傳輸，系統架構圖如圖 5-1 所示：

圖 5-1: 分散式語音辨識系統架構圖

我們將系統分割為三個模組。第一個為語音辨識的前端，其內容包括了對雜訊語音進行強化、抽取語音特徵參數以及無線網路通訊的傳輸，此部分也是本論文實作之重心。第二與第三個則是屬於後端的部分，第二個模組是以HTK為基礎的語音辨識模型以及語者模型調適[25]，第三個模組則是將辨識結果進行構詞的語言模型[26]。

5.2 系統前端實作

分散式語音辨識系統的困難點在於如何平均分散運算量在兩端的設備上，並在有限的網路傳輸頻寬內，達到快速反應辨識結果給使用者的目的。本系統是依照AURORA提出的分散式語音辨識系統的架構標準[17]進行設計的，以PDA作為系統的前端(front-end)，而伺服器是為系統的後端(back-end)。首先利用PDA上有限的硬體資源，抽取語音的特徵參數，並透過無線網路將這些參數傳輸到遠端伺服器上，進行辨識、構詞，再把辨識詞串傳回到終端設備上。系統處理流程如圖5-2 所示：

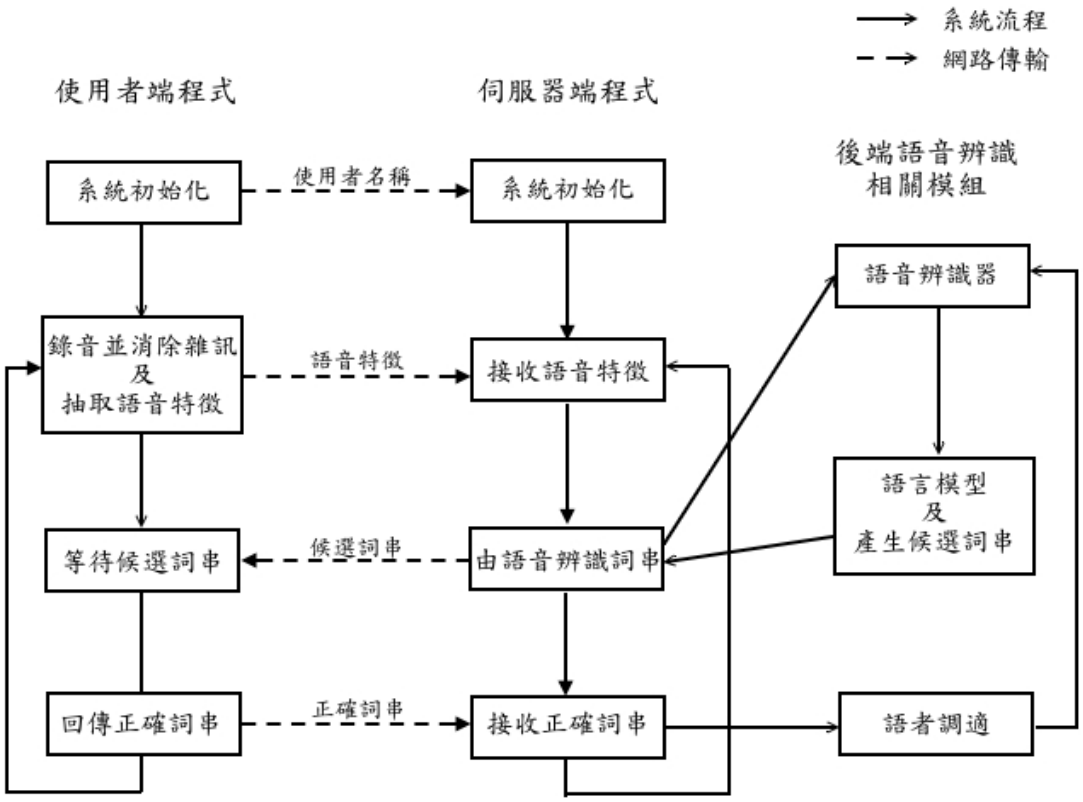

圖 5-2: 系統處理流程圖

圖中的實線箭頭代表系統的處理流程，而虛線箭頭代表網路傳輸的方向，其箭頭上的文字即為傳輸的資訊。

(1) 首先，使用者端與伺服器端的程式先進行系統初始化，建立兩端的網路連線，並由使用者端傳送使用者代號到伺服器端。

(2) 使用者從 PDA 輸入語音，由使用者端程式消除雜訊並抽取語音特徵，再把語音特徵傳送到伺服器端。

