國 立 交 通 大 學

電機與控制工程研究所

碩 士 論 文

使用麥克風陣列實現即時語音純化與真人語音活動偵測系統

A Real-time Speech Purification and Voice Activity Detection System Using Microphone Array

研 究 生： 楊 佳 興

指導教授： 胡 竹 生 博士

中 華 民 國 九 十 四 年 六 月

使用麥克風陣列實現即時語音純化與真人語音活動偵測系統

A Real-time Speech Purification and Voice Activity Detection System Using Microphone Array

研 究 生：楊 佳 興　　Student：Chia-Hsing Yang

指導教授：胡 竹 生 博士　Advisor：Jwu-Sheng Hu

國立交通大學
電機與控制工程學系

碩 士 論 文

A Thesis
Submitted to Institute of Electrical and Control Engineering
College of Electrical Engineering and Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of Master
in

Electrical and Control Engineering

June 2005

Hsinchu, Taiwan, Republic of China

中 華 民 國 九 十 四 年 六 月

國立交通大學

論文口試委員會審定書

本校 電機與控制工程 學系碩士班 楊佳興 君

所提論文 使用麥克風陣列實現即時語音純化與真人語音活動偵測系統
A Real-time Speech Purification and Voice Activity Detection System Using Microphone Array

合於碩士資格標準、業經本委員會評審認可。

口試委員：

_____ _____

_____ _____

_____ _____

_____ _____

_____ _____

指導教授： 教授

系 主 任： 教授

中 華 民 國 九十四 年 月 日

使用麥克風陣列實現即時語音純化與真人語音活動偵測系統

研究生：楊　佳　興　　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

本論文針對桌面及車內吵雜環境提出一即時抑制噪音源與真人語音活動偵測系統。並提出一個符合實驗室規模、使用 USB1.1 介面、8 通道之即時性麥克風陣列訊號處理實驗平臺，且已實作完成。為了讓系統能適應於噪音源的變化與環境的特徵，論文中設計出一適應性空間濾波器並實做完成。適應性訊號處理的啟動時機則由真人語音活動偵測結果所決定。演算法皆實現於個人電腦上，並利用最佳化的函式庫來達到即時的效果。為了確認系統效能，此語音純化系統與兩種語音辨識器做結合，由實驗證明，此演算法能抑制干擾源的影響，並用來提升語音辨識率。

A Real-time Speech Purification and Voice Activity Detection System Using Microphone Array

Student： Chia-Hsing Yang Advisor： Prof. Jwu-Sheng Hu

Institute of Electrical and Control Engineering

ABSTRACT

This thesis proposes a real-time interference suppression and voice activity detection (VAD) system for applications in desktop and vehicle environment where speech reception is contaminated by various noise and sound sources. A microphone array of 8 microphones with USB 1.1 interface was made for the implementation platform. To cope with the changes of noise and environmental characteristics, an adaptive beamformer with on-line noise channel calibration is designed and implemented. The VAD result is used to ensure the correct timing for on-line calibration. The algorithms are implemented on a PC platform and to guarantee real-time, optimization of intensive computations is also studied. To verify the effectiveness of the proposed system, the purified speech signal is combined with two ASR (Automatic Speech Recognition) system. The experimental results show that the algorithms are able to reduce the interference and enhance the ASR rate.

誌　謝

對於本論文的完成，首先，感謝指導老師胡竹生教授兩年來的辛勤指導。在這兩年中，老師不僅教導我在學術上的知識，更教導我一個研究生應有的研究態度，解決問題的方法，讓我了解到研究生就是應該對自己的研究負責。

而在這些日子，必須感謝實驗室眾多學長、同學和學弟妹的陪伴與協助。感謝价呈與維瀚兩位語音組的帥哥先驅，不僅教導我學術上的知識外更提供我許多財經方面相關訊息，跟你們一起做事，我真的很開心。謝謝立偉與宗敏學長，你們的熱心協助，真的讓我很感動。還有幾位一起奮鬥的同學，士奇，晏榮、群棋、岑思和鏗元，有了你們的陪伴，讓研究生活添加了許多歡樂。另外也謝謝學弟妹們，朱木、螞蟻、鳥蕙、佩靜、耀賢和恆嘉，謝謝你們的陪伴。

另外，還有眾多好友們。謝謝怡君，有了妳在大學及研究所生活的陪伴及捅了許多摟子，讓我在交大的生活能更快樂更完美。也謝謝我的室友們，明峰、裕隆、達哥和阿同，你們讓我在回寢室後能忘記所有的煩惱與壓力，讓我了解到豁達的生活態度。還有一路從國中陪伴到現在的好友們，貝克漢、丁丁丁、英傑、毅修、鳥佐及胖子，真的謝謝你們。

最感謝的就是我的家人，父親楊忠民先生、母親李碧雪女士及哥哥佳元，真摯的感謝你們給我一個無憂無慮的生活與溫暖的家，讓我能夠專心於學業上而無後顧之憂，真的最謝謝你們。也謝謝我的叔叔楊孝志先生及楊仁祥先生還有姑姑楊雪娥女士，謝謝你們在生活上的鼓勵與幫助，讓我的人生能更有信心與方向。謹以本論文向家人獻上最誠摯的謝意。

目　錄

表　列

圖　列

第一章 緒論

1.1 研究動機

環境中的語音訊號干擾源總是存在，例如冷氣機、電腦風扇、喇叭、密閉空間反射等等。當語音訊號遭到干擾時，若用於語音辨識中，辨識率會大為降低，若用於通訊中，通話品質也大受影響。因此若能設計出一語音輸入介面，降低環境中干擾源的影響，達到語音純化的效果，則在生活中將會有很大的應用面。

在論文中，我們利用麥克風陣列來對語音作純化的動作，只要能對當時的環境訊號作適應性空間濾波（Spatial Filter），則可對不同角度入射的訊號有不同的增益，以降低干擾源對語音訊號的影響，達到提升訊噪比（SNR）的作用。除了適應性空間濾波的功能外，我們額外加入真人語音活動偵測（Voice Activity Detection，VAD）的功能，讓系統能依據真人語音有無自動地適應性調整空間濾波器係數。

1.2 研究目標

本論文目標將分為

1. 選定真人語音活動偵測及適應性空間濾波器演算法
2. 發展一套麥克風陣列平台，能夠將語音訊號透過 USB 介面傳回電腦，並在電腦作演算法處理及即時性喇叭輸出。其系統簡圖如圖 1-1 所示。
3. 將演算法實現於平台，作即時性的處理。
4. 將平台與語音辨識器做整合

1.3 文獻回顧

陣列訊號處理技術早於第一次世界大戰時被提出並加以利用[1]，當時法國人 Sergent Jean Perrinm 用了兩組感測器，每組感測器由六組次感測器所組成，此發明是用來偵測敵機。之後，陣列訊號處理技術也被用於聲納[2]，陣列望遠鏡（如美國新墨西哥洲沙漠中的特大天線陣列（Very Large Array），它由 27 個碟形天線以Y字形分佈）等等。而早年的陣列訊號處理技術皆用於軍事或大型儀器上，直到最近，在電子元件普及與運算能力越來越強大的趨勢下，陣列訊號處理技術也慢慢走向消費性產品話，如麥克風陣列。

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[3]，接著慢慢衍生出諸如 Fourier Beamformer[4]、MVDR(Minimum Variance Distortionless Response Beamformer)[5][6]、Robust MVDR[7]、MCMV(Multiply Constrained Minimum Variance Beamformer)[8]、MMSE(Minimum Mean Square Error Beamformer) [9]、MSNR(Maximum SNR)[7]、ML(Maximum Likelihood Beamformer)[7]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。

在國內，麥克風陣列的製作廠商幾乎沒有，而本實驗室 1999 年時，曾經以 Fourier beamformer 為基礎，設計一組 real-time spatial filter and DOA estimation system[10][11]，此系統包含一個 16 channel 的 microphone array 以及對應的 signal conditioner module、sampler module 以及 DSP Module，並能即時估測空間之聲源方向。並在 2004 年設計出以 USB1.1 為介面之語音純化系統[12]。

1.4 論文貢獻

本論文已實作完成一以 USB1.1 為傳輸介面之八通道麥克風陣列平台，此平台有低成本、低消耗功率且隨插即用等優點。本論文在演算法上，將真人語音活動偵測（VAD）與空間濾波器（Beamformer）做整合，達到自動適應性調整空間濾波器功能，並將演算法實作完成於八通道麥克風陣列平台上，擁有即時的效能。論文中，麥克風陣列平台與語音辨識器做結合，並在吵雜的環境中做測試，由實驗證明，麥克風陣列平台能夠純化語音，用來提升語音辨識率。

