國 立 交 通 大 學
電機與控制工程研究所

碩 士 論 文

結合適應性波束形成與後濾波進行語音強化

Combining adaptive beamforming and post-filtering for speech enhancement

研 究 生： 李 明 唐
指導教授： 胡 竹 生 博士

中 華 民 國 九 十 七 年 七 月

結合適應性波束形成與後濾波進行語音強化

研究生：李　明　唐　　指導教授：胡　竹　生　博士

國立交通大學電機與控制工程研究所碩士班

摘　要

本論文提出一套結合適應性空間濾波與後濾波的方法進行語音強化。利用麥克風陣列訊號的空間資訊，以空間濾波方式對聲源方向純化語音，論文中採用成效較佳的 Dahl's 濾波器。為了進一步純化語音，我們使用單聲道語音強化的方法進行後濾波。後濾波主要包含雜訊估測與增益函數兩部份，雜訊估測將分別使用長時間語音活動偵測與最小控制遞迴平均法，搭配以頻譜刪減法與對數頻譜幅值計算出之增益函數，並實際比較在高速公路雜訊與音樂雜訊下的純化效果。

Combining adaptive beamforming and post-filtering for speech enhancement

Student： Ming-Tang Lee Advisor： Prof. Jwu-Sheng Hu

Institute of Electrical and Control Engineering

ABSTRACT

An approach combined adaptive spatial filtering and post-filtering for speech enhancement is proposed in this thesis. Using the spatial information of microphone array signals, we purify the speech in the sound source direction by applying spatial filtering. For the spatial filter, we choose Dahl's beamformer due to it's relatively better performance. To further purify the speech, we use sigle-channel speech enhancement methods for post-filtering. Post-filtering mainly contains noise estimation and gain function parts. We will use long-term voice activity detection (LTVAD) and minima controlled recursive averaging (MCRA) for noise estimation respectively, cooperating with gain functions computed by spectral subtraction (SS) and log-spectral amplitude (LSA) algorithms. And we will compare the purification results under musical noise and noise from freeway.

誌 謝

對於本論文的完成，首先感謝胡竹生老師為我指引方向，碰到困難時也是老師為我指點迷津，讓我的研究得以順利完成。此外，老師也教導我作為一個研究生的態度與觀念，讓我學習如何去認識、研究並解決一個問題，也讓我了解必須對自己的研究負責。在此，向老師致上最誠摯的謝意。

另外，也感謝實驗室中陪伴我的學長姐、同學與學弟們。感謝興哥一直以來的照顧；感謝 papa 讓我看到什麼叫做認真的男人；感謝啟揚讓我知道什麼叫做酷；感謝永融給我的關懷與幫助；感謝畢業的劉大幫助我完成進實驗室第一件任務；感謝宗敏學長與你的相遇，讓我更確定地進入了 Xlab；感謝育德學長提供我這麼好的打工機會；另外還有親切的楷祥、唱歌很 high 的鴻齡、有想法的新好男人 alphar、很會辦聯誼的晉源、有著甜美笑容的鏗元、食量頗大的俊宇、實驗室最強的阿吉、很愛唱歌的 gum、唱歌都會跟去的瓊文、人脈廣大的 Dodo、很會把妹的 Lundy、打球很準的肉鬆、自由搏擊冠軍的 Judo，有你們的陪伴，讓我的實驗室生活更精采豐富。

另外，感謝我的爸爸、媽媽多年來的支持，感謝那些陪我打球的大學同學們，還有所有我愛和愛我的人，謝謝你們。

目　錄

表　列

圖　列

第一章 緒論

1.1 研究動機

環境中聲音的雜訊與干擾源總是無所不在，舉凡冷氣機、電腦風扇、喇叭、空間反射訊號等，不論是在語音辨識或是通訊方面，都有很大的影響。因此，若能設計出一套方法，有效降低雜訊與干擾源影響以達到語音純化效果，將會有很大的應用面。

我們利用麥克風陣列的優勢，透過空間濾波器，可針對聲源方向作純化，並對其他方向干擾源做壓抑。其中，適應性空間濾波器的效果尤其顯著。但由於聲源方向仍夾雜部份干擾源，因此我們需要利用單聲道語音強化的方法做進一步純化。

單聲道語音強化方法對非穩態的噪音壓抑效果較差，這是由於雜訊估測誤差的影響。藉由空間濾波器的幫助，在非聲源方向的干擾源被壓抑後，使得單聲道語音強化效果可以更進一步提升。

1.2 研究目標

本論文目標將分為：

1. 選定及探討適應性空間濾波器之演算法。
2. 探討與比較不同雜訊估測方式與增益函數的結果。
3. 結合空間濾波器與單聲道語音強化方法(後濾波)。

1.3 文獻回顧

麥克風陣列可達到空間濾波的功能，一般而言稱之為 Beamformer[1]，Beamformer 用於麥克風陣列早用於第二次世界大戰[2]，接著慢慢衍生出諸如 Fourier Beamformer[3]、MVDR(Minimum Variance Distortionless Response Beamformer)[4][5]、Robust MVDR[6]、MCMV(Multiply Constrained Minimum Variance Beamformer)[7]、MMSE(Minimum Mean Square Error Beamformer) [8]、MSNR(Maximum SNR)[6]、ML(Maximum Likelihood Beamformer)[6]等。在各種 Beamformer 中最簡單實現的技術為 Fourier Beamformer，它具有較高的 SNR，但是它需要較大的麥克風陣列才可以達到較好的效果，這是因為越多的麥克風可以形成較尖銳的 beam pattern，進而減少其他非聲源角度之干擾源影響。這樣的缺點會造成為了增加效果而必須一直擴大麥克風陣列的體積，因而提出了一種可以自動消除干擾源的 beamformer—MVDR，它除了可以將所量測出之聲源角度作完整聲音之接收，並且還可讓非聲源角度之聲音接收達到最低。此法跟 Fourier Beamformer 有相同之 SNR，然而卻增加了抑制干擾源的效果。然而，如果接收到的訊號是 coherence 或者是作聲源判斷時產生錯誤(pointing error)，MVDR 這方法所形成的效果將大打折扣，甚至會使得原本要接收之聲源變成完全沒有接收。接下來所提出之 Robust MVDR 便是加入 pseudo noise 以減少 pointing error 的影響。另外還有 MCMV 的方法，這個方法需先計算出想要接收的角度以及干擾源的角度，Beamformer 的技術針對此聲源收音並且濾除其他方向之雜訊，則此系統將會變得更為實用，而這方面的系統複雜程度以及運算量相當的龐大，如何去利用 Beamformer 和 DOA 定義出想接收度，或者是不想接收的角度，然後產生一個 beam 於想要接收之角度，並且產生 null 於不想接收之角度，此法便可將不想接收的聲源消除，只是此法還需計算其他之角度，如此增加之計算量將是整體系統的

負擔。

在單聲道語音強化方法中包含雜訊估測與增益函數兩部份。雜訊估測一般常見到的是用 VAD 的方法，而我們所選用的 VAD 是利用長時間語音資訊來判斷[10]。除了 VAD 的方法外，另外有利用估測區域最小值的統計特性來判斷是否有語音成分的方法，如 MCRA(Minima Controlled Recursive Averaging)[11]、MS(Minimum Statistcs)[12]。增益函數最常使用的方法為 SS(Spectral Subtraction)[13]，另外，LSA (Log-Spectral Amplitude)[14]有很顯著的雜訊壓抑效果。

1.4 論文架構

本論文將分別介紹空間濾波器與單聲道語音強化方法，並透過實驗比較，主要內容如下：

第二章：適應性陣列訊號處理。

第三章：後濾波。

第四章：實驗結果與分析

第五章：結論

第二章 適應性陣列訊號處理

2.1 陣列訊號處理

數個感應器排成特定的形狀，接收來自空間中所傳遞的訊號，並經過訊號處理，此技術稱為陣列訊號處理。在陣列訊號處理領域中，依照其目的不同，大致可以將其研究領域分為兩大類，第一種類的研究著重於估測訊號的數量或在空間中的方位，此類研究一般來說稱為到達角估測（Direction of arrival estimation）。而另一種類的研究則是利用訊號的空間關係，希望能夠對不同方向的訊號作出不同的增益，以達到空間濾波的效果，藉以分離空間中不同方向聲源的訊號, 這一類的研究一般稱之為波束形成（Beamforming），也就是一種空間濾波器（Spatial Filter）。

在陣列訊號處理理論中，基於兩個假設

1. 窄頻訊號（Narrow band signal）
2. 遠場平面波（Far field plane wave）

假設一陣列感應器排置如圖 2-1 所示，s(t)為原始訊號，n(t)為雜訊

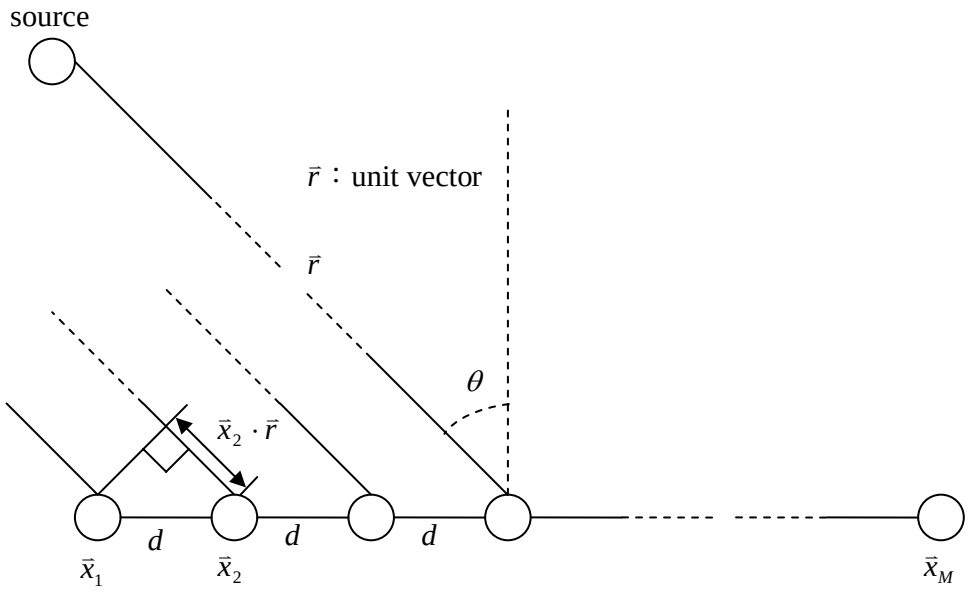

圖 2-1：陣列模型

則 M 個感應器輸出可寫成下列向量形式

$$x(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_M(t) \end{bmatrix} = \begin{bmatrix} s(t)e^{jw_c \frac{\vec{x}_1 \cdot \vec{r}}{c}} \\ \vdots \\ s(t)e^{jw_c \frac{\vec{x}_M \cdot \vec{r}}{c}} \end{bmatrix} + \begin{bmatrix} n_1(t) \\ \vdots \\ n_M(t) \end{bmatrix}$$

$$= \begin{bmatrix} e^{jk_c \vec{x}_1 \cdot \vec{r}} \\ \vdots \\ e^{jk_c \vec{x}_M \cdot \vec{r}} \end{bmatrix} s(t) + \begin{bmatrix} n_1(t) \\ \vdots \\ n_M(t) \end{bmatrix} = a(\vec{r})s(t) + n(t)$$

（2-1）

$k_c = \frac{w_c}{c} = \frac{2\pi}{\lambda_c}$ k_c稱為 wavenumber 而λ_c為波長，c 為波速，$a(\vec{r})$稱為 array manifold vector 包含了訊號傳遞到感應器之間時間關係。

不同的陣列型態會造成不同的空間響應，並會決定陣列的空間解析度，舉例來說，一維的陣列只能解析一維的空間維度，而二維的陣列就可解析二維的空間維度，論文中所實現的陣列型態屬於一維陣列的一部分，因此本章節將介紹屬一維陣列的均勻線性陣列。

均勻線性陣列（Uniform Linear Array），是指一組陣列感應器以線性方式排列，並且感應器之間的距離相等，圖 2-1 其實就是表示一個均勻線性陣列。

若以第一個感應器當作參考點，每個感應器對於訊號源相對角度皆為θ，則第 M 個感應器收到的時間為訊號到達第一個感應器後延遲$\frac{(M-1)\cdot d \cdot \sin\theta}{c}$，因此均勻線性陣列的 Array manifold vector 可寫成如（2-2）式，均勻線性陣列的優點是容易實現且公式容易推導，運算量較其它多維陣列型態低，但缺點為只能對一維空間作解析。

$$a(\theta)=\begin{bmatrix}1\\ e^{jk_c d\sin\theta}\\ \vdots\\ e^{jk_c(M-1)d\sin\theta}\end{bmatrix} \tag{2-2}$$

空間濾波器（Spatial Filter）指的就是將感應器輸出乘上各自加權值的線性組合，因此均勻線性陣列的總輸出可寫成如下形式：

$$p(\theta)=\sum_{i=1}^{M} W_i \cdot e^{jk_c(i-1)d\sin\theta} \tag{2-3}$$

此種線性組合的空間濾波器可稱為波束形成（beamforming），若將（2-3）式中的加權值都設為 1，則 $p(\theta)$ 可化簡成如下所示：

$$\begin{aligned} p(\theta)&=\sum_{i=1}^{M} e^{jk_c(i-1)d\sin\theta}=\frac{e^{jk_cMd\sin\theta}-1}{e^{jk_cd\sin\theta}-1}\\ &=e^{j\frac{k_c(M-1)d}{2}\sin\theta}\frac{\sin\left(\frac{k_cMd}{2}\sin\theta\right)}{\sin\left(\frac{k_cd}{2}\sin\theta\right)} \end{aligned} \tag{2-4}$$

若將 $p(\theta)$ 取 Magnitude 可得其 beampattern，如圖 2-2 所示。

從圖 2-2 可看出，不同角度入射的訊號會有不同的增益，而角度和增益的關係是由陣列的加權值所決定，因此波束形成（beamforming）就可達到空間濾波的效果，而在波束形成理論中，就是用適當的方法去計算出加權值，將訊號作空間濾波，就可得到想要的訊號。

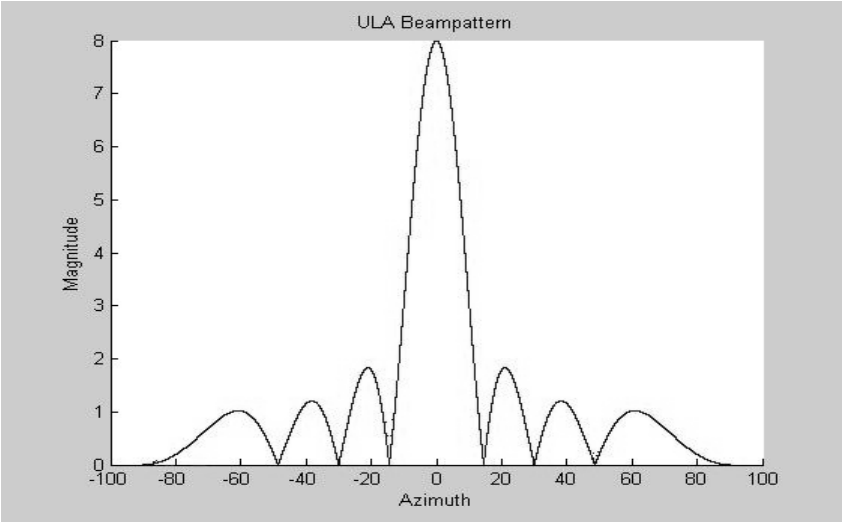

圖 2-2：均勻線性陣列之空間響應（M=8，frequency=100Hz，d=10）

將（2-4）式取絕對值可得

$$\left|p(\theta)\right| = \left| \frac{\sin\left(\frac{k_c M d}{2}\sin\theta\right)}{\sin\left(\frac{k_c d}{2}\sin\theta\right)} \right| \qquad (2\text{-}5)$$