(3) 伺服器端接收到語音特徵參數後將之送入語音模型進行辨識，並透過語言模型依辨識結果產生可能的候選詞串，再將此候選詞串傳送到使用者端。

(4) 使用者端接將接收到的可能候選詞串顯示在螢幕上，由使用者點選正確詞串傳送給伺服器端。

(5) 伺服器端利用使用者傳送來的正確詞串進行語者調適的工作。這樣即完

成了一次語音辨識的運作。

爲了在網路頻寬以及運算複雜度之間取得平衡，AURORA 將抽取語音特徵參數的運算分開由使用者端與伺服器端各負責一部分。系統所使用的語音特徵參數為 12 維梅爾刻度倒頻譜係數加上 1 維對數能量及其衍生出來的一階差量參數和二階差量參數，總共 39 維。我們將語音強化模組整合在語音參數的抽取過程中，故本系統的語音參數流程可由圖 2-2 改為下圖 5-3。

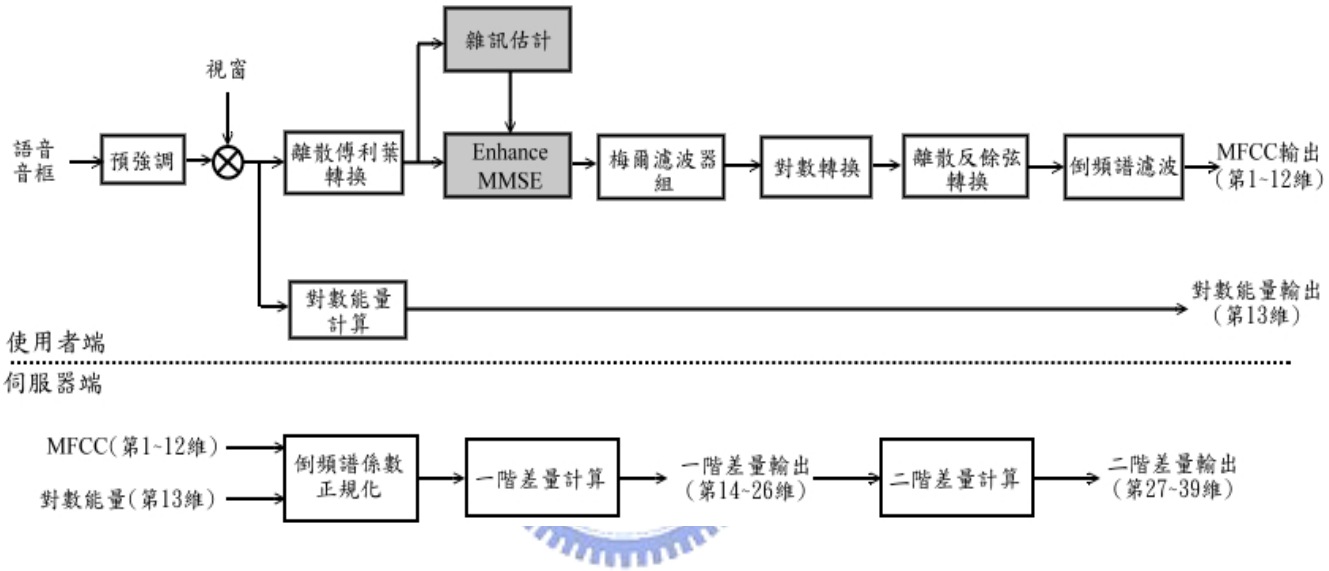

圖 5-3: 整合語音強化於分散式語音辨識系統之語音參數抽取流程

其中先由使用者端對含雜訊語音進行強化，然後抽取 12 維的 MFCC 和 1 維的對數能量，再傳送到伺服器端，由伺服器端產生後續的一階和二階差量參數。

能夠將語音強化模組整合在語音參數的抽取過程中是因為語音強化以及特徵參數抽取都是以音框為基本單位，且都需要轉換到頻域進行處理。因此在音框進行傅利葉轉換之後，進行語音強化，再將強化後的語音頻譜送進梅爾濾波器，完成抽取梅爾刻度倒頻譜參數的步驟。而使用者端就將 12 維的 MFCC 和 1 維的對數能量傳送到伺服器端。