1.5 論文架構

本篇論文包含了三個主要的部分，分別是即時性演算法的理論、實驗平臺的架構與實現與即時性演算法的驗證。底下將大致描述三個主要部分的內容：

第二章：將介紹陣列訊號處理概念、語音活動偵測演算法、適應性訊號處理簡介和適應性陣列訊號處理-Dahl's Algorithm

第三章：介紹語音辨識器與 IBM ViaVoice

第四章：介紹實驗平台架構

第五章：演算法在實驗平台的驗證

第六章：結論

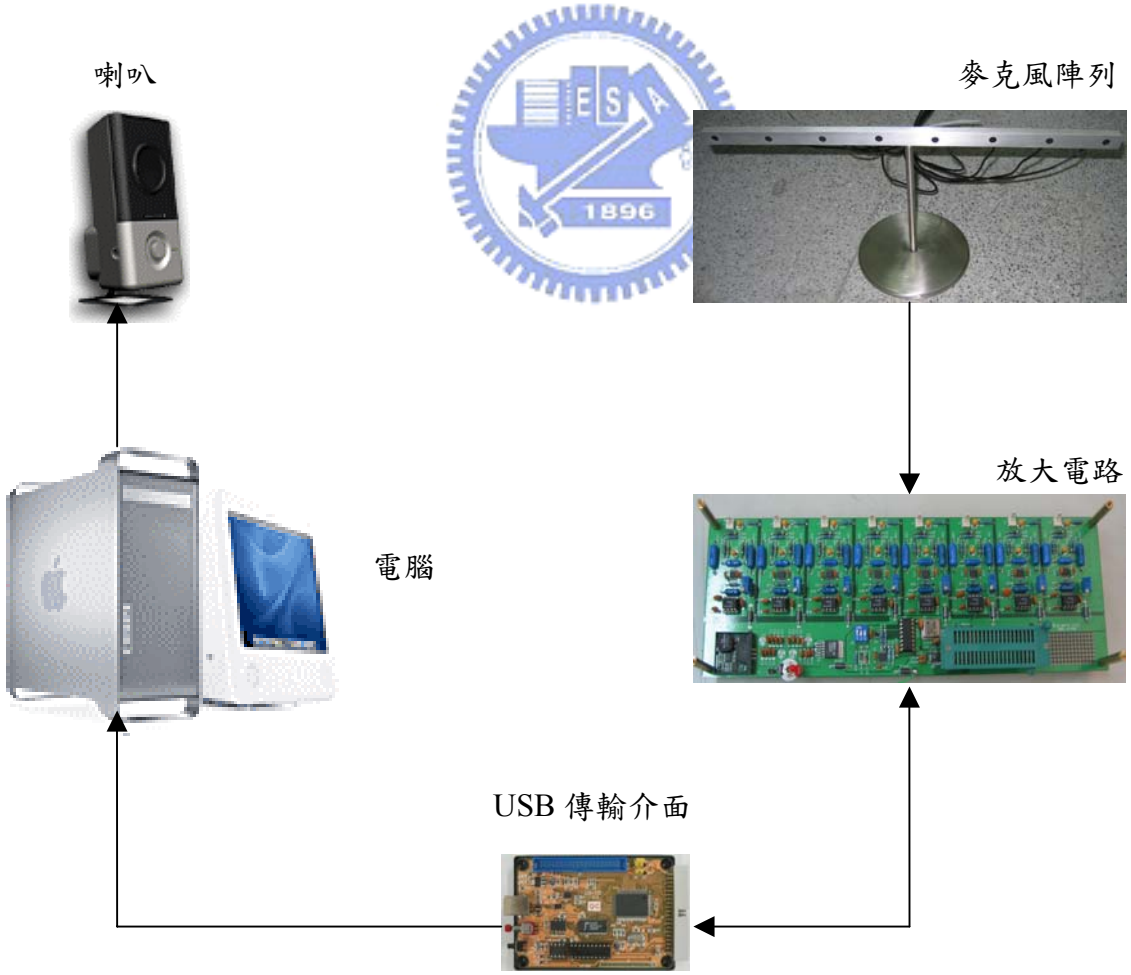

圖 1-1：本論文系統架構簡圖

第二章 適應性陣列訊號處理

2.1 陣列訊號處理

2.1.1 陣列式訊號處理簡介

在傳統數位訊號處理研究中，大多著重於時域訊號的處理技巧，通常先將連續訊號進行取樣，接著通過濾波器以區分訊號中不同的成分，但當原始訊號與雜訊在頻譜上極相似時，一般的時域濾波器很難將原始訊號與雜訊分開，以單一麥克風為例子，若同時接收到兩人說話的聲音，因為兩人聲音頻帶重疊性很高的原因，時域訊號的處理方式很難將兩人聲音分開，因此若需要還原原始訊號，則需要對訊號進行空間取樣以獲得空間資訊[1]。

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[1]。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

- 窄頻訊號（Narrowband signal）
- 遠場平面波（Far field plane wave）

假設一陣列感應器排置如圖一所示，s(t)為原始訊號，n(t)為雜訊

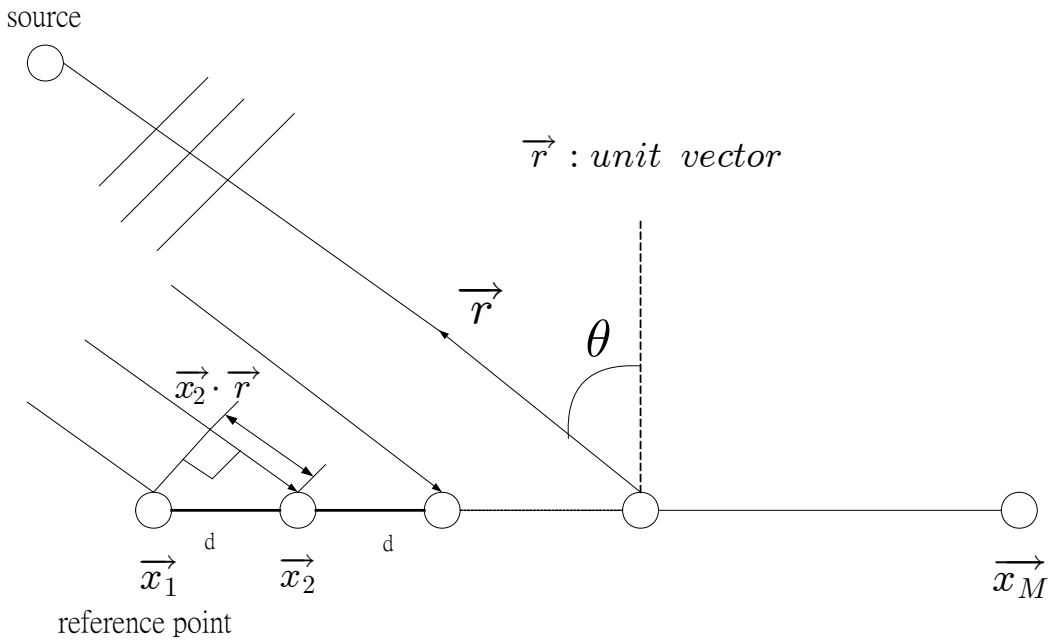

圖 2-1：一維陣列與平面波入射關係圖

則 M 個感應器輸出可寫成下列向量形式

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

$$= \begin{bmatrix} e^{jk_c \vec{x}_1 \cdot \vec{r}} \\ \vdots \\ e^{jk_c \vec{x}_M \cdot \vec{r}} \end{bmatrix} s(t) + \begin{bmatrix} n_1(t) \\ \vdots \\ n_M(t) \end{bmatrix} = a(\vec{r})s(t) + n(t) \qquad (2\text{-}1)$$

$k_c = \frac{w_c}{c} = \frac{2\pi}{\lambda_c}$ 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

2.1.2 陣列型態：均勻線性陣列（Uniform Linear Array）

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，其架構圖如圖 2-2 所示。

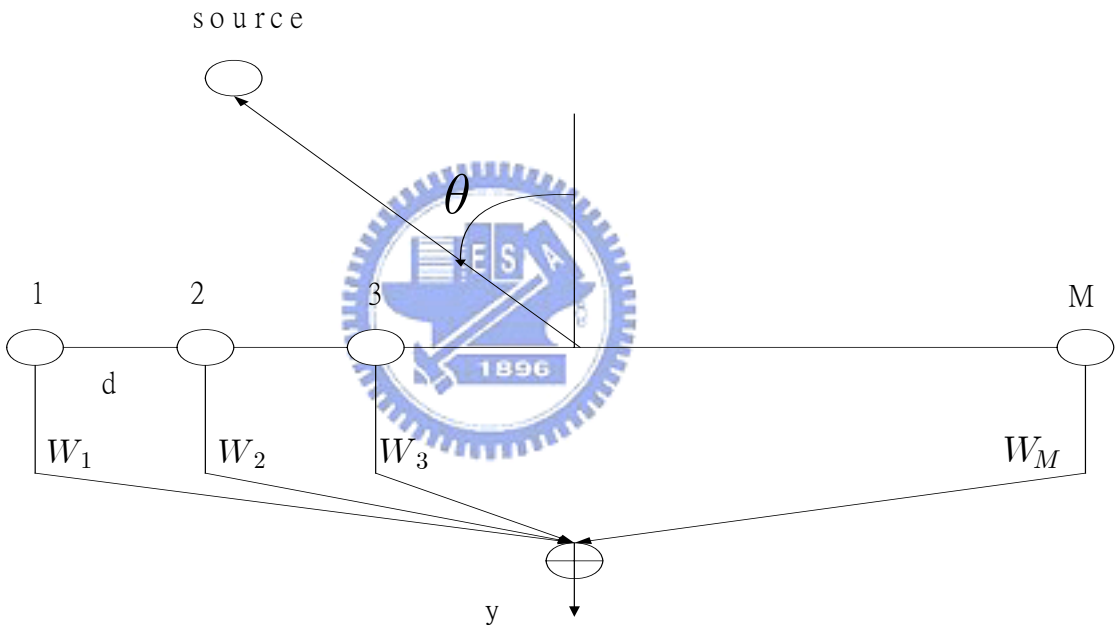

圖 2-2：均勻線性陣列架構圖

若以第一個感應器當作參考點，每個感應器對於訊號源的相對角度皆為 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。