由（2-5）式可看出 $\left|p(\theta)\right|$ 對 $\sin\theta$ 是一週期為 λ_c / d 的週期性的函式，關係圖如圖 2-3 所示。

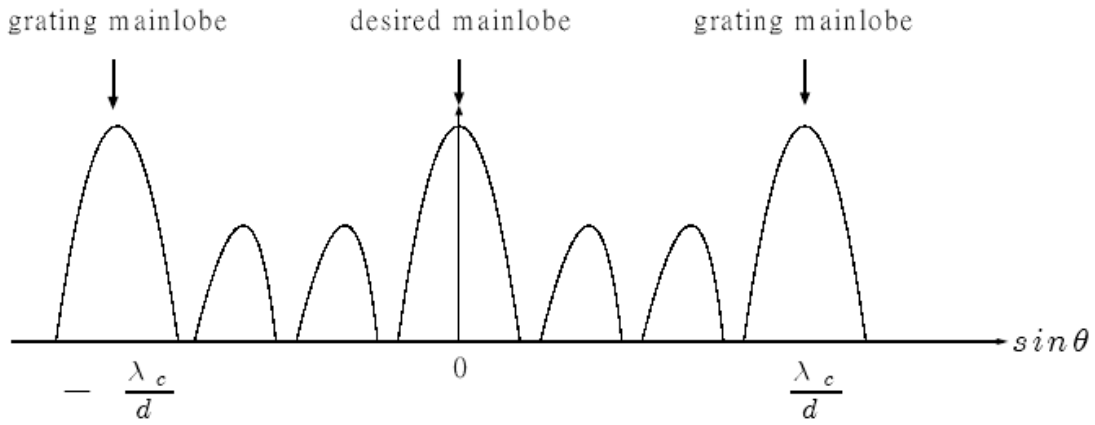

圖 2-3：Grating Lobe 示意圖

在均勻線性陣列中，預期訊號的角度在±90°間，而在這角度之間我們希望 Mainlobe 只會出現一次，如果 Mainlobe 出現兩次以上，則會照成不預期的訊號被接收近來。從圖 2-4 得知，Grating Lobe 發生在 $\sin\theta = \lambda_c / d$ 的時候，因此若讓 $\lambda_c / d > 1$，則可避免在±90°間出現兩個以上的 Mainlobe。而通常我們都會選取 $d = \lambda_c / 2$，以避免 Grating Lobe 的問題。此現象類似於 Nyquist Sampling Theorem，取樣頻率必須是訊號頻率的兩倍以上。

2.2 適應性訊號處理

一般而言，濾波器的係數設計出來後都是固定的，並不會自動的變動。而適應性濾波器指的是能根據輸入信號，用訊號處理的技巧來適應性地調整濾波器係數，讓濾波效果更能適應現在環境，以完成某些特定的需要。

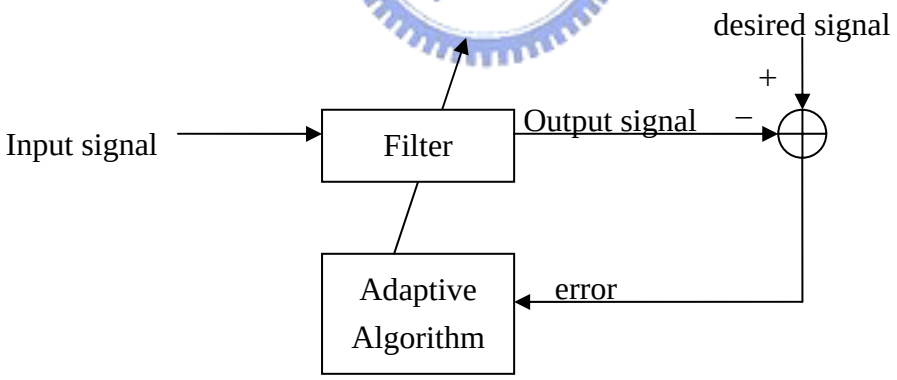

圖 2-4：適應性濾波器處理架構圖

適應性濾波器處理架構圖如圖 2-4 所示，當訊號輸入適應性濾波器處理之後，輸出訊號與希望達成的訊號不同，產生誤差訊號，將誤差訊號代入適應性演算法，即可調整適應性濾波器的係數，如此經由誤差訊號及適應性演算法不斷的調整適應性濾波器的係數，係數會不斷的變動，最後達

到某個穩定的值，此時系統輸出訊號與希望達成的訊號就會非常接近。

關於適應性演算法部份，我們採用計算量較小的最小平均平方法(Least-Mean-Square, LMS)。

LMS 演算法指的是，找出一組權重 W 使得誤差平方項最小[8]。假設希望達成的訊號為 zero-mean，其變異量為 σ_d^2，而輸入訊號 x 為一組 $M \times 1$ 向量，並定義其共異量矩陣和互共異量矩陣

$$E\{d\} = 0, \sigma_d^2 = E\left\{|d|^2\right\}$$

$$R_x = E\{x^* x\} \ , \ R_{dx} = E\{dx^*\}$$

因此目標函數如 2-6 式所示

$$J(w) \equiv \min_w E\{d - xw\}^2 = E(d - xw)(d - xw)^* \qquad (2\text{-}6)$$

(2-6)式的意義就是找出一組 W 使誤差平方項最小，而 W 的找法則需用 Steepest-Descend Method，如(2-7)式，

$$w_i = w_{i-1} + \mu p, \qquad i \geq 0 \qquad (2\text{-}7)$$

其中(2-7)式意義為從 w_{i-1} 出發，並前進 μp 的距離，μ 為一個比重稱為 stepsize。而 p 的選取必須從(2-6)式開始推導，將(2-6)式展開可得

$$J(w) = \sigma_d^2 - R_{dx}^* w - w^* R_{dx} + w^* R_x w \qquad (2\text{-}8)$$

為了找一組 W 使 $J(w)$ 最小，對(2-8)式取 ∇_w 得

$$\nabla_w J(w) = w^* R_x - R_{dx}^* \qquad (2\text{-}9)$$

因此，為了讓 w 往 $J(w)$ 最低處的方向與強度前進，我們取

$$p = -\left[\nabla_w J\left(w_{i-1}\right)\right]^* = R_{dx} - R_x w_{i-1} \qquad (2\text{-}10)$$

故(2-7)式可寫為

$$w_i = w_{i-1} + \mu\left[R_{dx} - R_x w_{i-1}\right] \qquad i \geq 0 \qquad (2\text{-}11)$$

在實做上，R_{dx}和R_x可用離散形式近似於瞬間值：

$$R_{dx} = d(i)x^*(i) \qquad R_x = x^*(i)x(i) \qquad (2\text{-}12)$$

所以(2-11)式可寫為：

$$w(i) = w(i-1) + \mu x^*(i)\left[d(i) - x(i)w(i-1)\right] \qquad i \geq 0 \qquad (2\text{-}13)$$

因此，LMS Algorithm 可整理如下：

Filter out ： $y(i) = x(i)w(i)$ (2-14)

Error function: $e(i) = d(i) - y(i)$ (2-15)

Update weight: $w(i) = w(i-1) + \mu x^*(i)e(i) \qquad i \geq 0$ (2-16)

在 LMS 演算法中，為了確保其收斂，μ 的範圍必須為$0 < \mu < \frac{2}{\lambda_{\max}}$，$\lambda_{\max}$為$R_x$的最大特徵值，若所需濾波器階數愈高，則解$R_x$的特徵值就愈複雜，以實作方面來講，如此大的運算量會造成龐大的負擔，因此為了簡化其運算量，衍生出另一種演算法，Normalize LMS Algorithm：

Filter out ： $y(i) = x(i)w(i)$ (2-17)

Error function: $e(i) = d(i) - y(i)$ (2-18)

Update weight: $w(i) = w(i-1) + \dfrac{\alpha x(i)w(i)}{\gamma + x^*(i)x(i)} \qquad i \geq 0$ (2-19)

與 LMS 演算法比較，Normalize LMS 演算法只有在更新權重的部分不一樣，原有的 μ 被 $\frac{\alpha}{\gamma + x^*(i)x(i)}$ 所取代，其中，0< α <2，γ 為一個微小的數，目的只是確保分母項不為零，如此即可確保 Normalize LMS 演算法收斂，而且如此的運算即不用解 R_x 的特徵值，讓運算量降低許多，但在硬體實現上，除法仍會消耗較大的硬體資源。

2.3 適應性陣列訊號處理

一般在作陣列訊號處理時，會假設兩條件：

1. 窄頻訊號（Narrowband signal）
2. 遠場平面波（Far field plane wave）

當此兩條件成立時，系統數學式子會簡化許多，空間濾波器的設計也較為簡單，但若感應器陣列所收到的訊號並非遠場平面波，則空間濾波器的設計會變的非常複雜，因此為了簡化空間濾波器的設計方法，則將陣列訊號處理結合了適應性訊號處理的觀念。因為適應性訊號處理只須知道希望達到訊號在空間上的特徵，則可利用演算法去自動調整適應性濾波器；若將此觀念用於陣列訊號處理，則只須先用感應器陣列得知希望達到訊號的空間特徵，在利用適應性訊號處理演算法來設計「適應性空間濾波器的係數」，於是將適應性觀念用於空間濾波器中。如此，就算感應陣列所收到的訊號並非遠場平面波，但只要知道訊號在空間的特徵，那麼即可利用適應性空間濾波器來專門接收某方向的訊號，並且不斷地作適應性調適，使誤差訊號愈來愈小。

本章節將介紹用於麥克風陣列的適應性空間濾波器設計方法，稱作 Dahl's Algorithm。依據適應性訊號處理的觀念，必須先得到希望達到訊號

的特性，而Dahl's Algorithm的訊號擷取架構圖如圖 2-5 所示。

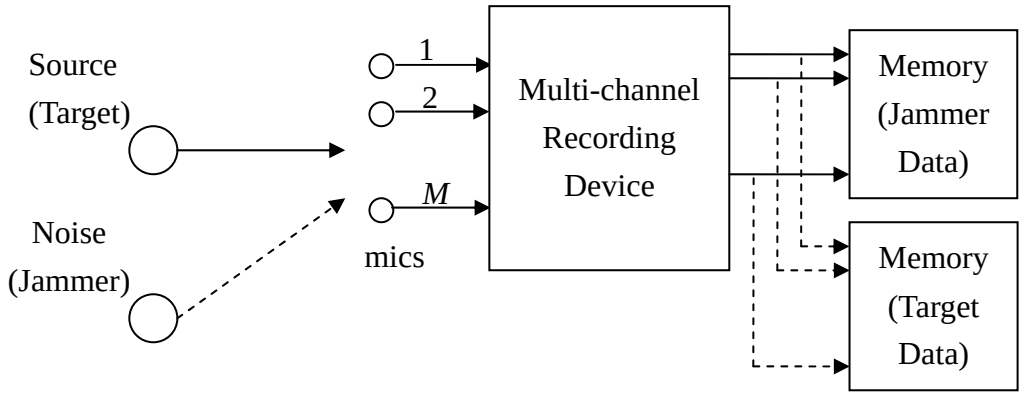

圖 2-5：Dahl's Algorithm 訊號擷取架構圖

Dahl's Algorithm的訊號擷取架構圖分兩部分來操作，首先利用M個麥克風，在安靜的環境下錄製希望達到的訊號，也就是特定方向的語音訊號，再將此訊號儲存至記憶體。第二步驟就是錄製固定干擾源，也就是希望空間濾波器濾掉的訊號，並將此固定干擾源儲存至記憶體。舉例來說，若環境中有人的講話聲和喇叭所撥放的音樂聲，則Dahl's Algorithm的操作方式為先用麥克風陣列在安靜環境下錄製幾秒鐘人講話的聲音，秒數可自己設定，接下來也在安靜環境下錄製幾秒鐘喇叭所撥放的音樂聲，這樣則完成Dahl's Algorithm的預錄部分。

而Dahl's Algorithm架構圖如圖 2-6 示，此架構用虛線分為兩部分，上半部分為將麥克風陣列收到的訊號乘上空間濾波器的係數作為輸出，而下半部分則為空間濾波器係數的更新。更新空間濾波器係數方式為，將麥克風陣列即時錄製到的訊號與希望達到的訊號和固定干擾源作相加，相加的結果當作 LMS Algorithm 的輸入，再利用 LMS Algorithm 去調變空間濾波器係數，係數會不斷變動，最後收斂到某一範圍，如此適應性空間濾波器的輸出訊號會和希望達到的訊號誤差最小，也就是說空間濾波器在希望達到訊號的方向增益最高，而固定干擾源的方向增益會被壓低，

達到濾除干擾源的效果。

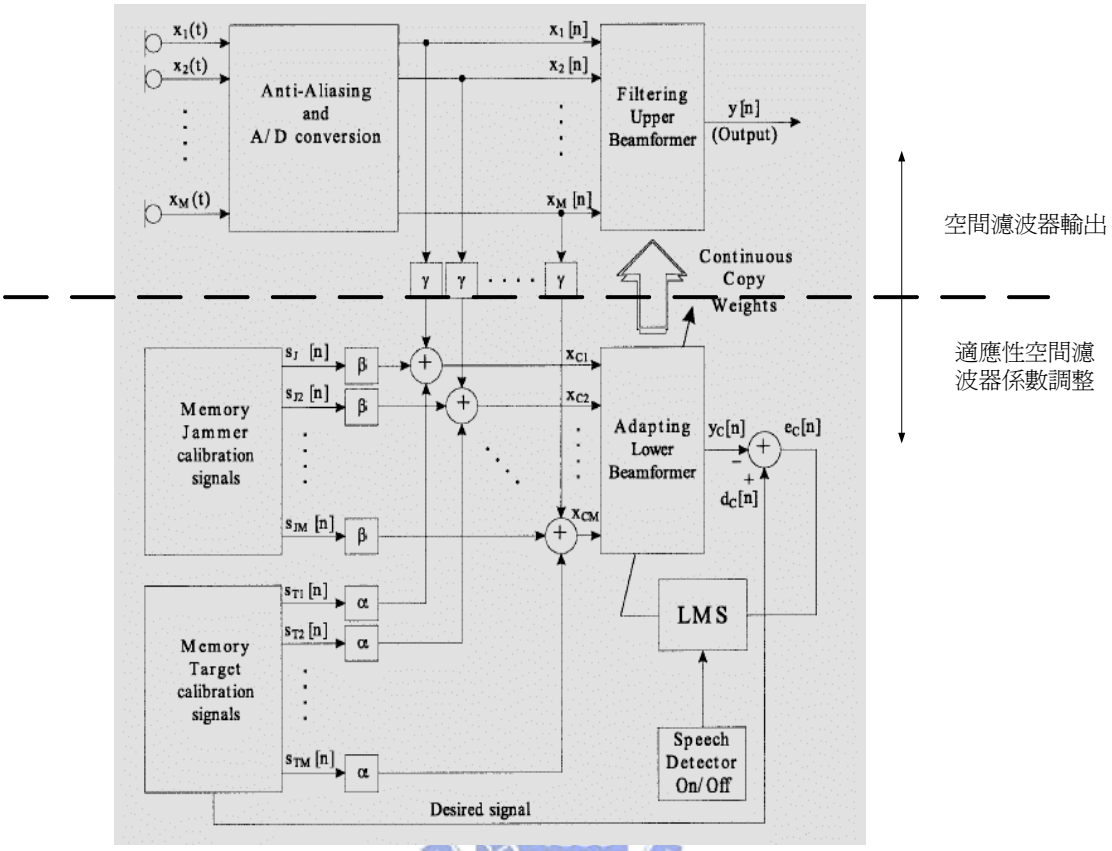

圖 2-6：Dahl's Algorithm 架構圖

在Dahl's Algorithm中，適應性空間濾波器調適和空間濾波器輸出這兩部分不可同時進行，這部份由Speech Detector判斷進行適應性空間濾波器調適或是空間濾波器輸出。當判斷為語音時則進行空間濾波器輸出進行空間濾波器輸出時；反之若判斷為非語音則進行適應性空間濾波器調適。而當干擾源方向改變或有新的干擾源發生，則必須重新啟動適應性空間濾波器係數調整的功能並關閉空間濾波器輸出，調整出適合新干擾源方向的空間濾波器係數。