分散式架構的語音辨識系統與一般的語音辨識系統的另一個差別是，在分散式架構中，使用者輸入的語音還需經過手持式設備的錄音通道(channel)，而此通

道亦會對語音產生失真，而造成辨識結果不佳。爲了補償此效應，我們採用了倒頻譜係數正規化(Cepstrum Mean Normalization, CMN)的技術[23]。因此伺服器端將前 13 維的特徵向量經過CMN處理後，先計算其一階差量，產生語音特徵的第 14 到 26 維，再計算二階差量，產生語音特徵的第 27 到 39 維。這樣就完成了語音參數的抽取流程。

由於iPAQ 5550 內的處理器Intel PXA255 是以 400Hz速度在執行定點(fixed-point)運算，但是抽取語音特徵參數需要用到大量的浮點(floating-point)運算，因此會造成系統執行速度上很大的負擔。爲了克服這個速度的瓶頸，在使用者端抽取語音特徵參數的實作上，我們使用定點運算去模擬浮點運算[24]。雖然這樣的做法會損失一些精準度，但是卻能明顯地提升系統速度，且在經過實驗後，這樣的做法仍能保有不錯的辨識率。

5.3 系統效能評估

本系統透過使用者實際操作，並紀錄其辨識結果來評估效能。共用有十位語者進行測試，每個語者使用相同的測試文稿，文稿中有 20 句中文的查詢語句。測試完畢後，以中文字為單位，計算每位語者辨識結果的正確率與準確率，並加以平均。最後結果的正確率為 88.62%，準確率為 85.52%。

第六章 結論與未來展望

6.1 結論

爲了消除相加性雜訊對語音的干擾，導致語音辨識率降低，本論文提出了EMMSE 語音強化法。此方法以最小平方誤差短時頻譜振幅估計法為基礎，並依據含雜訊語音及雜訊估計值的頻譜在某段時間中的變動程度，去調整濾波器的頻率響應，以達到強調語音訊號並壓抑雜訊訊號的目的。經由實驗可知，EMMSE 確實能夠加強語音特徵中差量參數受雜訊干擾的能力。在低訊雜比的情況下，EMMSE 比其他語音強化法能夠消除更多雜訊造成的干擾，而使語音辨識率提升，且隨著使用更多階次的差量參數，提升的效能更為顯著。

本論文並將提出 EMMSE 語音強化法應用在分散式語音辨識系統上，扮演著強健式前端(robust front-end)的角色。透過實際環境的使用，此系統確實能夠有不錯的辨識效能。

6.2 未來展望

本論文驗證了藉由語音與雜訊隨時間變化的特性去消除雜訊的強化方式，確實能夠提升辨識率。EMMSE 語音強化法相比於其他的語音強化法雖然能使語音辨識率有所提升，但是在低訊雜比的情況，因為雜訊干擾相當嚴重，辨識的結果與實際應用仍有一段差距，而還存在著相當大的提升空間。我們希望將來能夠研究出一個會隨雜訊干擾情況自動去設定 EMMSE 語音強化法所需參數的方法(這些參數包括了用來判斷頻率槽的變動是由語音主導或是雜訊主導的門檻值c，以

及用來控制 $f(x)$ 隨 x 變化速度的參數 a)，而不再是依靠實驗經驗去設定，因而能在低訊雜比情況下的辨識率，能有更大的提升。

參考文獻

[1] Boll S., “Suppression of acoustic noise in speech using spectral subtraction”, *IEEE Trans. on ASSP*, vol. 27, 1979.

[2] P. Lockwood and J. Boudy, “Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and projection, for robust recognition in cars,” *Speech Communications*, vol. 11, pp. 215–228, June 1992.