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

2.1.3 均勻線性陣列空間響應

均勻線性陣列的架構如圖 2-2 所示，其中 W 指的是每個感應器輸出乘上的加權，而空間濾波器（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：

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

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：

$$p(\theta)=\sum_{i=1}^{M} e^{jk_c(i-1)d\sin\theta}=\frac{e^{jk_c Md\sin\theta}-1}{e^{jk_c d\sin\theta}-1}$$

$$=e^{j\frac{k_c(M-1)d}{2}\sin\theta}\frac{\sin\left(\frac{k_c Md}{2}\sin\theta\right)}{\sin\left(\frac{k_c d}{2}\sin\theta\right)} \qquad (2\text{-}4)$$

若將 $p(\theta)$ 取 Magnitude 可得其 beampattern，如圖 2-3 所示

從圖 2-3 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。

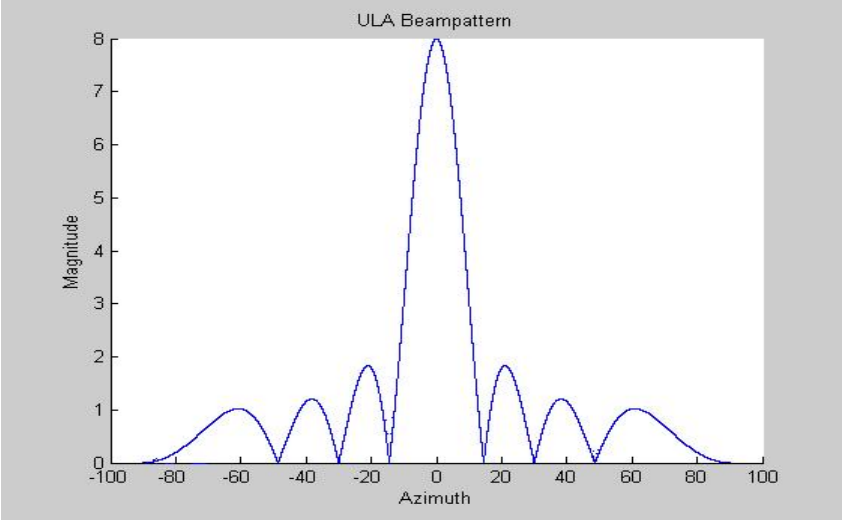

圖 2-3：均勻線性陣列之空間響應（M=8，frequency=100Hz，d=10）

2.1.4 均勻線性陣列特性

和時域濾波器一樣，空間濾波器也會有一些基本的特性，本章節將針對均勻線性陣列，介紹其基本特性[13]。

■ **Grating Lobe 問題**

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 2-4 所示。

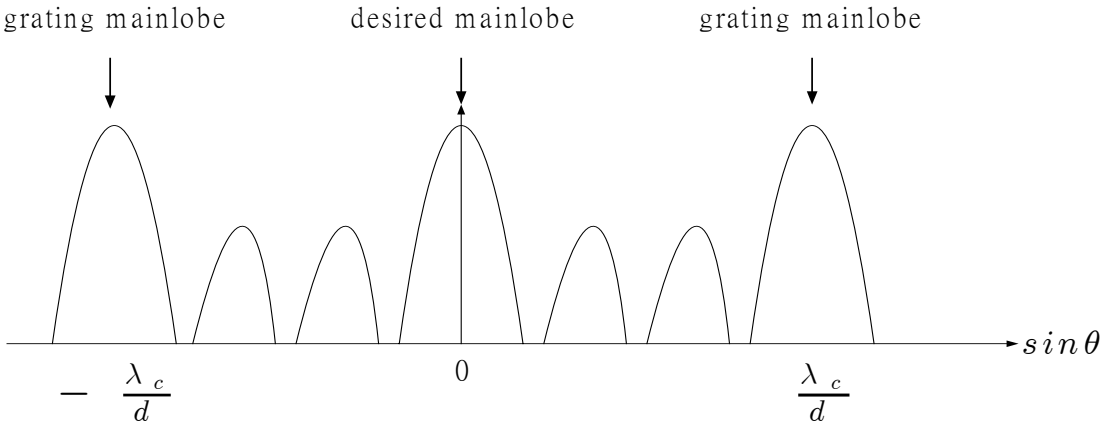

圖 2-4：均勻線性陣列 Grating Lobe 示意圖

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 語音活動偵測（Voice Activity Detection，VAD）

語音活動偵測是用來判定是否有真人語音，近年來已廣泛用於通訊上達到節省能量耗損的目的。若用於語音辨識方面是屬於語音辨識的前處理，對辨識結果的影響很大，精確的語音活動偵測可降低噪音影響並提高辨識率。傳統的語音活動偵測大多使用語音能量或過零率（zero-crossing rate）等資訊來判別，本節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[14]。

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，針對長時間語音資訊，定義出下列定義。

■ Long-Trem Spectrum 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，那麼 N 階的 LTSE 定義為：

$$\mathrm{LTSE}_{\mathrm{N}}(k,l)=\max\{X(k,l+j)\}_{j=-N}^{j=+N} \qquad (2\text{-}6)$$

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 個頻域絕對值內的最大值。而 LTSE 則代表了長時間語音資訊的意義，因為 LTSE 不只是對單一音框取值，而是針對 2N+1 個音框取最大值，這樣的好處是不容易忽略某些字頭的子音或是摩擦音。除了 LTSE 外，為了判定是否為真人語音，必須定義另一項定義 LTSD。

■ Long-Term Spectral Divergence（LTSD）

LTSD 的定義如（2-7）式：

$$\mathrm{LTSD}_{\mathrm{N}}(l)=10\log_{10}\left(\frac{1}{NFFT}\sum_{k=0}^{NFFT-1}\frac{LTSE^{2}(k,l)}{N^{2}(k)}\right) \qquad (2\text{-}7)$$

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 2-8 式：

$$N_{K}(k)=\frac{1}{2K+1}\sum_{j=-K}^{j=K}X(k,l+j) \qquad (2\text{-}8)$$

從（2-8）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：

$$\gamma=\begin{cases}\gamma_{0} & E\le E_{0}\\ \gamma_{0}+\dfrac{\gamma_{1}-\gamma_{0}}{E_{1}-E_{0}}(E-E_{0}) & E_{0}<E<E_{1}\\ \gamma_{1} & E\ge E_{1}\end{cases} \qquad (2\text{-}9)$$

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

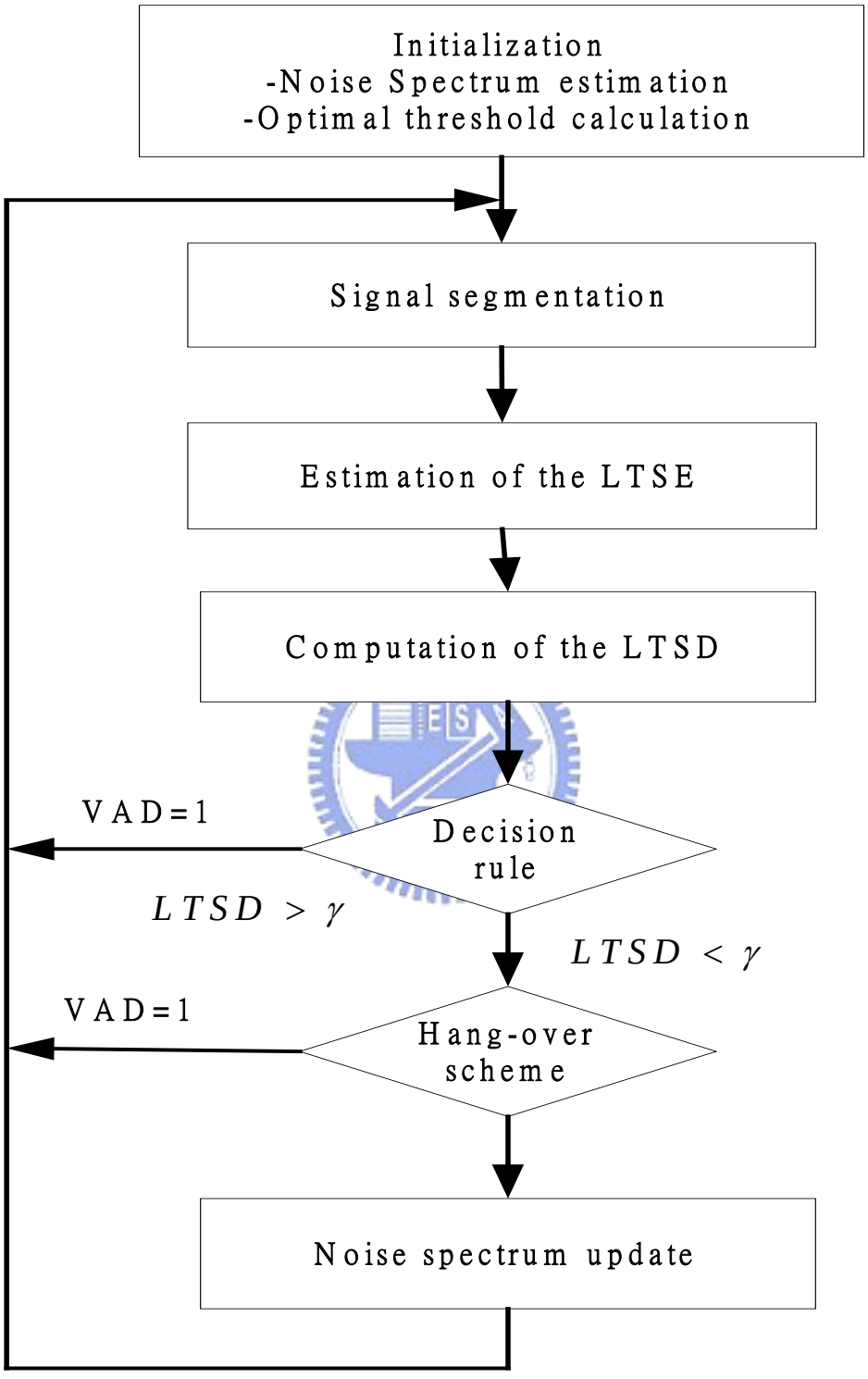

圖 2-5：VAD 演算法流程圖

臨界值，因此E_0, E_1, γ_0和γ_1是先設定好的初始值。從（2-9）式可觀察出當現階段雜訊能量介於E_0和E_1時，則γ會依$E-E_0$在E_1-E_0所佔的比例，作出γ_0的線性調整。而 VAD 演算法的流程如圖 2-5 所示：