第三章 後濾波(Post-filtering)

後濾波(Post-filtering)指的就是在空間濾波後做的濾波動作，更進一步純化語音。而後濾波本身是單聲道語音強化的方法，主要分成兩部份：

1. 噪音估測(Noise estimation)
2. 增益函數(Gain function)

透過噪音估測後，再利用估測時的參數產生增益函數用來濾波。在噪音估測方面，最常見的就是用語音活動偵測(Voice activity detection, VAD)的方法，可根據語音資訊如語音能量或過零率等來判斷該音框是否包含語音,而在此將介紹的語音活動偵測演算法是使用長時間語音資訊（long-term speech information）來判別是否有真人語音[10]。除了語音活動偵測之外，尚有一些方法利用統計的概念去估測語音是否存在，如最小控制遞迴平均法(minima controlled recursive averaging, MCRA)[11]、最小統計法(minimum statistics)[12]。在此也將會特別介紹最小控制遞迴平均法(MCRA)。

至於增益函數部份，一般常用的有 Wiener filter，頻譜刪減法(spectral subtraction, SS)或是最大概似法(Maximum likelihood)等。由於 Wiener filter 對於噪音估測的結果非常敏感，較不穩健，而最大概似法的結果較差，在此針對較穩健且效果不錯的頻譜刪減法(SS)做介紹。另外，一種基於最小平均平方誤差(minimum mean-square error, MMSE)在對數頻譜(log-spectra)下的方法將被介紹，也就是所謂的對數頻譜幅值(log-spectral amplitude)。對數頻譜幅值非常適合使用在聲音的處理上，像是語音辨識中的 MFCC(Mel frequency cepstral coefficients)也是在對數頻譜幅值下做運算，而在此我們拿來作語音純化。

3.1 噪音估測

3.1.1 長時間語音活動偵測(LTVAD)

近年來語音活動偵測(voice activity detection, VAD)的技術已廣泛應用在通訊上，最常見的判定真人語音資訊為語音能量和過零率，雜訊及氣音的過零率都很高，語音能量都較低。例如，由歐洲電信標準協會（ETSI）所制定用於 GSM（Global System for Mobile Communications）系統中的 AMR（Adaptive Multi Rate）VAD 判定方法就採用了能量、週期、頻譜失真等三種參數來判定[15][16]。另外由國際電信聯盟（ITU）所制定的 G.729-VAD 採用了全頻帶能量差、低頻帶能量差、頻譜失真和過零率四種參數來判定[17][18]。論文中使用的 VAD 演算法是使用長時間語音的資訊而非傳統瞬間音框（instantaneous frame）資訊，以下將針對長時間語音資訊做下列定義：

- **Long-Term Spectral Envelope (LTSE)**

若 $x(n)$ 為一段包含有雜訊的語音訊號，而 $X(k,l)$ 代表著 $x(n)$ 中第 l 個音框第 k 個頻率的值，則 N 階的 LTSE 定義為：

$$\mathrm{LTSE}_{\mathrm{N}}(k,l)=\max\{X(k,l+j)\}_{j=-N}^{j=+N} \tag{3-1}$$

其 $\mathrm{LTSE}_{\mathrm{N}}(k,l)$ 代表的意義為，從第 l-N 個音框到第 l+N 個音框，這 2N+1 個音框分別對其取頻譜絕對值（Amplitude Spectrum）後，在第 k 個頻率下，這 2N+1 個頻域絕對值音框內的最大值，這樣的好處是不容易忽略某些字頭的子音或是摩擦音。

- **Long-Term Spectral Divergence (LTSD)**

$$\mathrm{LTSD}_{\mathrm{N}}(l)=10\log_{10}\left(\frac{1}{NFFT}\sum_{k=0}^{NFFT-1}\frac{LTSE^{2}(k,l)}{N^{2}(k)}\right) \tag{3-2}$$

其中 NFFT 代表了作 FFT（Fast Fourier Transform）的點數，而 $N(k)$ 代表了雜訊的頻譜絕對值平均，定義如(3-3)式：

$$N_{K}(k)=\frac{1}{2K+1}\sum_{j=-K}^{j=K}X(k,l+j) \tag{3-3}$$

從(3-3)式可看出，$N_{K}(k)$ 代表在第 k 個頻率下，第 l 個音框及前後 K 個音框的頻譜絕對值平均，$X(k,l)$ 和先前定義一樣，代表現階段語音的頻譜絕對值。因此 LTSD 的意義為：現階段長時間語音的頻譜能量佔了雜訊頻譜能量的比例，換句話說判定是否為真人語音是用了現階語音能量的大小來判定，而此能量大小包含了長時間語音資訊，並非只有單一音框資訊。當 LTSD 大於某個臨界值則判定為真人語音，反之則非真人語音，而此臨界值 γ 定義如下：

$$\gamma=\begin{cases}\gamma_{0} & E\le E_{0}\\ \gamma_{0}+\dfrac{\gamma_{1}-\gamma_{0}}{E_{1}-E_{0}}(E-E_{0}) & E_{0}<E<E_{1}\\ \gamma_{1} & E\ge E_{1}\end{cases} \tag{3-4}$$

其中 E_{0} 和 E_{1} 代表了在最乾淨和最吵雜的情況下，雜訊的能量，而 E 是指現階段雜訊的能量。γ_{0} 和 γ_{1} 代表在最乾淨和最吵雜的情況下與 LTSD 比較的臨界值，因此 E_{0},E_{1},γ_{0} 和 γ_{1} 是先設定好的初始值。

圖 3-1 為 LTVAD 演算法流程圖，其中 Hang-over 機制是為了延長字母尾音的機制，而雜訊頻譜更新方式是使用一般遞迴(recursive)方式更新：

$$N(k,l)=\alpha N(k,l-1)+(1-\alpha)N(k) \tag{3-5}$$

表示在第 k 個頻率下，以 α 為權重，根據上一個音框的雜訊頻譜，更新第 l 個音框的雜訊頻譜。

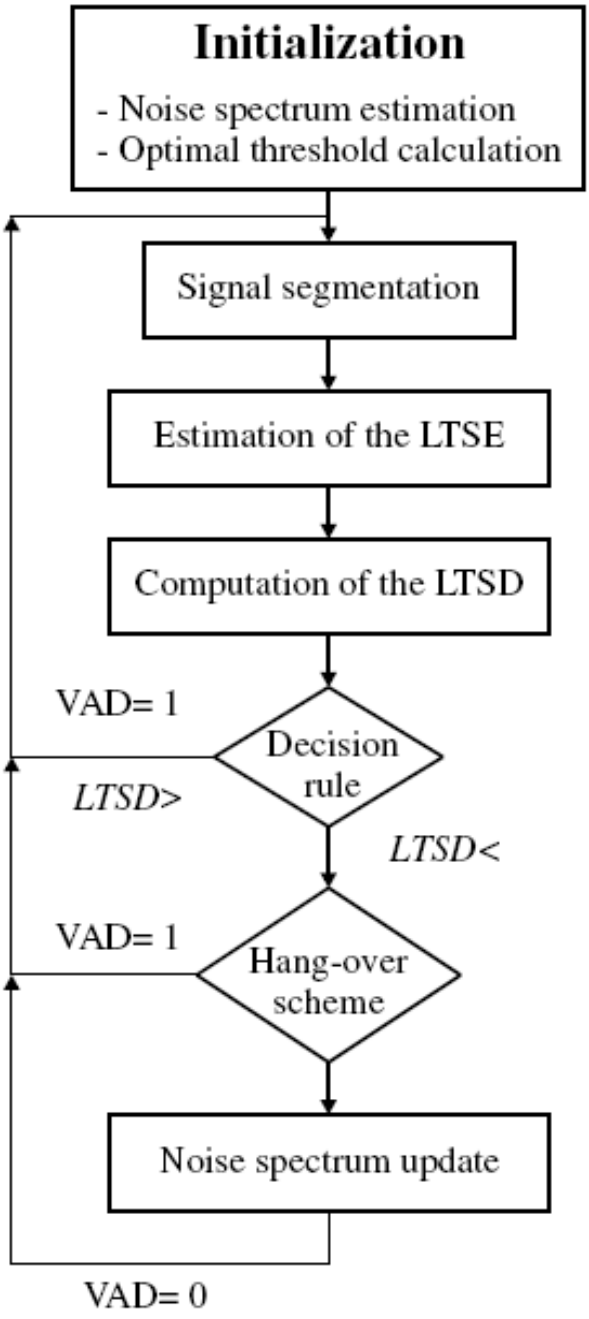

圖 3-1：LTVAD 演算法流程圖

圖 3-2 是 VAD 的模擬結果，此圖尚未加入 hang-over 機制，所以可看出第 5 個發聲的尾音被判定成非語音，加上 hang-over 機制即可解決此問

題。另外，由圖中上半部的 LTSD 與 γ 的關係可知，當 LTSD 大於 γ 時，即判定為語音，然而 γ 變動的劇烈程度取決於初始值的選取，也決定了在不同噪音環境下演算法的穩健性。

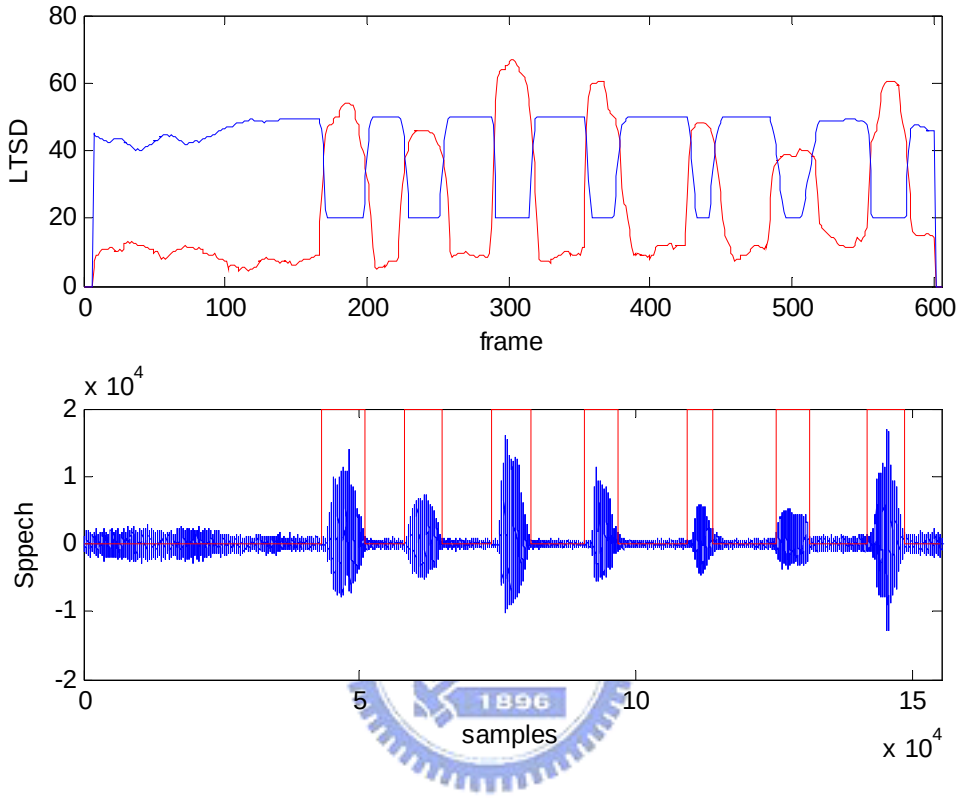

圖 3-2：上半部：LTSD 與 γ 關係圖，下半部：VAD 模擬結果

3.1.2 最小控制遞迴平均法(MCRA)

最小控制遞迴平均法其實是一種實驗性的方法，從頻譜來看，聲音有一塊塊明顯的聲紋，如圖 3-3。若能在一段範圍中，找出該範圍的最小能量，藉由與範圍中最小能量比較，再加上一些統計或是遞迴的方法更新，就可以保留聲紋部份並對雜訊做估測。

最小控制遞迴平均法主要先平緩輸入的能量頻譜，估測一段音框中能量的最小值，接著算出語音可能存在的機率，最後更新雜訊頻譜。以下將一一介紹。

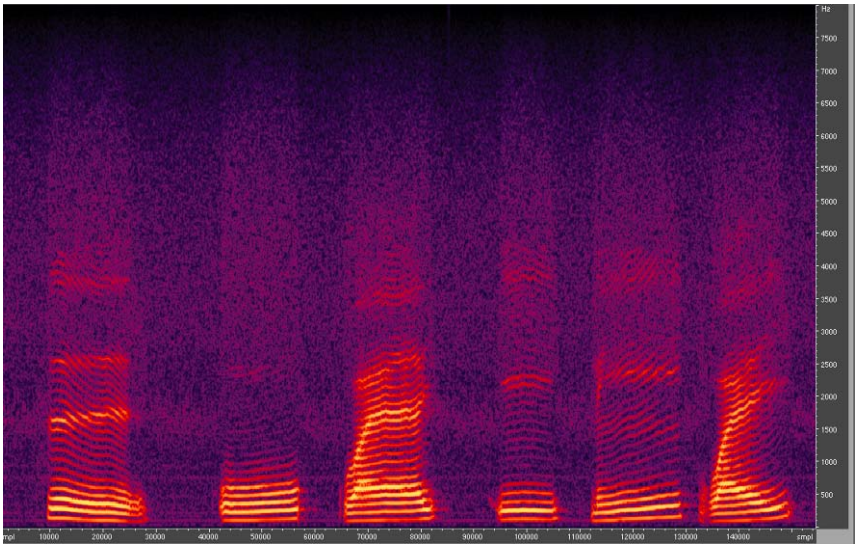

圖 3-3：語音頻譜圖

為了估測雜訊頻譜，我們將音框判定為有語音與沒有語音的情況，即 H_1'和 H_0'，並以常見的遞迴平均法更新估測的雜訊頻譜，如(3-6)式

$$\begin{aligned} H_0'(k,l):\hat{\lambda}_d(k,l+1) &= \alpha_d\hat{\lambda}_d(k,l)+(1-\alpha_d)|X(k,l)|^2 \\ H_1'(k,l):\hat{\lambda}_d(k,l+1) &= \hat{\lambda}_d(k,l) \end{aligned} \tag{3-6}$$