[3] Y. Ephraim and H. L. V. Trees, “A signal subspace approach for speech enhancement,” *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 251–266, July 1995.

[4] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, “Speech enhancement based on the subspace method,” *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 497–507, Sep. 2000.

[5] R. Martin, “Spectral subtraction based on minimum statistics,” in *Proc. Eur. Signal Processing Conf.*, 1994, pp. 1182–1185.

[6] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistic,” *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504-512, July 2001.

[7] R. Martin, “Statistical methods for the enhancement of noisy speech,” *International Workshop on Acoustic Echo and Noise Control*, Kyoto, Japan, Sept. 2003.

[8] R. Prasad, H. Saruwatari, K. Shikano, “Noise estimation using negentropy based voice-activity detector,” *IEEE International Midwest Symposium on Circuites and Systems*, vol. 2, pp. 25-28, July 2004.

[9] F. Beritelli. S. Casale, A. Cavallaero, “A robust voice activity detector for wireless communications using soft computing,” *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 1818-1827, December 1998.

[10] J. Yang, “Frequency domain noise suppression approach in mobile telephone systems,” in *Proceeding of the International Conference Acoustic, Speech, and Signal Processing*, pp. 363-366, 1993

[11] 王小川, “語音訊號處理”, 全華科技圖書, 2004.

[12] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[13] P. J. Wolfe and S. J. Godsill, “Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement,” in *Proc. 11th IEEE Workshop on Statistical Signal Processing*, pp. 496–499, Orchid Country Club, Singapore, August 2001.

[14] N. Virag, “Single channel speech enhancement based on masking properties of the human auditory system,” *IEEE Trans. on speech and audio processing*, vol. 7, no. 2, pp. 126-137, March 1999.

[15] P. Lockwood and J. Boudy, “Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and projection, for robust recognition in cars,” *Speech Communications*, vol. 11, pp. 215–228, June 1992.

[16] X. Huang, A. Acero, H, Hon, “Spoken Language Processing - A Guide to Theory,

Algorithm and System Development," Prentice-Hall, 2001.

[17] "Speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithm," ETSI Standard: ETSI ES 201 108 v1.1.2, 2000, http://www.etsi.org/stq.

[18] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"*, Paris, France, September 2000.

[19] H. C. Wang, F. Seide, C. Y. Tseng, and L. S. Lee, "MAT2000 – Design, collection, and validation on a Mandarin 2000-speaker telephone speech database," in appear in ICSLP2000, Beijing, 2000.

[20] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book(for HTK Version 3.2.1)*, Cambridge University Engineering Department, December, 2002, http://htk.eng.cam.ac.uk.

[21] L. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–285, February, 1989.

[22] TCC-300 speech database, Association for Computational Linguistics and Chinese Language Processing, Institute of Information Science, Academia Sinica, Nangkang, Taipei, ROC. [Online]. Available: http://rocling.iis.sinica.edu.tw/ROCLING/MAT/TCC-300brief.htm.

[23] F. Liu, R. Stern, X. Huang, and A. Acero, "Efficient cepstral normalization for robust speech recognition," in *Proc. ARPA Speech Natural Language Workshop*, pp. 69-75, Princenton, NJ, March 1993.

[24] B. Delaney, N. Jayant, M. Hans, T. Simunic, A. Acquaviva, “A low-power, fixed-point, front-end feature extraction for a distributed speech recognition system,” in *Proceeding of the International Conference on Acoustic, Speech and Signal Processing*, vol. 1, pp. 793-796, 2002.

[25] 謝宗儒, “語者辨識中語者調適方法之研究”, 國立交通大學, 資訊工程系碩士論文, 2005

[26] 呂宜玲, “中文語音辨識中語言模型的強化之研究”, 國立交通大學, 資訊工程系碩士論文, 2005

附錄

本論文中式(3.1)至式(3.7)的推導過程如下：

$$\hat{S} = \tilde{S} + f(x)(\tilde{S} - E[\tilde{S}_L]) \qquad \text{原式(3.1)}$$

其中 $\hat{S}$ 是經 EMMSE 強化後的語音能量，$\tilde{S}$ 是用 MMSE 估計法得到的語音能量，可寫為 $\tilde{S} = YH_{MMSE}^2$ ，H_{MMSE} 即為 MMSE 估計法的濾波器的濾波器。