■ VAD 演算法流程解說

1. 設定初始值E_0, E_1, γ_0和γ_1。
2. 將語音作切割，一個音框為 30ms，而音框和音框的交疊為 20ms。
3. 計算 LTSE 和 LTSD。
4. 將 LTSD 與γ作比較，若 LTSD > γ則判定為真人語音，若 LTSD < γ，則經過 Hang-Over 機制。
5. 經過 Hang-Over 機制，若為非真人語音，則更新雜訊頻譜絕對值平均$N(k)$。

Hang-Over 機制是為了延長字母尾音判定為真人語音的機制，因為字母尾音部分通常能量較小，容易被判定為非真人尾音，因此系統中加入 Hang-Over 機制，彌補字母尾音能量小的問題。另外在更新雜訊頻譜絕對值平均$N(k)$方面，並非完全的更新，而是利用了適應性訊好處理的觀念，定義如下：

$$N(k,l)=\alpha N(k,l-1)+(1-\alpha)N(k) \qquad (2\text{-}10)$$

其中，k 到表頻率，l 代表音框，從（2-10）式可看出，$N(k)$的更新，除了有現階段$N(k)$的資訊外，也包含了上一個音框的$N(k)$資訊，而此權重α可依照環境自行調整。

2.2.1 VAD 模擬

本章節將上述 VAD 演算法，用 Matlab 模擬，先以取樣頻率為 16k Hz，用單一麥克風錄製 12 s 的語音，並對其作 VAD 的判定。圖 2-6 展示出有做長時間語音資訊 VAD 的結果，也就是 2-6 式取 N=6。而圖 2-7 展示出沒做長時間語音資訊 VAD 的結果，也就是 2-6 式取 N=0。

圖 2-6 和圖 2-7 的上半部分展示出 LTSD 與 γ 的關係，變動較大的為 LTSD，從圖中可觀察出，沒做長時間語音資訊（N=0）的 LTSD 變動較大，這是因為每前進一個音框，LTSD 都會有新的值。比較兩張圖可發現，沒做長時間語音資訊（N=0）的 VAD 結果很容易將字頭字尾部分，並判定為非真人語音，而有做長時間語音資訊（N=6）的 VAD 就不容易忽略字頭字尾部分。