其中 $\hat{\lambda}_d(k,l)$ 為第 k 個頻率，第 l 個音框下的雜訊頻譜。$X(k,l)$ 代表著 $x(n)$ 中第 l 個音框第 k 個頻率的值。α_d ($0<\alpha_d<1$)為平滑參數。

定義語音存在的條件機率為 $p'(k,l)=P(H_1'(k,l)|X(k,l))$，則(3-6)可改寫為

$$\begin{aligned} \hat{\lambda}_d(k,l+1) &= \hat{\lambda}_d(k,l)p'(k,l) \\ &\quad +\left[\alpha_d\hat{\lambda}_d(k,l)+(1-\alpha_d)|X(k,l)|^2\right](1-p'(k,l)) \\ &= \tilde{\alpha}_d(k,l)\hat{\lambda}_d(k,l)+[1-\tilde{\alpha}_d(k,l)]|X(k,l)|^2 \end{aligned} \tag{3-7}$$

其中

$$\tilde{\alpha}_d(k,l)=\alpha_d+(1-\alpha_d)p'(k,l) \tag{3-8}$$

$\tilde{\alpha}_d(k,l)$是一個隨 $p'(k,l)$變動的平滑參數，因此我們只要對 $p'(k,l)$作估測，即可估測雜訊頻譜。

基於對輸入訊號從時域及頻域統計其區域性特性，可以估測語音存在的條件機率 $p'(k,l)$。首先，為了使估測的雜訊較穩健，我們先對輸入的能量頻譜在時域與頻域下做平滑(smoothing)動作。在頻域下，用一個視窗函數 b 決定平滑的範圍，長度為 2w+1，

$$S_{\mathrm{f}}(k,l)=\sum_{i=-w}^{w} b(i)\left|X(k-i,l)\right|^2 \tag{3-9}$$

其中 X 為輸入訊號做 STFT(short-time Fourier transform)後的訊號。

接著，用遞迴平均的方法對時域做平滑動作，

$$S(k,l)=\alpha_s S(k,l-1)+(1-\alpha_s)S_f(k,l) \tag{3-10}$$

其中 α_s ($0<\alpha_s<1$)是參數。

完成平滑動作後，我們要尋找區域性的最小值。一般來說，搜尋區域性最小值的視窗(window)長度適當的範圍通常為 1s ~ 1.5s 間，搜尋時必須包含到非語音的音框才能有準確的估測。然而，使用搜尋比較的方式非常耗運算量，因此在此使用較簡化的搜尋方式[19]。首先，設定區域能量最小值 $S_{\min}(k,0)=S(k,0)$與暫存變數 $S_{temp}(k,0)=S(k,0)$，接著，在每次的音框下與現在的輸入能量作比較，

$$S_{\min}(k,l)=\min\{S_{\min}(k,l-1),S(k,l)\} \tag{3-11}$$

$$S_{temp}(k,l)=\min\{S_{temp}(k,l-1),S(k,l)\} \tag{3-12}$$

當音框數到達我們設定的視窗長度 L 時，我們將區域能量最小值與暫存變數重新做初始化，

$$S_{\min}(k,l)=\min\{S_{temp}(k,l-1),S(k,l)\} \tag{3-13}$$

$$S_{temp}(k,l)=S(k,l) \tag{3-14}$$

接著，重複做(3-11)與(3-12)式直到下一次音框數到達視窗長度。此法找出的區域最小值並非如一般移動性視窗搜尋最小值，而是依視窗長度 L 將輸入訊號切割成一段一段的區域性最小值。暫存變數S_{temp}是確保在視窗間切換時仍能有一個較合理的區域性最小值。

有了區域性最小值與平滑後的輸入能量，我們可以計算比值以判別語音是否存在，

$$S_r(k,l)=S(k,l)/S_{\min}(k,l) \tag{3-15}$$

根據Bayes minimum-cost decision rule，我們可以得到

$$\frac{p(S_r\mid H_1)}{p(S_r\mid H_0)}\underset{H_0'}{\overset{H_1'}{\gtrless}}\frac{c_{10}P(H_0)}{c_{01}P(H_1)} \tag{3-16}$$

其中$P(H_0)$與$P(H_1)$為語音不存在和語音存在的事前機率，c_{10}表示在H_0下判斷成H_1的代價(cost)，c_{01}表示在H_1下判斷成H_0的代價。由於$p(S_r\mid H_1)/p(S_r\mid H_0)$為單調函數(monotonic function)，故(3-16)式可表示成

$$S_r(k,l)\underset{H_0'}{\overset{H_1'}{\gtrless}}\delta \tag{3-17}$$

其中δ為參數，不同種類的雜訊對此參數影響不大。

利用此比值決定此音框是否有語音，

$$I(k,l)=\begin{cases}1 & S_r(k,l)>\delta\\ 0 & \text{otherwise}\end{cases} \tag{3-18}$$

其中$I(k,l)$代表指標函數。$I(k,l)$=1 代表判斷為H_1'，表示語音存在的狀況；

$I(k,l)$=1 代表判斷為 $H_0^{'}$，表示語音不存在的狀況。

透過 $I(k,l)$，我們可以用以下遞迴平均的方式估測語音存在的條件機率

$$\hat{p}^{'}(k,l)=\alpha_p \hat{p}^{'}(k,l-1)+(1-\alpha_p)I(k,l) \tag{3-19}$$

其中 α_p ($0<\alpha_p<1$) 為平滑參數。

將此估測到的語音存在條件機率代回(3-7)，即可求得雜訊頻譜。

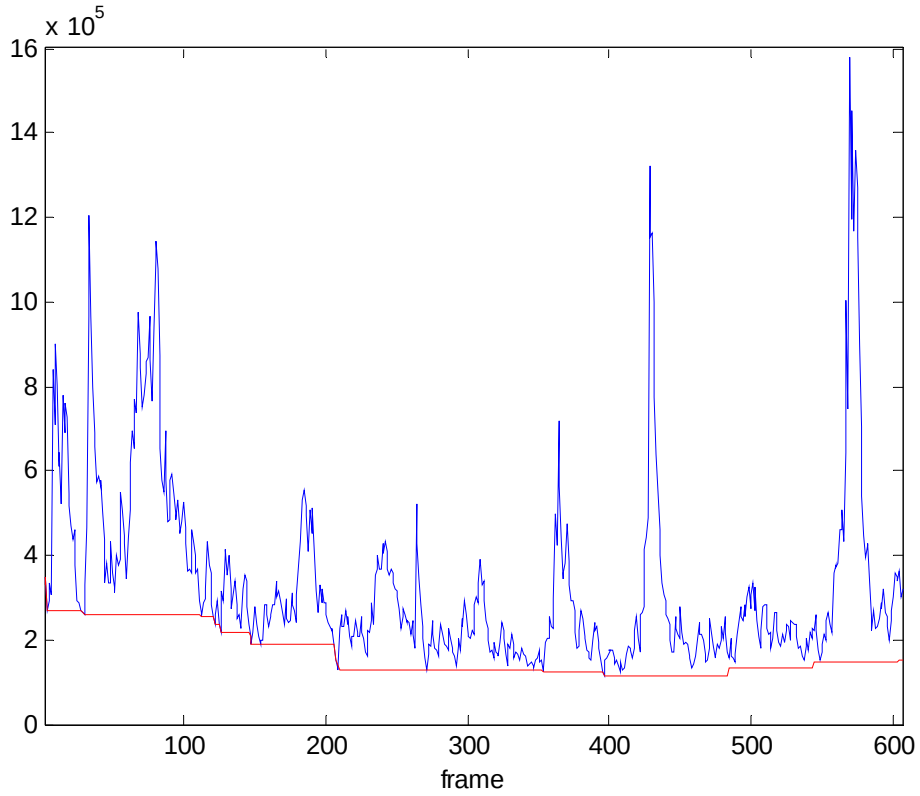

圖 3-4：第 200 個 frequency bin 的 S 與 $S_{\min}$，fft size = 512

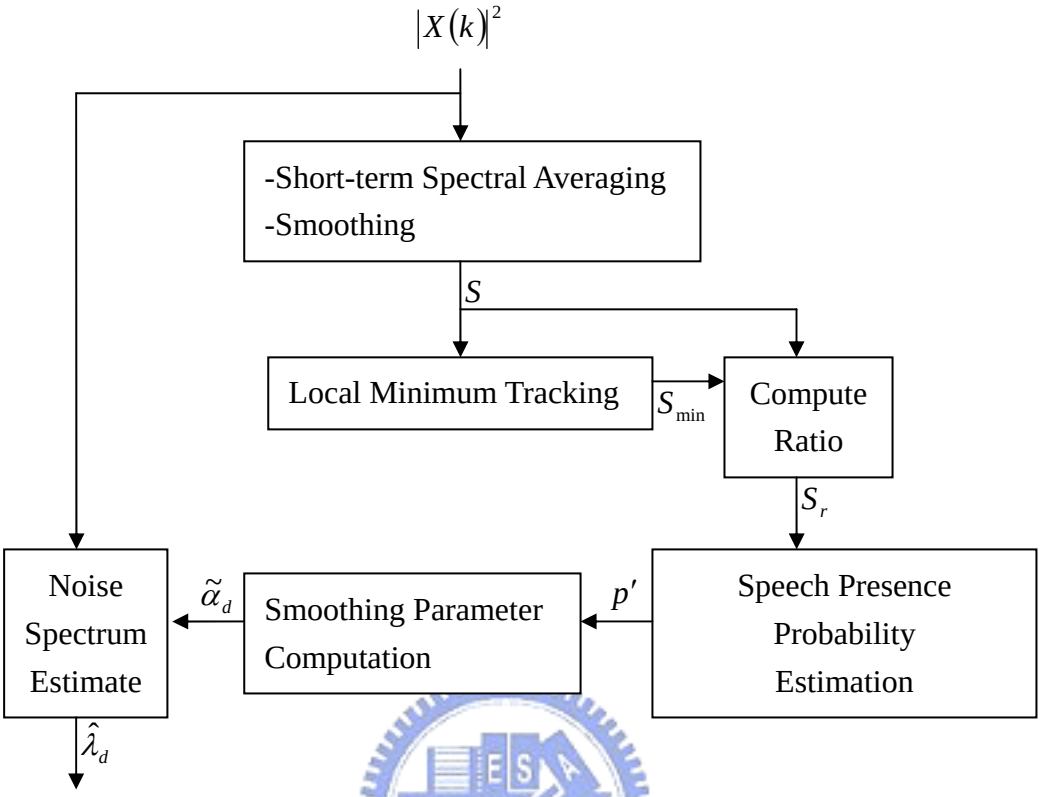

圖 3-5：MCRA 演算法流程圖

3.2 增益函數

3.2.1 頻譜刪減法(SS)

頻譜刪減法是一種在頻域中很常見的訊號處理方式[13]，一般假設輸入訊號的能量頻譜密度(power spectral density, PSD)是由原始訊號的 PSD 與雜訊的 PSD 相加。換句話說，即是假設原始訊號與雜訊為不相關。所以，若能將雜訊的 PSD 成分刪除，即可還原原始訊號的 PSD。

$$X(k) = S(k) + D(k) \tag{3-20}$$

其中 $X(k)$ 為輸入訊號做 DFT 後在頻率 k 下的值，$S(k)$ 與 $D(k)$ 則分別代表原始訊號與雜訊做 DFT 後在頻率 k 下的值。而且

$$X(k)=|X(k)|e^{j\theta(k)} \tag{3-21}$$

由於假設原始訊號與雜訊不相關，故輸入訊號的 PSD 也可表示成

$$S_{XX}(k)=S_{SS}(k)+S_{DD}(k) \tag{3-22}$$

其中 $S_{XX}(k)$、$S_{SS}(k)$、$S_{DD}(k)$ 分別代表輸入訊號、原始訊號與雜訊的 PSD。$S_{DD}(k)$ 可由 3.1 章節中的方法估測。因此我們可以根據輸入訊號與雜訊的 PSD 對原始訊號的 PSD 做估測，

$$\hat{S}_{SS}(k)=\begin{cases} S_{XX}(k)-S_{DD}(k) & \text{if } S_{XX}(k)-S_{DD}(k)>0 \\ 0 & \text{otherwise} \end{cases} \tag{3-23}$$

對(3-23)式開根號並乘上輸入訊號的相位後做 IDFT，即可得到估測的原始訊號，

$$\hat{s}(n)=F^{-1}\left\{\sqrt{\hat{S}_{SS}(k)}\,e^{j\theta(k)}\right\} \tag{3-24}$$

由(3-21)式可將(3-24)式改成，

$$\hat{s}(n)=F^{-1}\left\{\sqrt{\hat{S}_{SS}(k)}\frac{X(k)}{|X(k)|}\right\}=F^{-1}\{G(k)X(k)\} \tag{3-25}$$

$$G(k)=\frac{\sqrt{\hat{S}_{SS}(k)}}{|X(k)|} \tag{3-26}$$

由 Parseval's theorem 可知，

$$\sum_{-\infty}^{\infty}|x(n)|^2=\frac{1}{2\pi}\int_{-\pi}^{\pi}|X(k)|^2dk=\frac{1}{2\pi}\int_{-\pi}^{\pi}S_{XX}(k)dk \tag{3-27}$$

因此可將(3-26)式改寫成

$$G(k)=\begin{cases} \sqrt{1-\dfrac{S_{DD}(k)}{S_{XX}(k)}} & \text{if } S_{XX}(k)-S_{DD}(k)>0 \\ 0 & \text{otherwise} \end{cases} \tag{3-28}$$

其中$G(k)$就是頻譜刪減法中最基本的增益函數格式。這種格式也稱為能量頻譜刪減法(power spectral subtraction)。

利用類似概念，我們也可以對原始訊號 PSD 的指數次方做估測，

$$S_{XX}^{\gamma}(k)=S_{SS}^{\gamma}(k)+S_{DD}^{\gamma}(k) \tag{3-29}$$

$$\hat{S}_{SS}^{\gamma}(k)=\begin{cases} C\left[S_{XX}^{\gamma}(k)-S_{DD}^{\gamma}(k)\right] & \text{if } S_{XX}^{\gamma}(k)-S_{DD}^{\gamma}(k)>0 \\ 0 & \text{otherwise} \end{cases} \tag{3-30}$$