H_{MMSE}^2 可由溫尼濾波器 H_w 表示，$H_{MMSE}^2 = H_w(\frac{1}{\gamma} + H_w)$，其中 γ 是瞬時事後訊雜比，$\gamma = \frac{Y}{E[\hat{N}]}$ 。

將 $\tilde{S}$ 以 MMSE 估計法改寫，代入式(3.1)可得

$$\Rightarrow \hat{S} = YH_{MMSE}^2 + f(x)\left(YH_{MMSE}^2 - E[Y_L H_{MMSE\,L}^2]\right)$$
$$\Rightarrow \hat{S} = Y\left(H_{MMSE}^2 + f(x)H_{MMSE}^2\right) - f(x)E[Y_L H_{MMSE\,L}^2]$$
$$\Rightarrow \hat{S} = Y \times \underbrace{\frac{1}{Y}\left(Y\left(H_{MMSE}^2 + f(x)H_{MMSE}^2\right) - f(x)E[Y_L H_{MMSE\,L}^2]\right)}_{H_{EMMSE}^2}$$

其中 H_{EMMSE} 即為 EMMSE 語音強化法的濾波器頻率響應。假設 H_{MMSE} 在 L 個連續音框中是穩定的，也就是 $H_{MMSE\,L}^2 = H_{MMSE}^2$ ，因此 $E[Y_L H_{L\,MMSE}]$ 可改寫為下式

$$E[Y_L H^2_{MMSE\,L}] = E[Y_L\left(H_w(\frac{1}{\gamma} + H_w)\right)]$$
$$= E[H_w^2 Y_L] + E[H_w \frac{Y}{\gamma}]$$
$$= H_w^2 E[Y_L] + H_w E[\frac{Y}{\gamma}]$$
$$= H_w^2 E[Y_L] + H_w E[\hat{N}]$$

再將上式$E[Y_L H^2_{MMSE\,L}]$的結果代入H^2_{EMMSE}得下式

$$H^2_{EMMSE} = \frac{1}{Y}\left(Y\left(H^2_{MMSE} + f(x)H^2_{MMSE}\right) - f(x)\left(H_w^2 E[Y_L] + H_w E[\hat{N}]\right)\right)$$
$$= \frac{1}{Y}\left(Y\left(H^2_{MMSE} + f(x)H^2_{MMSE}\right) - f(x)\left(H_w^2 E[Y_L] + H_w E[\hat{N}]\right)\right)$$
$$= H^2_{MMSE} + f(x)H^2_{MMSE} - f(x)\left(H_w^2 \frac{E[Y_L]}{Y} + H_w \frac{E[\hat{N}]}{Y})\right)$$
$$= H^2_{MMSE} + f(x)\left(H^2_{MMSE} - H_w^2 \frac{E[Y_L]}{Y} - \frac{H_w}{\gamma})\right)$$
$$= H_w(\frac{1}{\gamma} + H_w) + f(x)\left(H_w(\frac{1}{\gamma} + H_w) - H_w^2 \frac{E[Y_L]}{Y} - \frac{H_w}{\gamma})\right)$$
$$= \frac{H_w}{\gamma} + H_w^2 + f(x)\left(\frac{H_w}{\gamma} + H_w^2 - H_w^2 \frac{E[Y_L]}{Y} - \frac{H_w}{\gamma})\right)$$
$$= \frac{H_w}{\gamma} + H_w^2 + f(x)H_w^2 - f(x)H_w^2 \frac{E[Y_L]}{Y}$$
$$= H_w\left(\frac{1}{\gamma} + H_w + f(x)H_w - f(x)H_w \frac{E[Y_L]}{Y}\right)$$
$$= H_w\left(\frac{1}{\gamma} + H_w + H_w f(x)(1 - \frac{E[Y_L]}{Y})\right)$$

$$\therefore H_{EMMSE} = \sqrt{H_w(\frac{1}{\gamma} + H_w + H_w f(x)(1 - \frac{\mathrm{E}[Y_L]}{Y}))} \qquad \text{原式(3.7)}$$