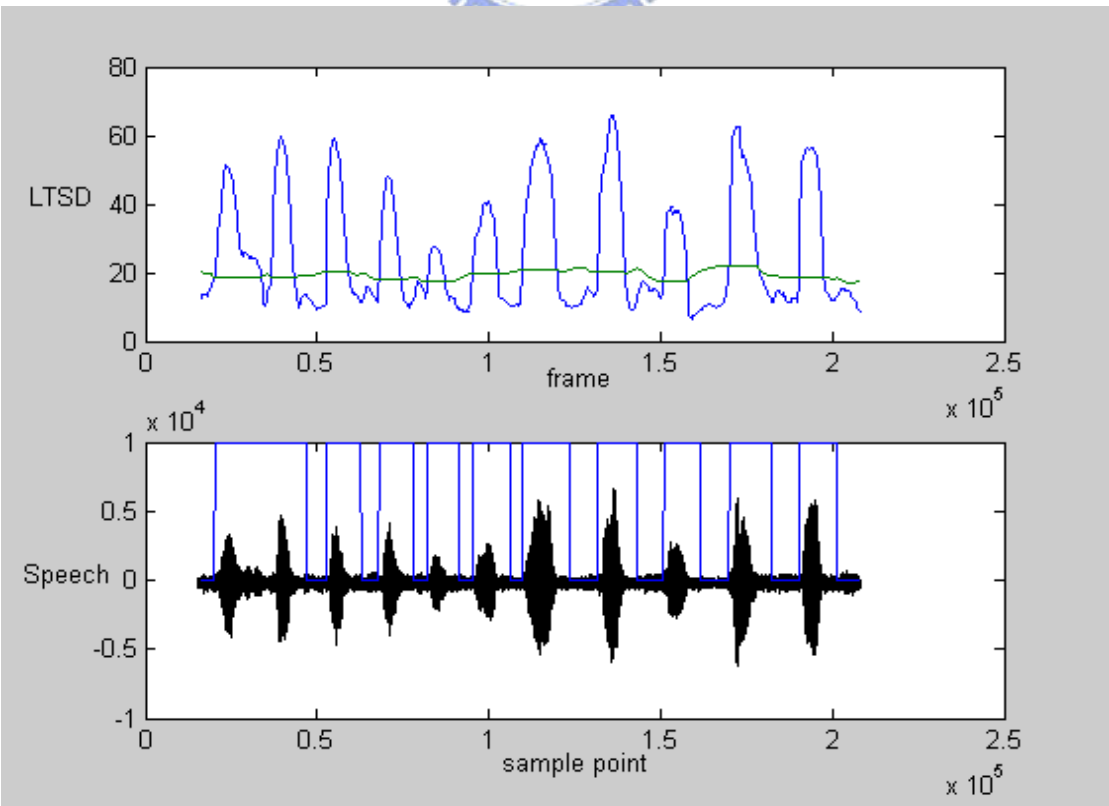

圖 2-6：上半部：LTSD 與 γ 關係圖　下半部：VAD 模擬結果 N=6

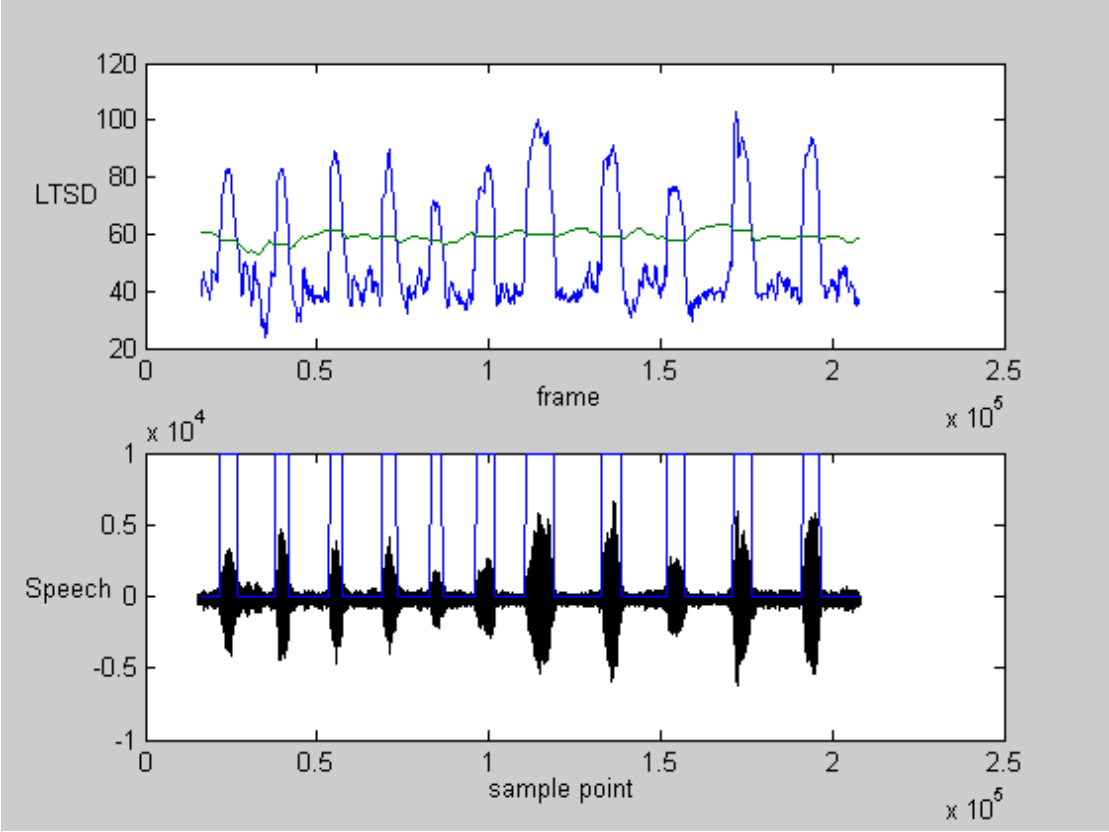

圖 2-7：上半部：LTSD 與 γ 關係圖　下半部：VAD 模擬結果 N=0

2.3 適應性訊號處理

2.3.1 適應性濾波器簡介

通常而言，濾波器的係數通常設計出來後皆為固定的，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。

2.3.2 適應性濾波器處理架構

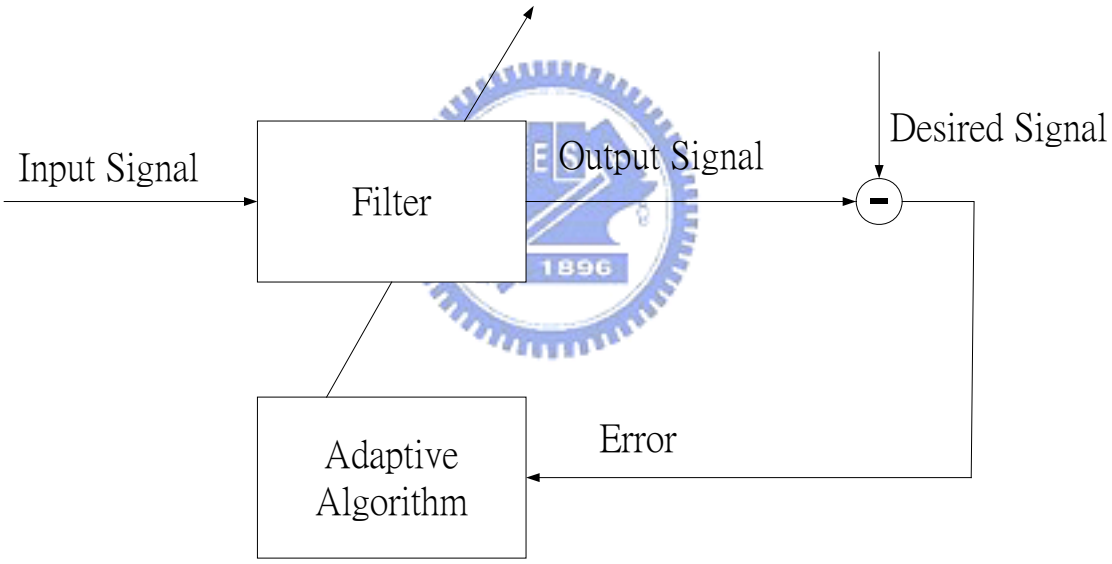

圖 2-8：適應性濾波器處理架構圖

適應性濾波器處理架構圖如圖 2-8 所示，當輸入訊號經過濾波器處理後，會與希望達成的訊號相減而產生誤差訊號，此誤差訊號經過適應性演算法的運算而調整濾波器的係數，使誤差訊號降低。如此反覆運算會讓濾波器係數不斷地變動，讓濾波器輸出訊號與希望達成的訊號愈來愈相近。

2.3.3 Least-Mean-Square（LMS）Algorithm

LMS 演算法指的是，找出一組權重 W 使得誤差平方項最小[19]。其基本架構如圖 2-8 所示。假設希望達成的訊號為 zero-mean，其變異量為 σ_d^2

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

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

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

因此目標函數如（2-11）式所示

$$J(w)\equiv\min_w E\{d-xw\}^2=E(d-xw)(d-xw)^* \qquad (2\text{-}11)$$

2-11 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，其標準式如下：

（new guess）=（old guess） ＋ （a correction term）

也就是

$$w_i=w_{i-1}+\mu p,\qquad i\geq 0 \qquad (2\text{-}12)$$

其中（2-12）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（2-11）式下手，將（2-11）式展開可得

$$J(w)=\sigma_d^2-R_{dx}^*w-w^*R_{dx}+w^*R_xw \qquad (2\text{-}13)$$

為了找組 W 使 $J(w)$ 最小，對 2-13 式取 ∇_w 得

$$\nabla_w J(w)=w^*R_x-R_{dx}^* \qquad (2\text{-}14)$$

因此，為了讓 w 往 $J(w)$ 最低處的方向與強度前進，我們取

$$p=-\left[\nabla_w J(w_{i-1})\right]^*=R_{dx}-R_xw_{i-1} \qquad (2\text{-}15)$$

（2-12）式可寫為

$$w_i = w_{i-1} + \mu\left[R_{dx} - R_x w_{i-1}\right] \qquad i \geq 0 \qquad \text{（2-16）}$$

在實做上，R_{dx}和R_x可用離散形式近似於瞬間值：

$$R_{dx} = d(i)x^*(i) \qquad R_x = x^*(i)x(i) \qquad \text{（2-17）}$$

所以（2-16）是可寫為：

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

（2-18）

因此，LMS Algorithm 可整理如下：

Filter out ： $y(i) = x(i)w(i)$ （2-18）

Error function： $e(i) = d(i) - y(i)$ （2-19）

Update weight： $w(i) = w(i-1) + \mu x^*(i)e(i) \qquad i \geq 0$ （2-20）

其方塊圖如下所示。

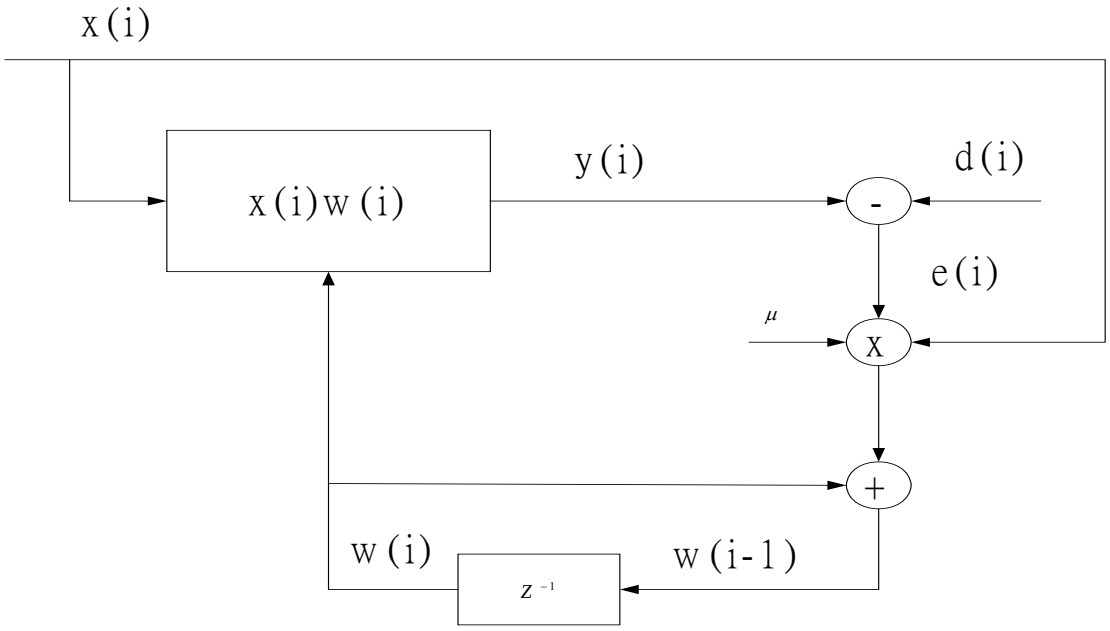

圖 2-9：LMS 演算法方塊圖

2.3.4 Normalize LMS Algorithm

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[19]：

Filter out ： $y(i) = x(i)w(i)$ （2-21）

Error function： $e(i) = d(i) - y(i)$ （2-22）

Update weight： $w(i) = w(i-1) + \frac{\alpha x(i) w(i)}{\gamma + x^*(i)x(i)} \qquad i \geq 0$ （2-23）

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，$0<\alpha<2$，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。

2.4 適應性陣列訊號處理

2.4.1 適應性陣列訊號處理簡介

在作陣列訊號處理時，會假設兩條件：

- 窄頻訊號（Narrowband signal）
- 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的訊號特徵，則可利用演算法去調整適應性濾波器，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，再利用適應性訊號處理演算法來設計「適應性空間濾波器」，及就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。

2.4.2 適應性空間濾波器：Dahl's Algorithm

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 2-10 所示。

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，再將此訊號儲存至硬碟。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硬碟。舉例來說，若環

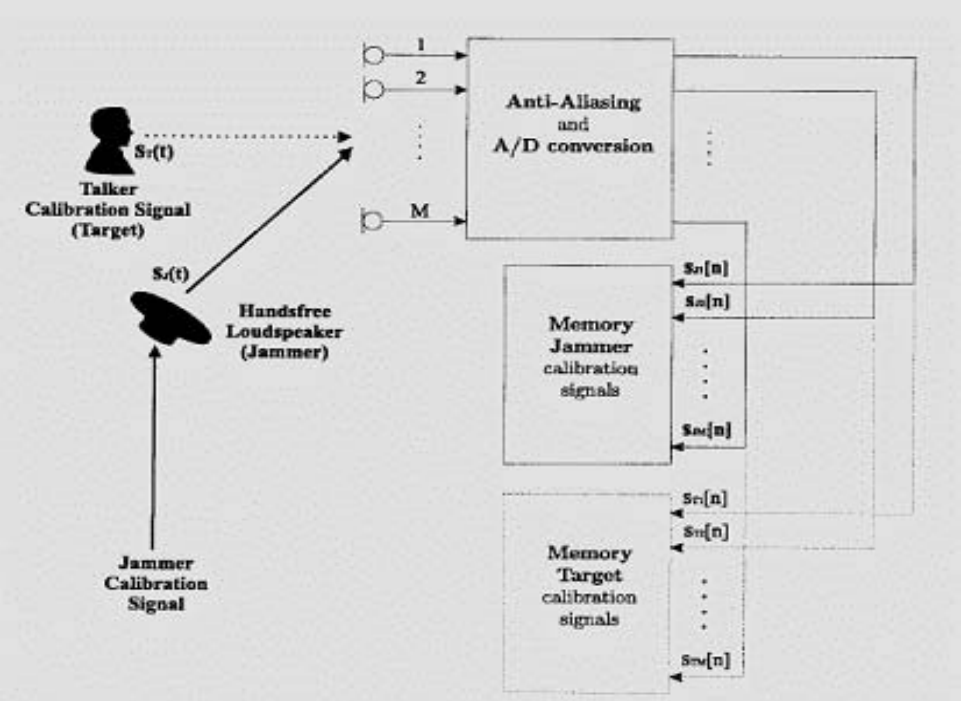

圖 2-10：Dahl's Algorithm 訊號擷取架構圖

境中有人講話聲和喇叭所播放的音樂聲，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鍾人講話的聲音，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播放的音樂聲，這樣則完成Dahl's Algorithm的預錄部分。