其中 C 為標準化係數(normalize factor)，當γ小於 1 時，頻譜受到刪減的影響隨著γ越小而增加，因此需要 C 去做補償，詳細說明在[13]。

經由推導，我們可以得到增益函數

$$G(k)=\begin{cases} \sqrt[\frac{1}{2\gamma}]{C\left[1-\left(\frac{S_{DD}(k)}{S_{XX}(k)}\right)^{\gamma}\right]} & \text{if } S_{XX}^{\gamma}(k)-S_{DD}^{\gamma}(k)>0 \\ 0 & \text{otherwise} \end{cases} \tag{3-31}$$

在頻譜刪減法中，幾乎都會產生一種稱為音樂性雜訊(musical noise)。這是由於雜訊頻譜估測的誤差，使得頻譜相減時產生了新的不必要的雜訊。

為了減少音樂性雜訊的產生，我們會犧牲一點訊噪比(SNR)來抑制。像是對雜訊的 PSD 前加入係數，來控制刪減的比例；或是使用 spectral floor 的概念，使得雜訊刪減的效果較緩和，如(3-32)、(3-33)式

$$S_{XX}^{\gamma}(k)=S_{SS}^{\gamma}(k)+\alpha S_{DD}^{\gamma}(k) \tag{3-32}$$

$$\hat{S}_{SS}^{\gamma}(k)=\begin{cases} C\left[S_{XX}^{\gamma}(k)-\alpha S_{DD}^{\gamma}(k)\right] & \text{if } S_{XX}^{\gamma}(k)-S_{NN}^{\gamma}(k)>\beta S_{DD}(k) \\ \beta S_{DD}(k) & \text{otherwise} \end{cases} \tag{3-33}$$

其中 α ($0<\alpha<1$)用來控制刪減的比例，$\beta S_{DD}(k)$ ($0<\beta<<1$)則是 spectral floor 的概念。

在第五章的實驗分析中，我們採用的是(3-28)式的能量頻譜刪減法。

3.2.2 對數頻譜幅值(LSA)

對數頻譜幅值估測法是由 Ephraim 與 Malah 所提出[14]，主要概念是用最小平均平方誤差(Minima Mean-Square Error, MMSE)估測法所導出。首先，我們先來看一個 MMSE 的估測器

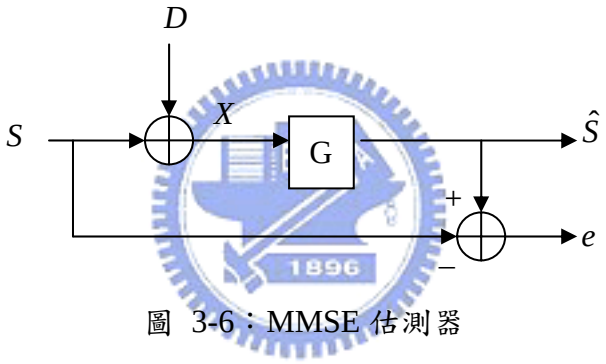

圖 3-6：MMSE 估測器

圖 3-6 中，S 為原始訊號，D 為雜訊，X 為輸入訊號，$\hat{S}$ 為估測結果，e 為估測誤差。為了估測 $\hat{S}$ 使得平均平方誤差 $E\{|e|^2\}$ 有最小值，其中

$$E\{|e|^2\}=E\left[(\hat{S}-S)^T(\hat{S}-S)\right]=E_X\left\{E\left[(\hat{S}-S)^T(\hat{S}-S)|X\right]\right\} \quad (3\text{-}34)$$

所以等同於使 $E_X\left\{E\left[(\hat{S}-S)^T(\hat{S}-S)|X\right]\right\}$ 有最小值，

$$E_X\left\{E\left[(\hat{S}-S)^T(\hat{S}-S)|X\right]\right\}=\hat{S}^T\hat{S}-2\hat{S}E(S|X)+E(S^TS|X)$$
$$=\left[\hat{S}-E(S|X)\right]^T\left[\hat{S}-E(S|X)\right]+E(S^TS|X)-\left[E(S|X)\right]^TE(S|X) \quad (3\text{-}35)$$

由(3-35)式可得到最佳估測解

$$\hat{S}_{opt} = E(S \mid X) \tag{3-36}$$

而在對數頻譜幅值下估測 MMSE，平均平方誤差為

$$E\left\{|e|^2\right\} = E\left[\left(\log \hat{A} - \log A\right)^T \left(\log \hat{A} - \log A\right)\right]$$
$$\text{where} \quad A = |S| \tag{3-37}$$

以相同方式推導可得到最佳估測解

$$\hat{A}_{opt} = \exp\left[E(\log A \mid X)\right] \tag{3-38}$$

在此提出兩個前提，$H_0(k,l)$與$H_1(k,l)$分別代表在第 k 個頻率下，第 l 個音框不含語音與包含語音的情況，

$$H_0(k,l):\ X(k,l) = D(k,l)$$
$$H_1(k,l):\ X(k,l) = S(k,l) + D(k,l) \tag{3-39}$$

假設訊號與雜訊的 STFT 的係數皆為複數的高斯變數[20]，則輸入訊號的條件機率密度函數可表示成

$$p(X(k,l) \mid H_0(k,l)) = \frac{1}{\pi\lambda_d(k,l)} \exp\left\{-\frac{|X(k,l)|^2}{\lambda_d(k,l)}\right\}$$
$$p(X(k,l) \mid H_1(k,l)) = \frac{1}{\pi(\lambda_s(k,l) + \lambda_d(k,l))} \exp\left\{-\frac{|X(k,l)|^2}{\lambda_s(k,l) + \lambda_d(k,l)}\right\} \tag{3-40}$$

由 Bayes rule 可知，

$$p(H_1(k,l) \mid X(k,l))$$
$$= \frac{p(X(k,l) \mid H_1(k,l))p(H_1(k,l))}{p(X(k,l) \mid H_1(k,l))p(H_1(k,l)) + p(X(k,l) \mid H_0(k,l))p(H_0(k,l))} \tag{3-41}$$

在此定義

$$q(k,l)=P(H_0(k,l)),\ \xi(k,l)=\frac{\lambda_s(k,l)}{\lambda_d(k,l)},\ \gamma(k,l)=\frac{|X(k,l)|^2}{\lambda_d(k,l)}$$
$$v(k,l)=\frac{\gamma(k,l)\xi(k,l)}{1+\xi(k,l)} \tag{3-42}$$

則將(3-40)、(3-42)式代入(3-41)式可得

$$p(H_1(k,l)\mid X(k,l))=\left\{1+\frac{q(k,l)}{1-q(k,l)}(1+\xi(k,l))\times\exp(-v(k,l))\right\}^{-1} \tag{3-43}$$

定義語音存在的條件機率 $p(k,l)\equiv p(H_1(k,l)\mid X(k,l))$，基於(3-39)式的假設，我們可將(3-38)式的 LSA 最佳估測解改成

$$\begin{aligned}\hat{A}(k,l)&=\exp\{E[\log A(k,l)\mid X(k,l),H_1(k,l)]p(k,l)\\&\quad+E[\log A(k,l)\mid X(k,l),H_0(k,l)](1-p(k,l))\ \}\\&=(\exp\{E[\log A(k,l)\mid X(k,l),H_1(k,l)]\ \})^{p(k,l)}\\&\quad\times(\exp\{E[\log A(k,l)\mid X(k,l),H_0(k,l)]\ \})^{(1-p(k,l))}\end{aligned} \tag{3-44}$$

因此，我們必須得知 $\exp\{E[\log A(k,l)\mid X(k,l),H_1(k,l)]\}$、$\exp\{E[\log A(k,l)\mid X(k,l),H_0(k,l)]\}$與 $p(k,l)$ 即可求得 LSA 最佳估測解。

- **求語音不存在時** $\exp\{E[\log A(k,l)\mid X(k,l),H_0(k,l)]\}$

當語音不存在時，根據聲音特徵的客觀標準，增益應該要大於一個門檻 $G_{\min}$ [13][21]。

$$\exp\{E[\log A(k,l)\mid X(k,l),H_0(k,l)]\}=G_{\min}|X(k,l)| \tag{3-45}$$

這種概念與頻譜刪減中(3-32)式所提出的 spectral floor 概念大致相同。

- **求語音存在時** $\exp\{E[\log A(k,l)|\, X(k,l), H_1(k,l)]\}$

根據 Ephraim 與 Malah 的推導[14]，最後可得到

$$\exp\{E[\log A(k,l)|\, X(k,l), H_1(k,l)]\} = G_{H_1}|X(k,l)| \tag{3-46}$$

其中

$$G_{H_1}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \exp\left(\frac{1}{2}\int_{v(k,l)}^{\infty} \frac{e^{-t}}{t} dt\right) \tag{3-47}$$

- **求語音存在條件機率函數** $p(k,l)$

從(3-42)、(3-43)式可得知，只要能求出事前訊噪比(a priori SNR)，$\xi(k,l)$，與事後訊噪比(a posteriori SNR)，$\gamma(k,l)$，還有語音不存在的事前機率，$p(H_0(k,l))$，即可求得 $p(k,l)$。

1. 事後訊噪比 $\gamma(k,l)$ 估測

根據(3-42)式的定義，我們可以利用第 3.1 章節中估測的訊號頻譜 $\lambda_d(k,l)$ 與輸入訊號振幅的平方 $|X(k,l)|^2$ 直接計算。

2. 事前訊噪比 $\xi(k,l)$ 估測

根據 Ephraim 與 Malah 提出的估測方法，Israel Cohen[22]將語音存在的不確定性加入考慮改良成

$$\hat{\xi}(k,l) = \alpha G_{H_1}^2(k,l-1)\gamma(k,l-1) + (1-\alpha)\max\{\gamma(k,l)-1, 0\} \tag{3-48}$$

其中 α ($0 < \alpha < 1$)。

3. 語音不存在的事前機率 $q(k,l)$ 估測

首先，我們對估測到的事前訊噪比做平滑動作，

$$\zeta(k,l)=\beta\zeta(k,l-1)+(1-\beta)\hat{\xi}(k,l) \tag{3-49}$$

接著，在頻域上分別對$\zeta(k,l)$作區域性與全域性的平均，求得$\zeta_{local}(k,l)$與$\zeta_{global}(k,l)$，

$$\zeta_{\lambda}(k,l)=\sum_{i=-w_{\lambda}}^{w_{\lambda}}h_{\lambda}(i)\zeta(k-i,l) \tag{3-50}$$

其中λ分別代表"local"與"global"。

接著，我們定義兩個變數$P_{local}(k,l)$與$P_{global}(k,l)$分別代表$\zeta_{local}(k,l)$、$\zeta_{global}(k,l)$與語音分布的相似程度，

$$P_{\lambda}(k,l)=\begin{cases}0 & \text{if } \zeta_{\lambda}(k,l)\le\zeta_{\min}\\ 1 & \text{if } \zeta_{\lambda}(k,l)\ge\zeta_{\max}\\ \dfrac{\log(\zeta_{\lambda}(k,l)/\zeta_{\min})}{\log(\zeta_{\max}/\zeta_{\min})} & otherwise\end{cases} \tag{3-51}$$

其中λ分別代表"local"與"global"，$\zeta_{\max}$與$\zeta_{\min}$是經驗調整出的常數。

若要進一步在完全沒有語音的音框中壓抑雜訊，我們可以用一個與音框有關的參數去比較與語音分布的相似程度，

$$\zeta_{frame}(l)=\underset{1\le k\le M/2+1}{mean}\{\zeta(k,l)\} \tag{3-52}$$

$$P_{frame}(k,l)=\begin{cases}0 & \text{if } \zeta_{frame}(l)\le\zeta_{peak}(l)\zeta_{p\min}\\ 1 & \text{if } \zeta_{frame}(l)\ge\zeta_{peak}(l)\zeta_{p\max}\\ \dfrac{\log(\zeta_{frame}(l)/\zeta_{peak}(l)/\zeta_{p\min})}{\log(\zeta_{p\max}/\zeta_{p\min})} & otherwise\end{cases} \tag{3-53}$$

其中$\zeta_{frame}(l)$表示$\zeta(k,l)$在一頻帶中的平均，$\zeta_{peak}(l)$是一個對$\zeta_{frame}(l)$作限制的值，$\zeta_{p\max}$與$\zeta_{p\min}$是經驗調整出的常數。

有了 (3-51)、(3-53)式的估測，我們可以得到語音不存在的事前機率，

$$\hat{q}(k,l)=1-P_{local}(k,l)P_{global}(k,l)P_{frame}(l) \tag{3-54}$$

為了降低語音失真的可能性，我們應將$\hat{q}(k,l)$限制在一個門檻$q_{\max}$下

($q_{\max}$<1)。

將(3-45)、(3-47)代入(3-44)，我們可得到最後的增益函數

$$\begin{aligned} G(k,l) &= \hat{A}(k,l) \\ &= \left(G_{H_1}(k,l)\right)^{p(k,l)} G_{\min}^{\ (1-p(k,l))} \end{aligned} \tag{3-55}$$

其中，$G_{H_1}(k,l)$的積分計算尤其重要(3-47 式)，由於估測事前訊噪比$\hat{\xi}(k,l)$會用到$G_{H_1}(k,l)$的回授，所以若是$G_{H_1}(k,l)$積分沒有收斂就會造成整體運算錯誤。至於實際運算$G_{H_1}(k,l)$並非直接積分至無限大，在[14]中可將此式轉成離散形式以便計算。

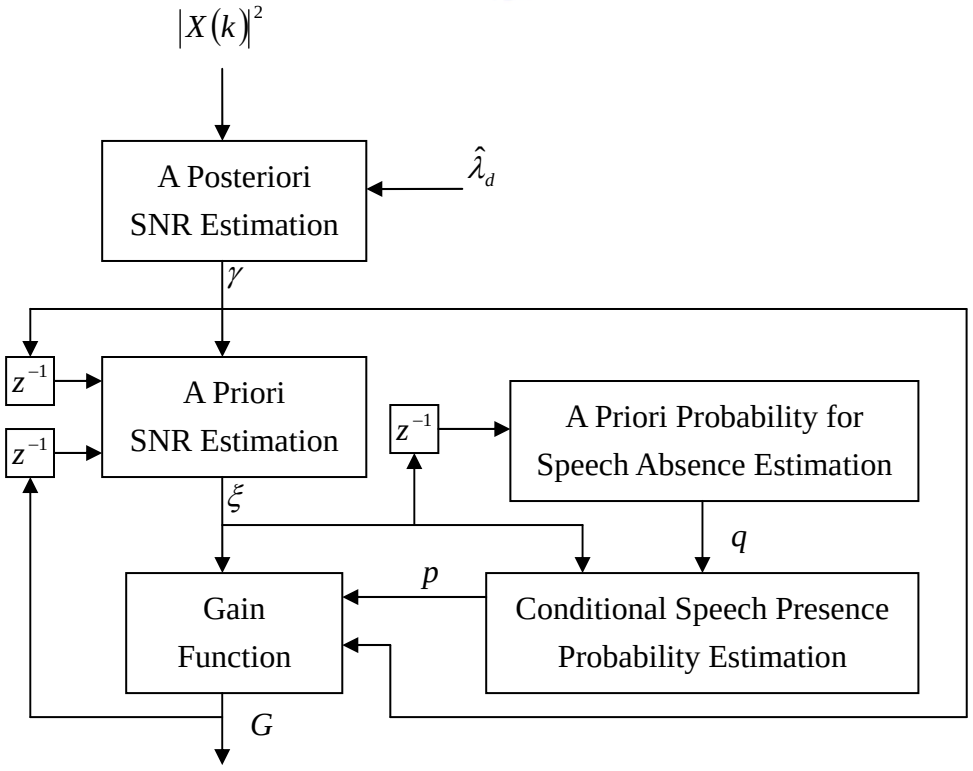

圖 3-7：LSA 演算法流程圖

第四章 實驗結果與分析

本章節將介紹將麥克風陣列平台於不同噪音環境情況下測試的結果，分別探討適應性空間濾波器、獨立成份分析法及本論文所提出之方法並做比較，而實驗環境則在室內環境。

實驗測試將針對 2 顆數位麥克風陣列輸入，模擬一般人在使用手機裝置時的情形。圖 4-1 為數位麥克風陣列輸入裝置照片，實驗中只取上下排的中間麥克風訊號做處理。圖 4-2 為實驗環境示意圖，在室內環境下以喇叭播放噪音源，測試人員手持手機型數位麥克風陣列裝置，模擬一般使用手機之情形。