而Dahl's Algorithm架構圖如圖 2-11 所示，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而當輸出，下半部分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與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。

在Dahl's Algorithm中，作適應性空間濾波器調適和空間濾波器輸

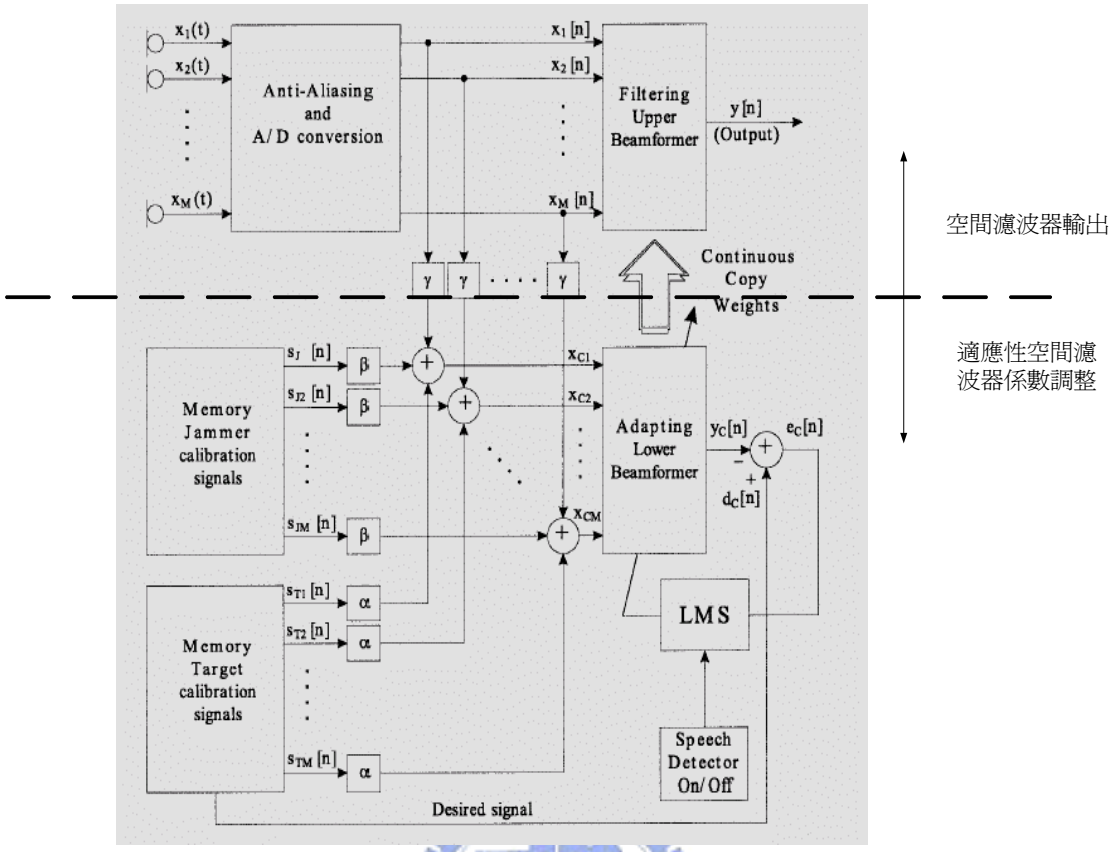

圖 2-11：Dahl's Algorithm 架構圖

出這兩部分不可同時進行，若在作空間濾波器輸出時，干擾源方向改變，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。

2.5 結合真人語音偵測與適應性陣列訊號處理

2.5.1 結合雙層真人語音偵測與適應性陣列訊號處理架構簡介

本章節將介紹結合雙層真人語音偵測與適應性陣列訊號處理架構，其架構圖如圖 2-12 所示：

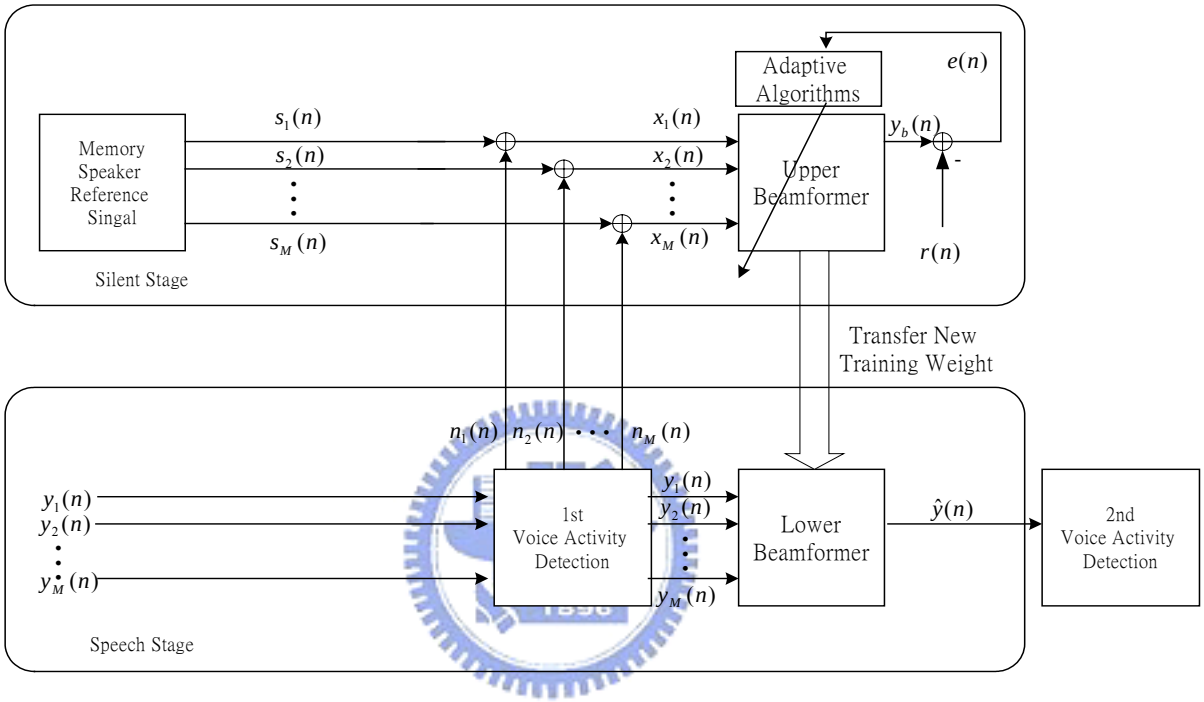

圖 2-12：結合雙層真人語音偵測與適應性陣列訊號處理架構圖

圖 2-12 為結合VAD與Dahl's Algorithm架構圖，此架構圖有兩層VAD，而此架構可由第一層VAD的判定來分為兩部分，第一部分稱為Silent Stage，第二部分稱為Speech Stage。當聲音訊號經過第一層VAD的判定，若為非真人語音，此時系統會進入Silent Stage，透過Normalize LMS的方法適應性調整Upper Beamformer係數。若聲音訊號經過第一層VAD的判定為真人語音，此時系統會進入Speech Stage，Silent Stage中的適應性訊號調整將會被關閉，並將Upper Beamformer係數傳遞給Lower Beamformer，讓真人語音通過空間濾波的處理，並再通過第二層VAD當輸出。

在現實環境中，噪音源聲量若與真人語音能量近似時，則第一層的

VAD 難免會判斷錯誤，因此我們在 Lower Beamformer 加入第二層 VAD，用來彌補當第一層 VAD 判斷錯誤的情形，並且聲音通過 Beamformer 後會提高 SNR，增加第二層 VAD 的準確率。因此第一層 VAD 的主要作用為用來判定是否須做適應性訊號處理的調整，而第二層 VAD 的作用為將非真人語音訊號濾除。

2.5.2 結合雙層真人語音偵測與適應性陣列訊號處理模擬

本章節將展示將聲音訊號於 2.5.1 節所敘述架構中的模擬結果，圖 2-13 為一真人語音與音樂混合之訊號，圖 2-14 為將此訊號通過第一層 VAD 與 Lower Beamformer 的結果，圖 2-15 為將 Lower Beamformer 的輸出再通過第二層 VAD 的結果。

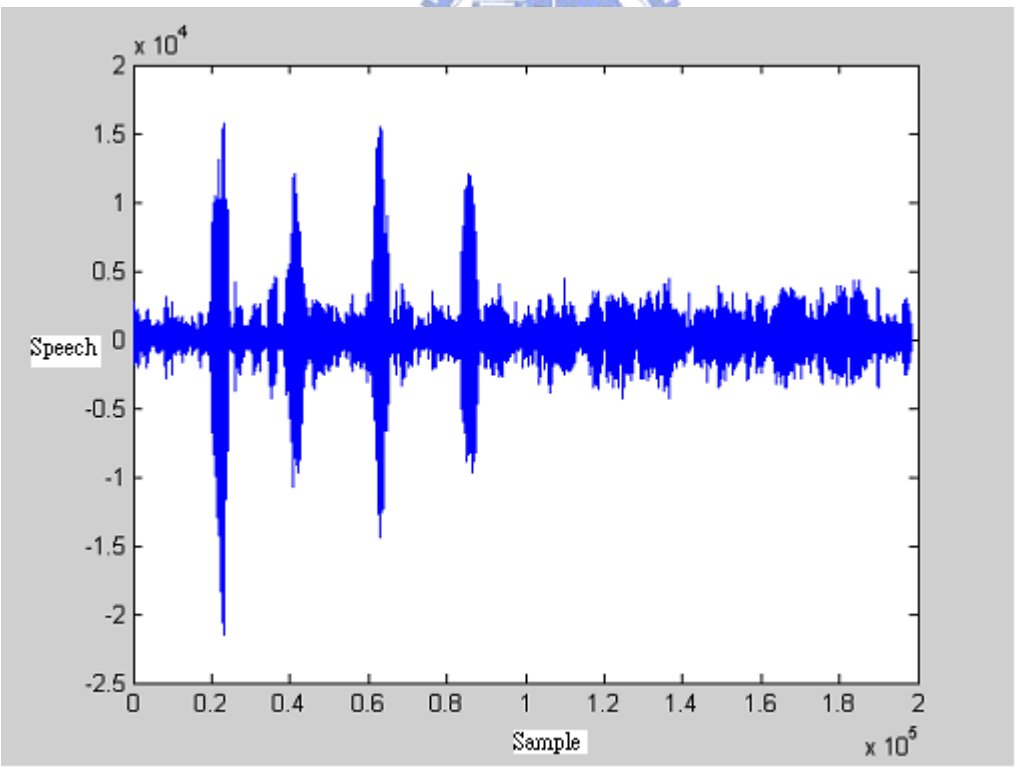

圖 2-13：真人語音與音樂混合之訊號

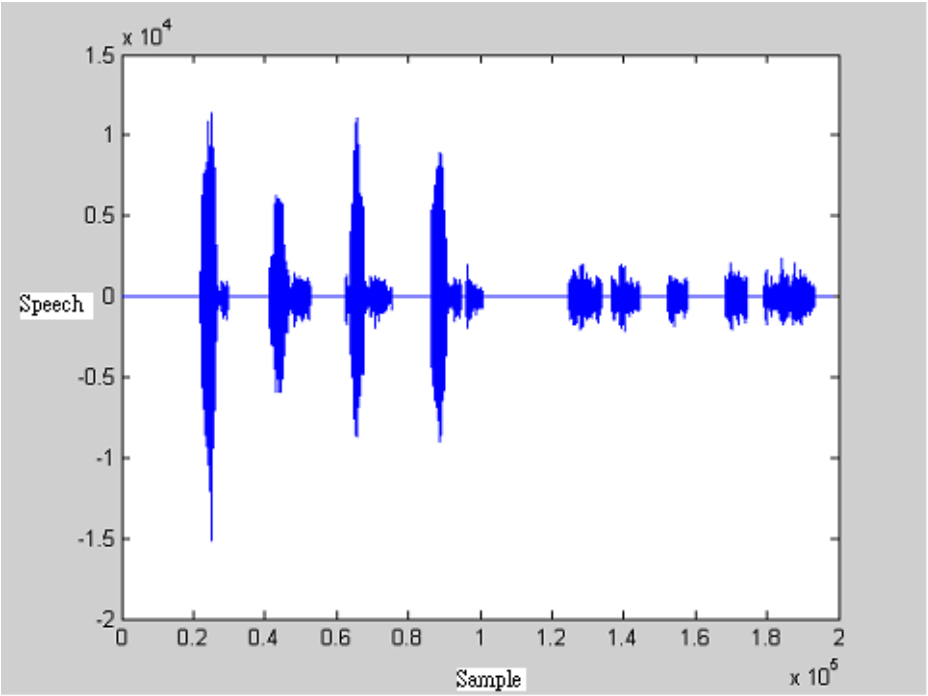

圖 2-14：混合訊號通過第一層 VAD 與 Lower Beamformer 結果

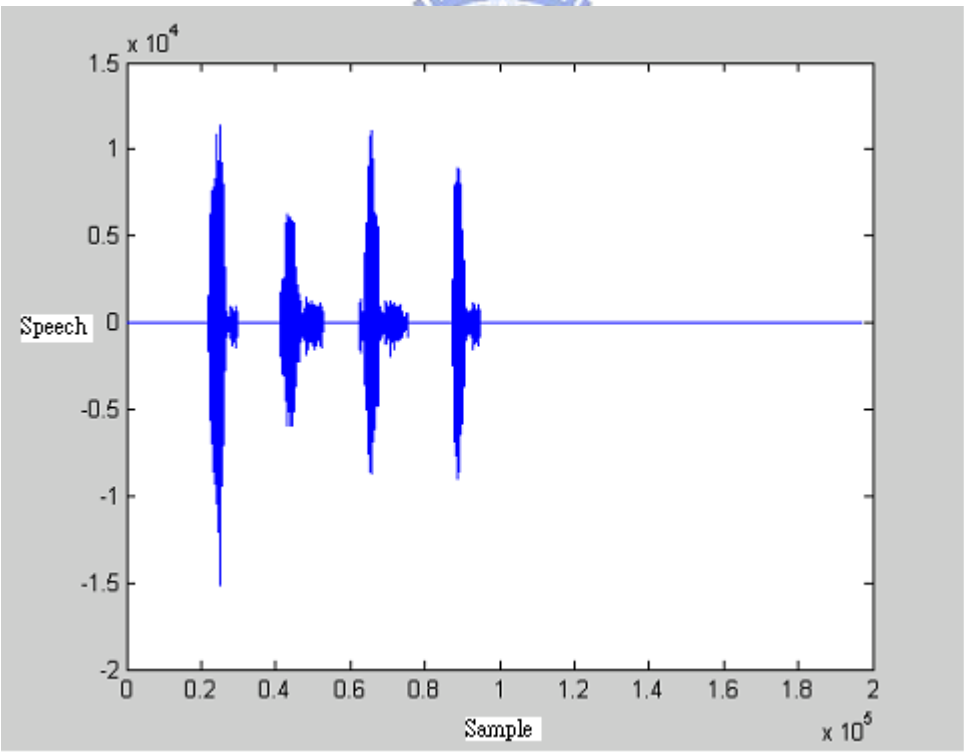

圖 2-15：混合訊號通過第二層 VAD 結果

從圖 2-14 中可發現當真人語音與音樂混合之訊號通過第一層 VAD 與 Lower Beamformer 後，正確的 VAD 判定應只有圖 2-13 中橫軸十萬點前的四個語音，但因為音樂聲在播放時，有起伏大小聲的緣故，因此從圖 2-14 中可看出，第一層 VAD 將一些音樂聲判定為真人語音。為了彌補第一層 VAD 的誤判，Lower Beamformer 後再加入第二層 VAD，因為通過 Lower Beamformer 後的緣故，因此音樂聲會被壓低，而再通過第二層 VAD 就會將音樂聲完全濾除，只輸出真人語音，如圖 2-15 所示。

2.5.3 結合單層真人語音偵測與適應性陣列訊號處理架構簡介

圖 2-12 中的架構，以即時的論點來講，有兩項缺點，分別為：

1、雙層的 VAD 造成計算量過於龐大，容易造成音訊訊號延遲

2、當音樂聲與真人語音能量接近時，易造成第一層 VAD 判斷錯誤，這樣系統會一直將訊號直接通過 Lower Beamformer，系統便不能自動做適應性訊號調整

因此針對上述兩項缺點，本論文提出了結合單層真人語音偵測與適應性陣列訊號處理架構，其架構圖如圖 2-16 所示，在圖 2-16 架構中，VAD 只用於 Lower Beamformer 後，因此麥克風陣列訊號皆會先經過 Lower Beamformer，再通過 VAD 判定，若判定為真人語音，則真人語音訊號會直接輸出，若為非真人語音，系統會將非真人語音的原始訊號（未通過 Lower Beamformer），傳遞給 Upper Beamformer 做適應性訊號調整，調整完畢後再將濾波器係數傳遞給 Lower Beamformer，更新 Lower Beamformer 濾波係數。此架構的好處為：