麥克風陣列裝置為 10cm × 5cm，上下各 3 個，如圖 4-1 所示。以下 4.1 與 4.2 節將各取上下排中間的麥克風資料作處理，即兩通道的麥克風陣列資料，上下麥克風距離約 9.5cm。

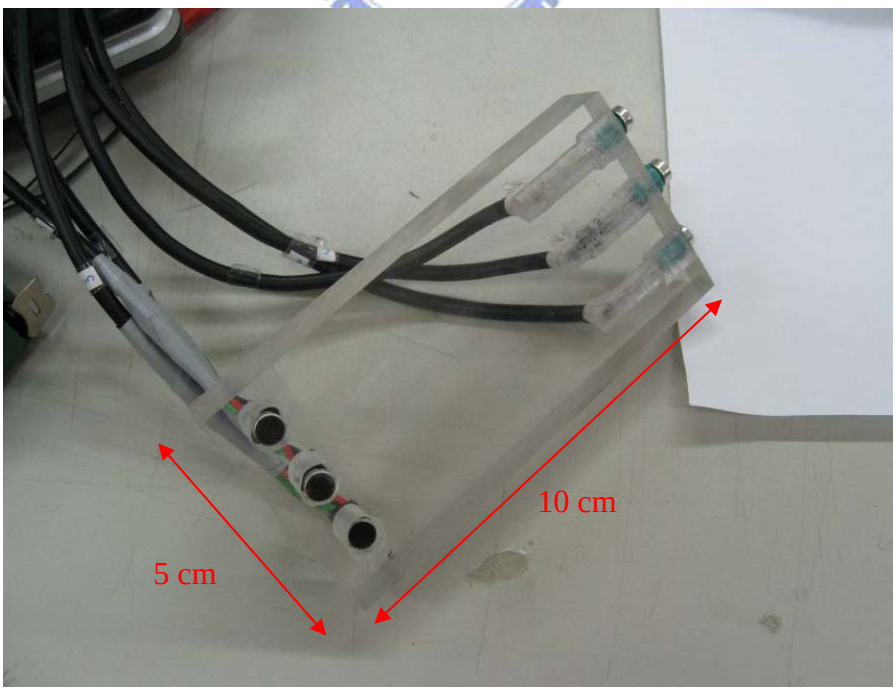

圖 4-1：數位麥克風陣列裝置(上下各有 3 顆數位麥克風)

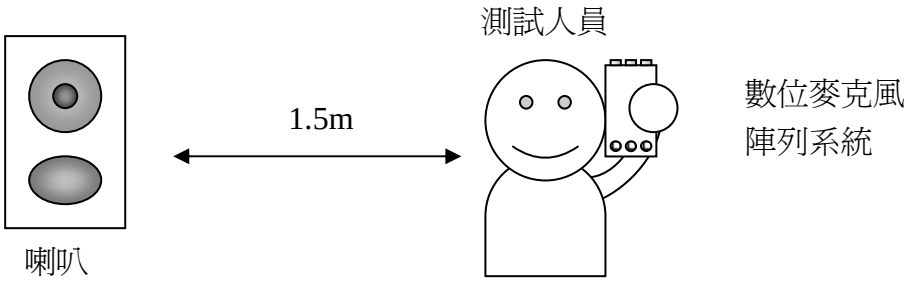

圖 4-2：實驗環境示意圖

實驗中 SNR 的計算方式如下：

$$10\log\left(\frac{\sum_{i=M}^{N} x^2(i)}{N-M+1}\right) \qquad (4\text{-}1)$$

假設雜訊為第 M_1 到第 N_1 筆，而語音加雜訊為第 M_2 到第 N_2 筆，其 SNR 為

$$10\log\left(\frac{\sum_{i=M_2}^{N_2} x^2(i)}{N_2-M_2+1}\right)-10\log\left(\frac{\sum_{i=M_1}^{N_1} x^2(i)}{N_1-M_1+1}\right) \quad \text{dB} \qquad (4\text{-}2)$$

4.1 適應性空間濾波器測試結果

本章節針對 2 通道數位麥克風訊號輸入資料作測試，利用頻域下的 Dahl's algorithm，取快速傅立葉轉換的長度為 512。

測試一：測試雜訊播放高速公路上錄製的聲音，且有預錄一段雜訊作為訓練資料使用。

圖 4-3 為空間濾波器處理前較靠近聲源的麥克風錄製的訊號，SNR = 3.85 dB。經由空間濾波器處理後，如圖 4-4，SNR = 14.91 dB，增加了 11.06 dB。

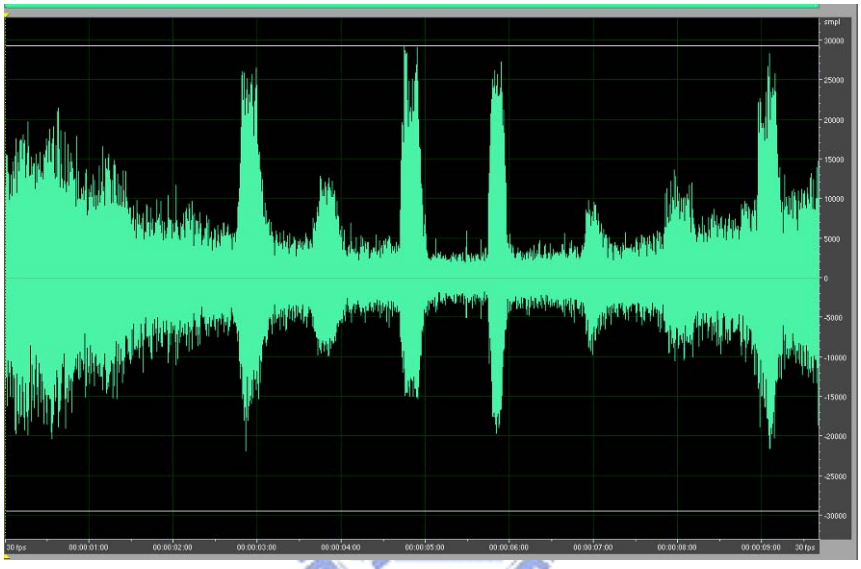

圖 4-3：空間濾波器處理前，高速公路雜訊

圖 4-4：空間濾波器處理後，高速公路雜訊

測試二：測試雜訊播放音樂(孫燕姿-奔)，且有預錄一段雜訊作為訓練資料使用。

圖 4-5 為空間濾波器處理前較靠近聲源的麥克風錄製的訊號，SNR = 8.61 dB。經由空間濾波器處理後，如圖 4-6，SNR = 13.28 dB，增加了 4.67dB。

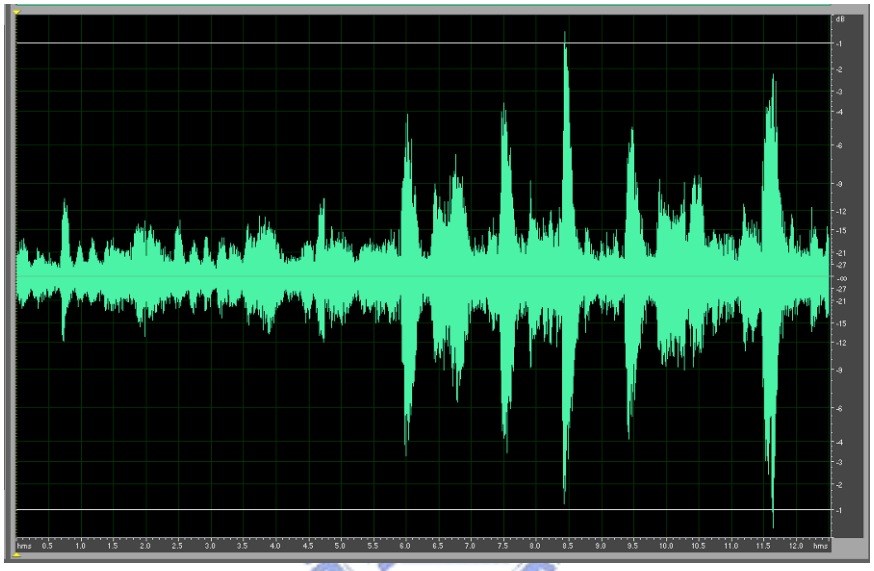

圖 4-5：空間濾波器處理前，音樂雜訊

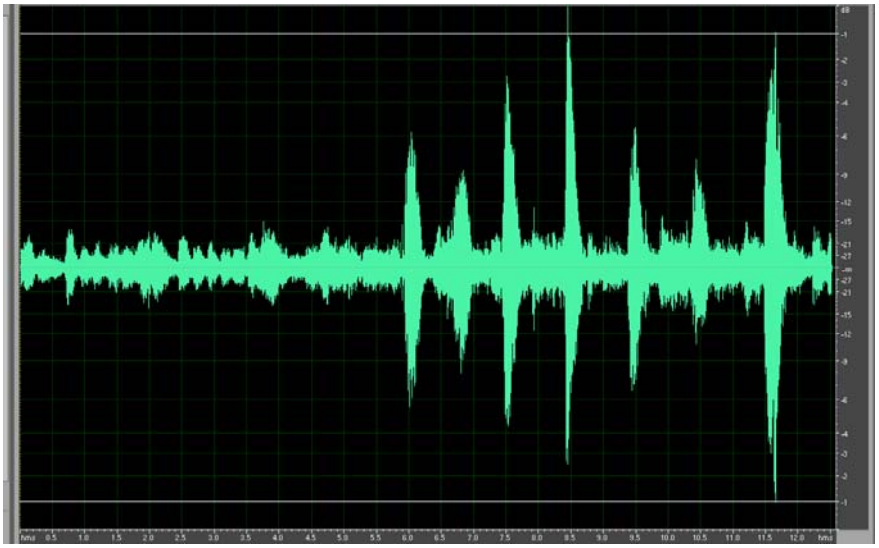

圖 4-6：空間濾波器處理後，音樂雜訊

適應性空間濾波器測試總結：

在測試一高速公路雜訊的效果最好，高速公路所錄製的雜訊除了能量變化會受到車輛來往影響之外，較為接近白色雜訊。在測試二中可明顯發現有兩處背景音樂的人聲明顯被壓抑下去，可看出適應性濾波器在處理同為人聲訊號上，利用空間資訊處理的優勢。

目前的應用上僅以兩通道訊號做處理，增加通道數可再提升效能，但由以上測試可知在聲音反射嚴重的室內，只針對主要聲源方向作純化能提升的 SNR 值有限。

4.2 結合空間濾波與 post-filtering 測試結果

為了進一步強化語音，在此結合了單聲道強化語音的方法做空間濾波的後處理。以下將分別用長時間語音活動偵測(long-term voice activity detection, LTVAD)與最小控制之遞迴平均法(minima controlled recursive averaging, MCRA)做噪音估測，並主要針對頻譜刪減(spectral subtraction, SS)與對數頻譜幅值(log-spectral amplitude, LSA)的結果做比較。以下分別用高速公路雜訊與音樂雜訊作測試。

測試一：高速公路雜訊

測試雜訊為高速公路錄製雜訊，經由空間濾波器處理後(如圖 4-4)，分別以不同雜訊估測方法搭配不同增益函數(gain function)測試語音純化效果。

1. LTVAD+SS

圖 4-7 為空間濾波器後再經由後處理的結果，其中雜訊以長時間語音活動偵測做雜訊估測，並以頻譜刪減做增益函數。訊號的 SNR = 21.57 dB，與空間濾波的結果(圖 4-4)做比較，增加了 6.66 dB。

圖 4-8 為處理後訊號的頻譜分布圖，可看出背景部份白訊號的成分已被消去，然而大部分的聲紋都被完整保留下來。

圖 4-7：高速公路雜訊空間濾波後再經過後處理，LTAD+SS

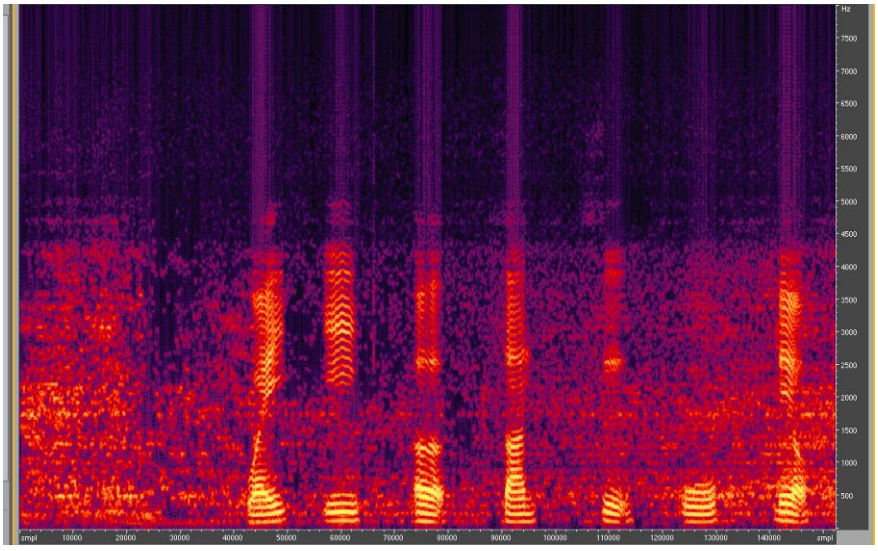

圖 4-8：頻譜分布圖。高速公路雜訊，LTVAD+SS

2. LTVAD+LSA

圖 4-9 為空間濾波器後再經由後處理的結果，其中雜訊以長時間語音活動偵測做雜訊估測，並以對數頻譜幅值做增益函數。訊號的 SNR = 33.13 dB，與空間濾波的結果(圖 4-4)做比較，增加了 18.22 dB。

圖 4-10 為處理後訊號的頻譜分布圖，可看到非語音的部份幾乎都被消去，但這也包含了一些聲音中氣音的部份，然而大部分的聲紋仍被保留下來。

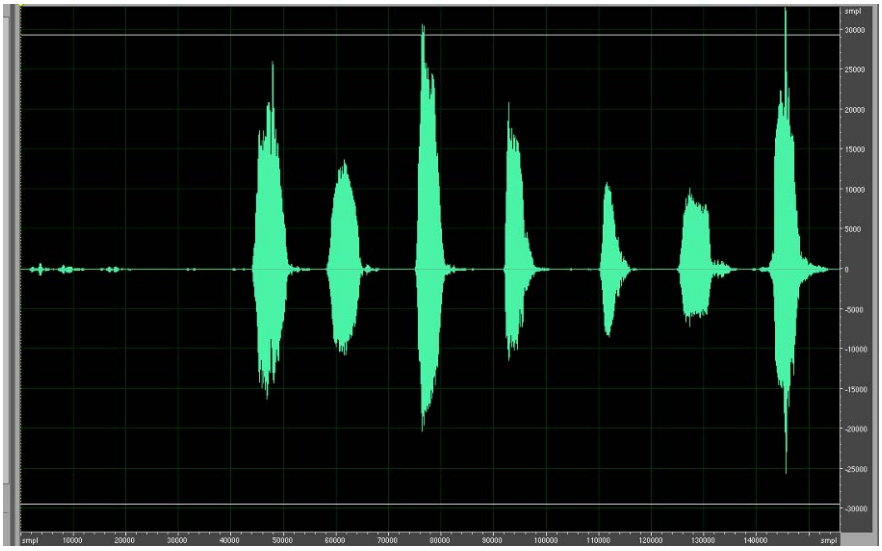

圖 4-9：高速公路雜訊空間濾波後再經過後處理，LTAD+LSA

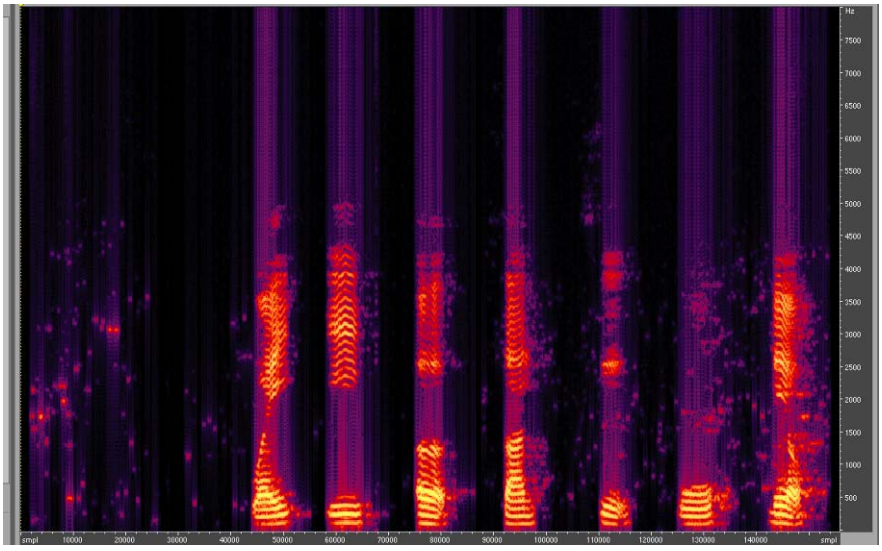

圖 4-10：頻譜分布圖。高速公路雜訊，LTVAD+LSA

3. MCRA+SS

圖 4-11 為空間濾波器後再經由後處理的結果，其中雜訊以最小控制之遞迴平均法做雜訊估測，並以頻譜刪減做增益函數。訊號的 SNR = 19.74 dB，與空間濾波的結果(圖 4-4)做比較，增加了 4.83 dB。

圖 4-12 為處理後訊號的頻譜分布圖，與 LTVAD+SS 的結果做比較(圖 4-8)，可發現在部分聲紋能量較強的頻帶之後沒有語音成分的音框，有明顯的壓抑，而其他部分則較不顯著。

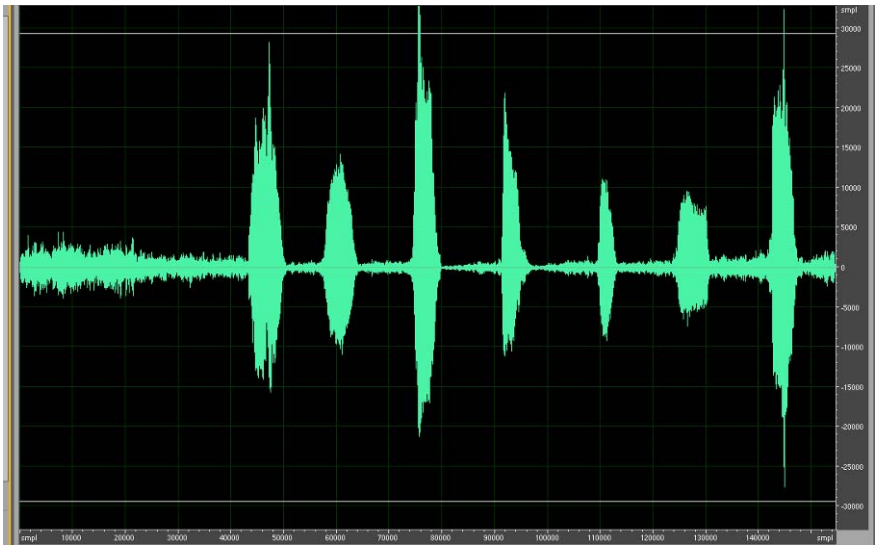

圖 4-11：高速公路雜訊空間濾波後再經過後處理，MCRA+SS

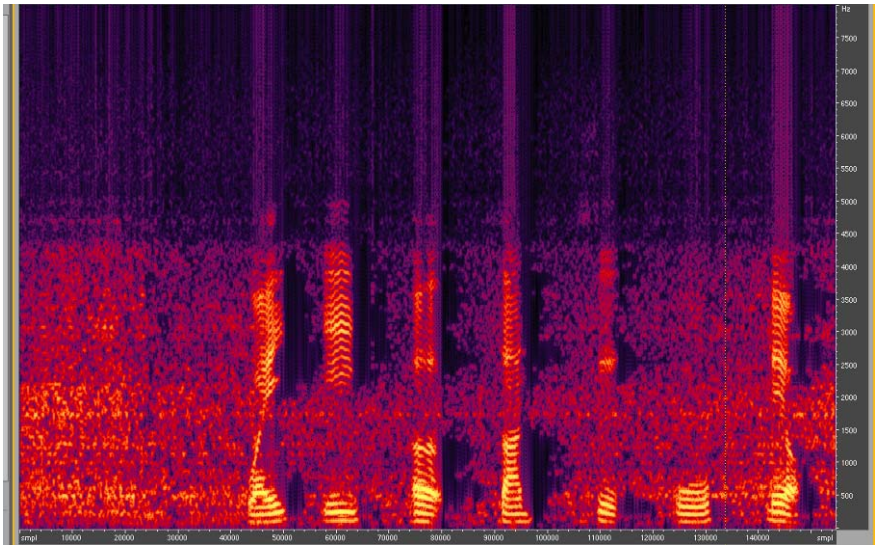

圖 4-12：頻譜分布圖。高速公路雜訊，MCRA+SS

4. MCRA+LSA

圖 4-13 為空間濾波器後再經由後處理的結果，其中雜訊以最小控制之遞迴平均法做雜訊估測，並以對數頻譜幅值做增益函數。訊號的 SNR = 33.93 dB，與空間濾波的結果(圖 4-4)做比較，增加了 19.02 dB。

圖 4-14 為處理後訊號的頻譜分布圖，可看到非語音的部份幾乎都被消去，與用 LTVAD+LSA 的結果相比(圖 4-10)，效果差不多。

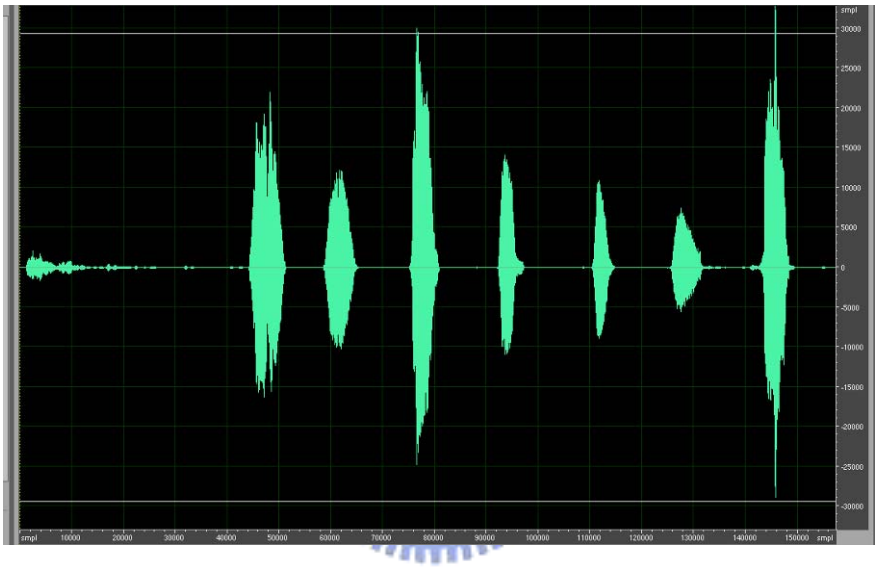

圖 4-13：高速公路雜訊空間濾波後再經過後處理，MCRA+LSA

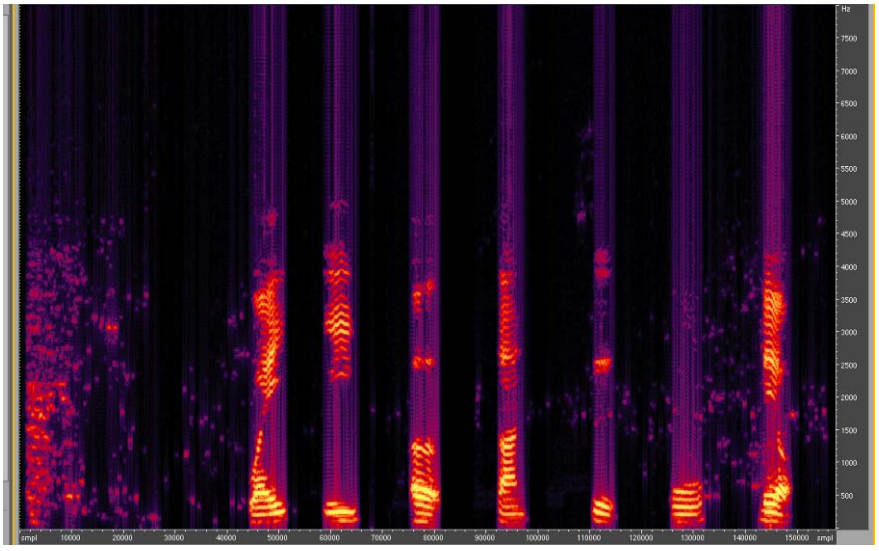

圖 4-14：頻譜分布圖。高速公路雜訊，MCRA+LSA

高速公路雜訊測試總結：

由表 4-1 可看出，高速公路雜訊類似白訊號(white noise)，沒有顯著的方向性，利用 Dahl's Beamformer 後，語音品質即有顯著的提升。純化聲源方向後，再利用單通道語音強化方法做後處理。由於這些方法在估測噪音時，是利用語音能量大小或是能量差來做判斷，對於這種穩態的雜訊可以做到很好的估測，使得後端通過增益函數能有不錯的效果。至於對數頻譜幅值(LSA)部分，幾乎是將雜訊消除，但一些語音資訊中的氣音容易被消去，造成的失真也較大。

	SNR(dB)	SNR improved from Beamformer output(dB)
Original Speech	3.85	-
Dahl's Beamformer	14.91	-
LTVAD+SS	21.57	6.66
LTVAD+LSA	33.13	18.22
MCRA+SS	19.74	4.83
MCRA+LSA	33.93	19.02

表 4-1：高速公路雜訊，訊噪比(SNR)比較表

測試二：音樂雜訊

測試雜訊為播放音樂(孫燕姿-奔)，經由空間濾波器處理後(如圖 4-6)，分別以不同雜訊估測方法搭配不同增益函數(gain function)測試語音純化效果。

1. LTVAD+SS

圖 4-15 為空間濾波器後再經由後處理的結果，其中雜訊以長時間語音活動偵測做雜訊估測，並以頻譜刪減做增益函數。訊號的 SNR = 15.66 dB，與空間濾波的結果(圖 4-6)做比較，增加了 2.38 dB。

圖 4-16 為處理後訊號的頻譜分布圖，背景雜訊部份被些微的壓抑，效果沒有很明顯。

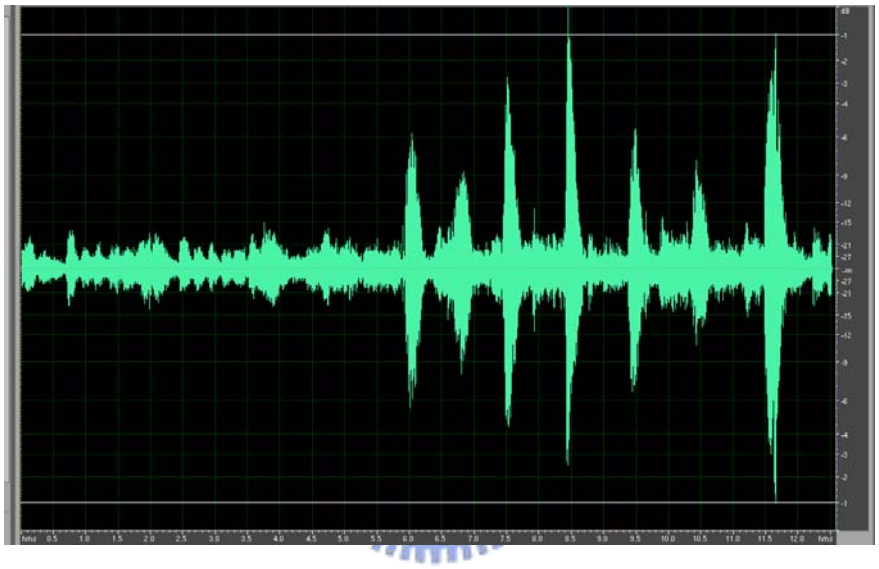

圖 4-15：音樂雜訊空間濾波後再經過後處理，LTVAD+SS

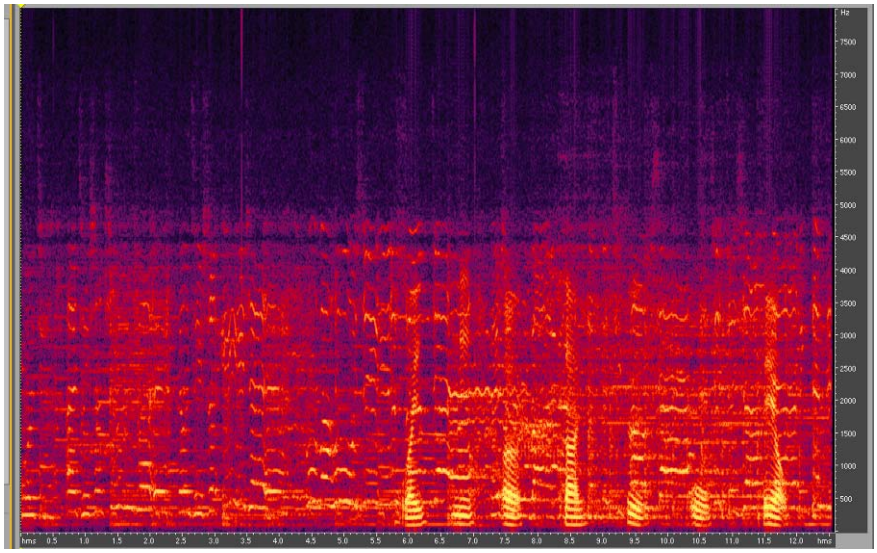

圖 4-16：頻譜分布圖。音樂雜訊，LTVAD+SS

2. LTVAD+LSA

圖 4-17 為空間濾波器後再經由後處理的結果，其中雜訊以長時間語音活動偵測做雜訊估測，並以對數頻譜幅值做增益函數。訊號的 SNR = 19.44 dB，與空間濾波的結果(圖 4-6)做比較，增加了 6.16 dB。

圖 4-18 為處理後訊號的頻譜分布圖，由於音樂中夾雜人聲，在判定有語音的音框中仍包含唱歌者的聲紋特徵；而判定沒有語音的音框中，由於空間濾波器已先對聲源方向作純化，因此只剩下一些聲紋特徵較明顯的部份。

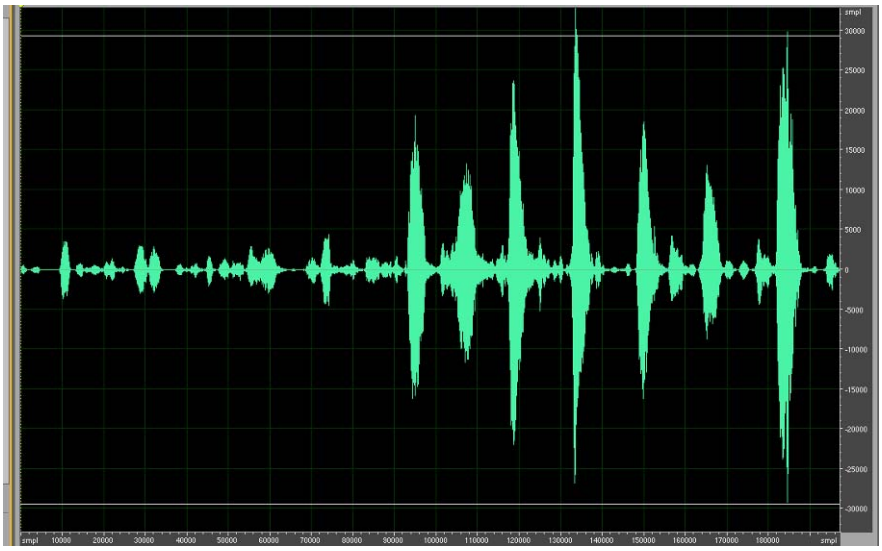

圖 4-17：音樂雜訊空間濾波後再經過後處理，LTVAD+LSA

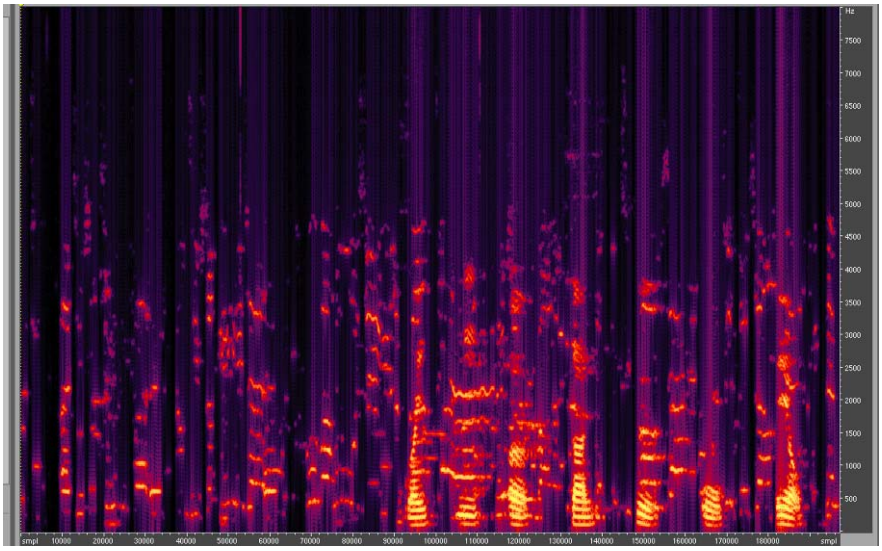

圖 4-18：頻譜分布圖。音樂雜訊，LTVAD+LSA

3. MCRA+SS

圖 4-19 為空間濾波器後再經由後處理的結果，其中雜訊以長最小控制之遞迴平均法做雜訊估測，並以頻譜刪減做增益函數。訊號的 SNR = 16.14 dB，與空間濾波的結果(圖 4-6)做比較，增加了 2.86 dB。

圖 4-20 為處理後訊號的頻譜分布圖，由圖中可見背景的顏色變深，頻譜圖看起來較清晰，表示背景偏白訊號的成分被消除。至於聲紋特徵明顯的部份幾乎沒有被壓抑，包含音樂中的人聲。與 LTVAD+SS 的結果比較，較能將訊號中白訊號的成分消除。

圖 4-19：音樂雜訊空間濾波後再經過後處理，MCRA+SS

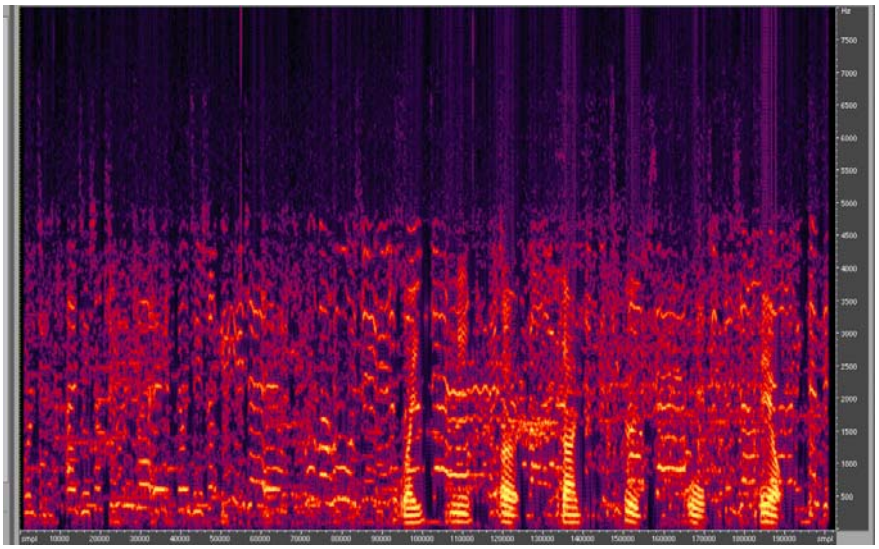

圖 4-20：頻譜分布圖。音樂雜訊，MCRA+SS

4. MCRA+LSA

圖 4-21 為空間濾波器後再經由後處理的結果，其中雜訊以最小控制之遞迴平均法做雜訊估測，並以對數頻譜幅值做增益函數。訊號的 SNR = 19.87 dB，與空間濾波的結果(圖 4-6)做比較，增加了 6.59 dB。

圖 4-22 為處理後訊號的頻譜分布圖，由圖可看出聲紋特徵較明顯的部份都被保留下來，效果比 LTVAD+LSA 要好一點。

圖 4-21：音樂雜訊空間濾波後再經過後處理，MCRA+LSA

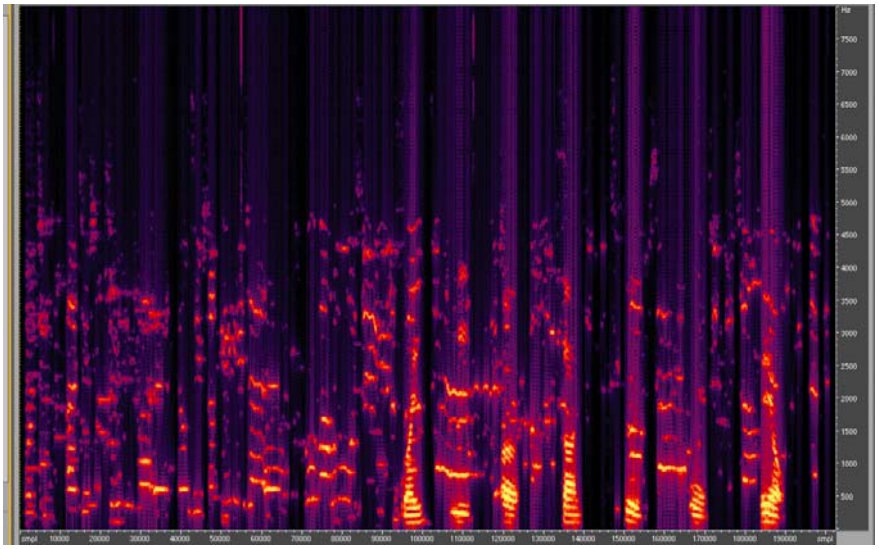

圖 4-22：頻譜分布圖。音樂雜訊，MCRA+LSA

音樂雜訊測試總結：

由於音樂中夾雜人聲，且為非穩態雜訊，因此語音純化的效果會比高速公路雜訊的情況差。在頻譜刪減(SS)下改善有限，以對數頻譜幅值(LSA)較有顯著的效果。

其中值得注意的是，若只做單聲道強化語音的方法(如 SS 或 LSA)，會因為背景音樂的人聲部分未被壓抑，使得大部分音樂中人聲的聲紋被保留，純化效果大大下降。這也是為什麼要結合空間濾波的原因及優勢。

	SNR(dB)	SNR improved from Beamformer output(dB)
Original Speech	8.61	-
Dahl's Beamformer	13.28	-
LTVAD+SS	15.66	2.38
LTVAD+LSA	19.44	6.16
MCRA+SS	16.14	2.86
MCRA+LSA	19.87	6.59

表 4-2：音樂雜訊，訊噪比(SNR)比較表

第五章 結論

5.1 研究成果

本論文結合適應性空間濾波與後濾波，並測試在不同雜訊估測與增益函數組合的結果。結合適應性空間濾波與後濾波最大的好處在於，適應性空間濾波是針對空間資訊對聲源方向作純化，但該方向仍包含雜訊，所以需要靠後濾波做進一步的純化。相對於後濾波而言，在音樂雜訊下包含語音成分，若單獨做沒辦法將不必要的語音成分壓抑。透過空間濾波將聲源方向外的語音成分壓抑後，後濾波才能再進一步對目標聲源做純化。

在雜訊估測方面，LTVAD 是對一個音框做判斷，而 MCRA 則是在各個頻率下各自判斷是否為雜訊，一般來說效果會比 LTVAD 好。

在增益函數方面，頻譜刪減法對於非穩態雜訊(如音樂雜訊)的效果較差，對數頻譜幅值則有不錯的雜訊壓抑效果。

5.2 未來展望

對數頻譜幅值能有不錯的訊噪比提升，但相對的也造成了較大的失真，針對這個部份還有改善的空間。另外，直接將後濾波的一些方法結合進空間濾波成為多通道處理的方法也是未來研究的一個方向。

Reference

[1] D. Johnson and D. Dudgeon, Array Signal Processing: Concepts and Techniques,Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[2] Barry D. Van Veen and Kevin M.Buckley,"Beamforming:A Versatile Approach to Spatial Filtering,"IEEE ASSP MAGAZINE April 1988.

[3] Zoltowski, M., "High resolution sensor array signal processing in the beamspace domain: novel techniques based on the poor resolution of Fourier beamforming," Spectrum Estimation and Modeling, 1988., Fourth Annual ASSP Workshop on , pp. 350 –355, 1988

[4] Byung-Chul Kim; I-Tai Lu , "High resolution broadband beamforming based on the MVDR method," OCEANS 2000 MTS/IEEE Conference and Exhibition , Volume: 3 , pp. 1673 –1676, 2000

[5] Marciano, J.S., Jr.; Vu, T.B., "Reduced complexity MVDR broadband beamspace beamforming under frequency invariant constraint," Antennas and Propagation Society International Symposium, 2000. IEEE , Volume: 2 , pp. 902 –905, 2000

[6] Pillai, S. Unnikrishna, Array signal processing, 1989

[7] Ta-Sung Lee; Tsui-Tsai Lin , "Coherent interference suppression with complementally transformed adaptive beamformer," Antennas and Propagation, IEEE Transactions on , Volume: 46 Issue: 5 , pp. 609 –617, May 1998

[8] Gollamudi, S.; Yih-Fang Huang , "Optimally combined nonlinear MMSE beamforming and interference cancellation for CDMA communications," Personal Wireless Communications, 2000 IEEE International Conference on , pp. 474 –478, 2000

[9] Dahl, M.; Claesson, I., "Acoustic noise and echo cancelling with microphone array," Vehicular Technology, IEEE Transactions on , Volume: 48 Issue: 5 , Sept.1999 Page(s): 1518 –1526

[10] Javier Ramírez , José C. Segura , Carmen Benítez , Ángel de la Torre and Antonio Rubio ,"Efficient voice activity detection algorithms using long-term speech information," Speech Communication, Volume 42, Issues 3-4, April 2004, Pages271-287

[11] Cohen I., Berdugo B., "Noise estimation by minima controlled recursive averaging for robust speech enhancement," Signal Processing Letters, IEEE, Issue:1, pp. 12-15, 2002.

[12] R. Martin, “Spectral subtraction based on minimum statistics,” in Proc. 7th EUSIPCO'94 Edinburgh, U.K., Sept. 13\–16, pp. 1182-1185, 1994.

[13] Berouti M., Schwartz R., Makhoul J., “Enhancement of speech corrupted by acoustic noise ,” Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79., Volume:4, pp. 208-211, 1979.

[14] Ephraim Y., Malah D., “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on, Volume:33, Issue:2, pp. 443-445, 1985.

[15]European Digital Cellular Telecommuni- cations System; Half rate speech; Voice Activity Detection (VAD), ETSI GSM 06.42 (ETS 300-581-6), 1995.

[16] European Digital Cellular Telecommuni- cations System; Half rate speech; Half rate speech transcoding, ETSI GSM 06.20 (ETS 300-581-2), 1995.

[17]ITU-T G.729, Coding of Speech at 8kbit/s Using CS-ACELP, March, 1996.

[18] A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin, and J. P. Petit, “ITU recommendation G.729 annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications,” IEEE Commun. Mag., vol. 35, pp. 64–73, Sept. 1997.

[19] R. Martin, “Spectral subtraction based on minimum statistics, Proceedings of the Seventh European Signal Processing Conference, EUSIPCO-94, Edinburgh, Scotland, 13–16 September 1994, pp. 1182–1185.

[20] Y. Ephraim, D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” IEEE Trans. Acoust. Speech Signal Process. ASSP-32 (6) (December 1984) 1109–1121.

[21] O. CappTe, “Elimination of the musical noise phenomenon with the phraim and Malah noise suppressor,” IEEE Trans. Speech Audio Process. 2 (2) (April 1994) 345–349.

[22] Israel Cohen, Baruch Berdugo, “Speech enhancement for non-stationary noise environments,” Signal Processing, Volume:81, Issue:11, pp. 2403-2418, 2001.