1、單層 VAD 的計算量相較於 2.5.1 節架構中雙層 VAD 小很多

2、當音樂聲與真人語音能量接近時，音樂聲與真人語音混合訊號經過 Lower Beamformer 後會將音樂聲壓抑，此時再由 VAD 判定是否須

做適應性訊號調整，會比 2.5.1 節架構準確許多。

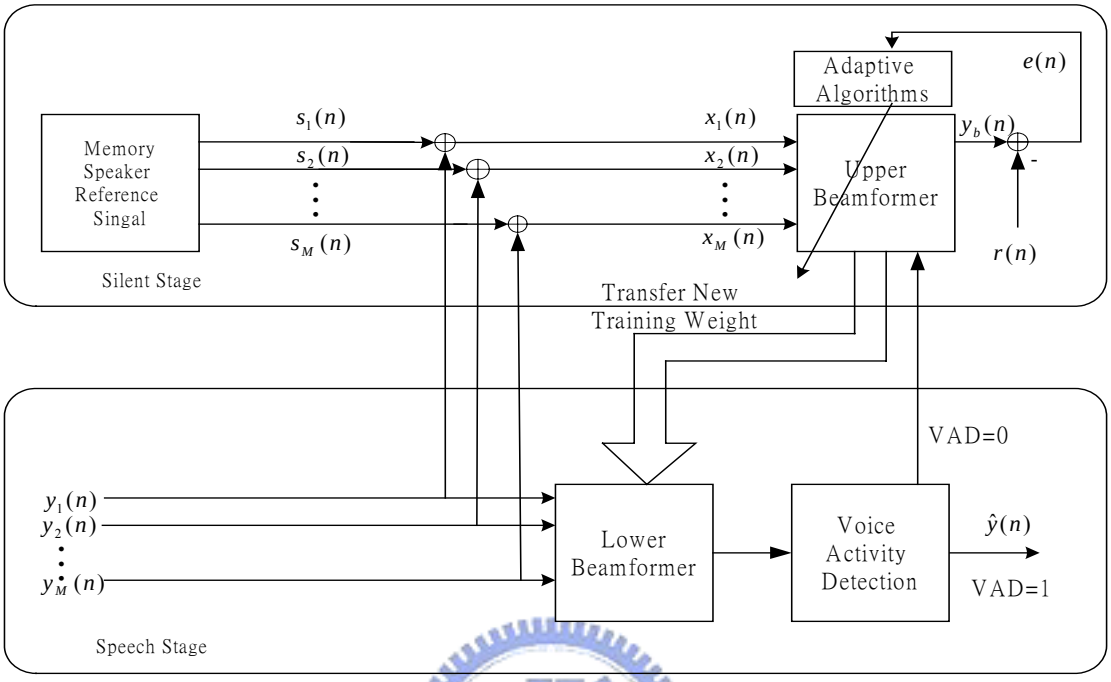

圖 2-16：結合單層真人語音偵測與適應性陣列訊號處理架構圖

2.5.4 結合單層真人語音偵測與適應性陣列訊號處理模擬

本章節將展示將聲音訊號於 2.5.3 節所敘述架構中的模擬結果，為了與 2.5.1 節所述架構做比較，通過 Lower Beamformer 的真人語音與音樂混合之訊號與圖 2-13 一樣，而圖 2-17 為真人語音與音樂混合之訊號通過 2.5.3 節所敘述架構中 Lower Beamformer（濾波器階數=10）的效果，圖 2-18 為通過 VAD 後的效果。圖 2-13 中的混合訊號 SNR 為 11.54dB，圖 2-17 的混合訊號 SNR 為 18.23dB，Lower Beamformer 將混合訊號 SNR 提高了 6.69dB。從圖 2-18 可觀察出，結合單層真人語音偵測與適應性陣列訊號處理可達到與結合雙層真人語音偵測與適應性陣列訊號處理一樣的效果，因此本論文採取 2.5.3 節所述架構，實現於麥克風陣列平台上。

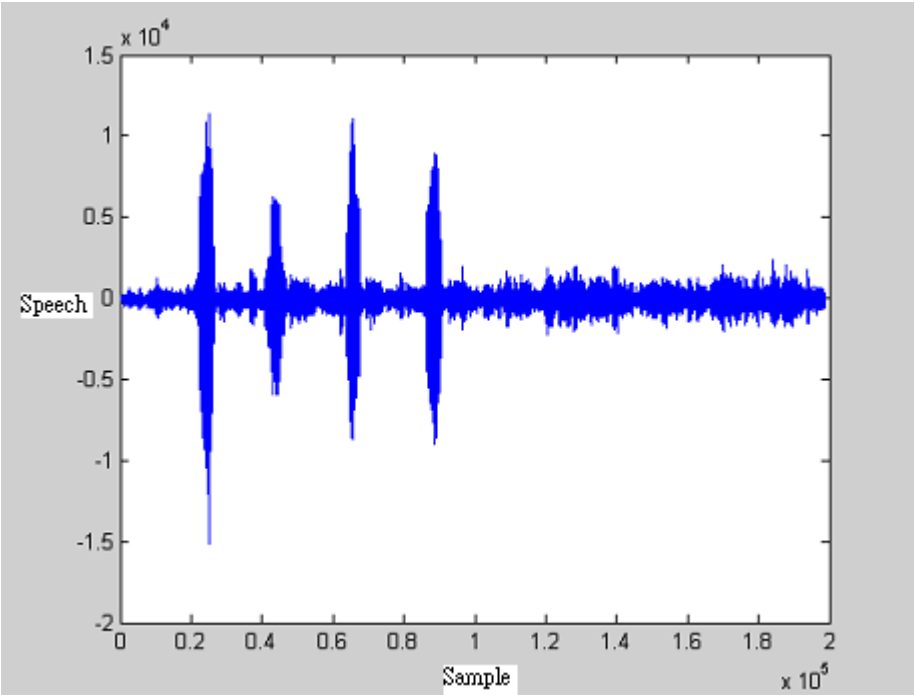

圖 2-17：真人語音與音樂混合之訊號通過 Lower Beamformer 結果

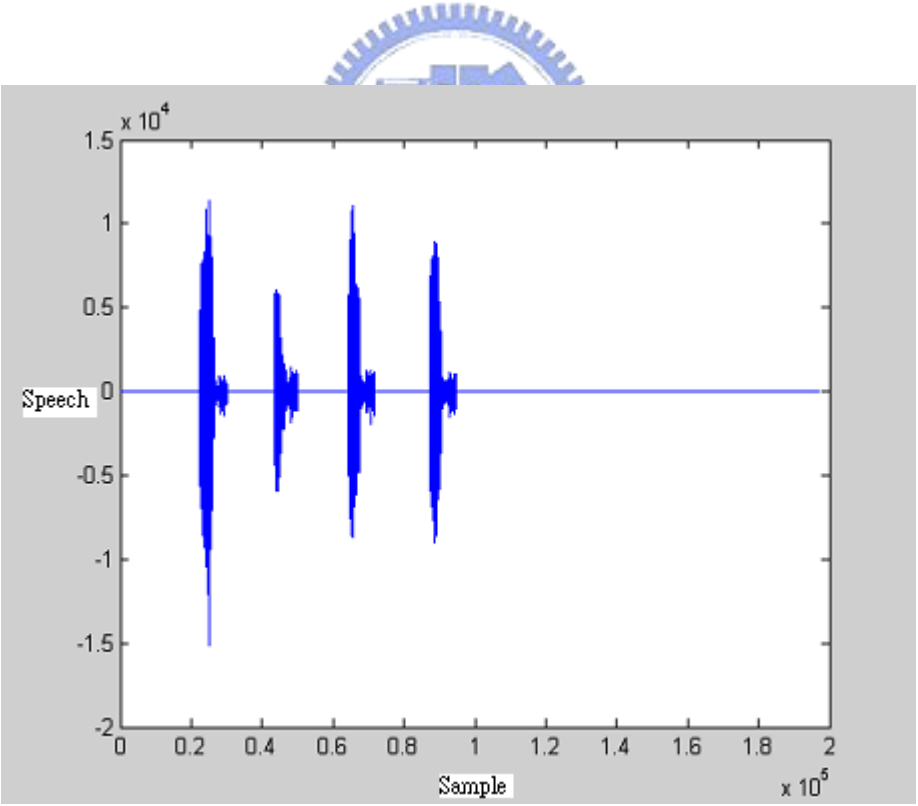

圖 2-18： Lower Beamformer 輸出通過 VAD 結果

第三章 自動語音辨識

3.1 語音辨識簡介[21]

對於一個語音辨識系統而言，可以簡單的將之分為：連續音、非連續音辨識系統;單一使用者、多使用者辨識系統;以及較少樣本數的資料庫、較多樣本數的的資料庫。

語音辨識最主要的目的是希望電腦聽懂人類說話的聲音，進而命令電腦執行相對應的工作。當聲音藉由類比到數位的轉換裝置輸入電腦內部，並以數值方式儲存後，語音辨識程式便開始已事先儲存好的聲音樣本與輸入的測試聲音樣本進行比對工作。比對完成後點腦集輸入一個它認為最“像”的聲音樣本序號，我們就可以知道使用者剛剛唸進去的聲音代表何意，進而命令電腦做事。

3.2 語音辨識系統架構

一般來說，語音辨識系統的架構分為兩個部分：1.語音樣本的訓練 2.語音信號測試。第一個步驟屬 off-line，就是將我們所要識別的語音之參數訓練出來後，儲存在系統中。第二個步驟為 on-line，即為輸入語音後，系統會將語音識別的結果正確地顯示出來。

3.2.1 語音特徵參數求取[22]

在圖 3-1 與圖 3-2 中模型建立前的步驟皆屬語音特徵參數的求取，在訊號辨識中，最常用的特徵參數是訊號在頻譜（Spectrum）上的能量值，這些在頻譜上的能量值便可稱為一種特徵值。然而，對語音訊號而言，另一種稱為倒頻譜的參數卻更能代表語音訊號的特性，而使辨識率提高。圖 3-3 為語音特徵參數求取流程圖，其過程分為六個步驟：

1. 預強調：

   人的口腔就像一個濾波器,會將語音的高頻部分濾除,因此將語音信號通過一高通濾波器來做為補償，其高通濾波器的方程式如（3-1）式所示：

$$H(Z) = 1 - 0.95Z^{-1} \qquad (3\text{-}1)$$

2. 音框化：

   將語音訊號每 30 ms 取一個音框，為防止相鄰音框的特性變化太過於迅速,令相鄰音框之間重疊 20ms

3. Hamming Window：

   使用 Hamming Window 降低音框中起始點與終點信號的不連續性，其 Hamming Window 的公式如（3-2）式所示：

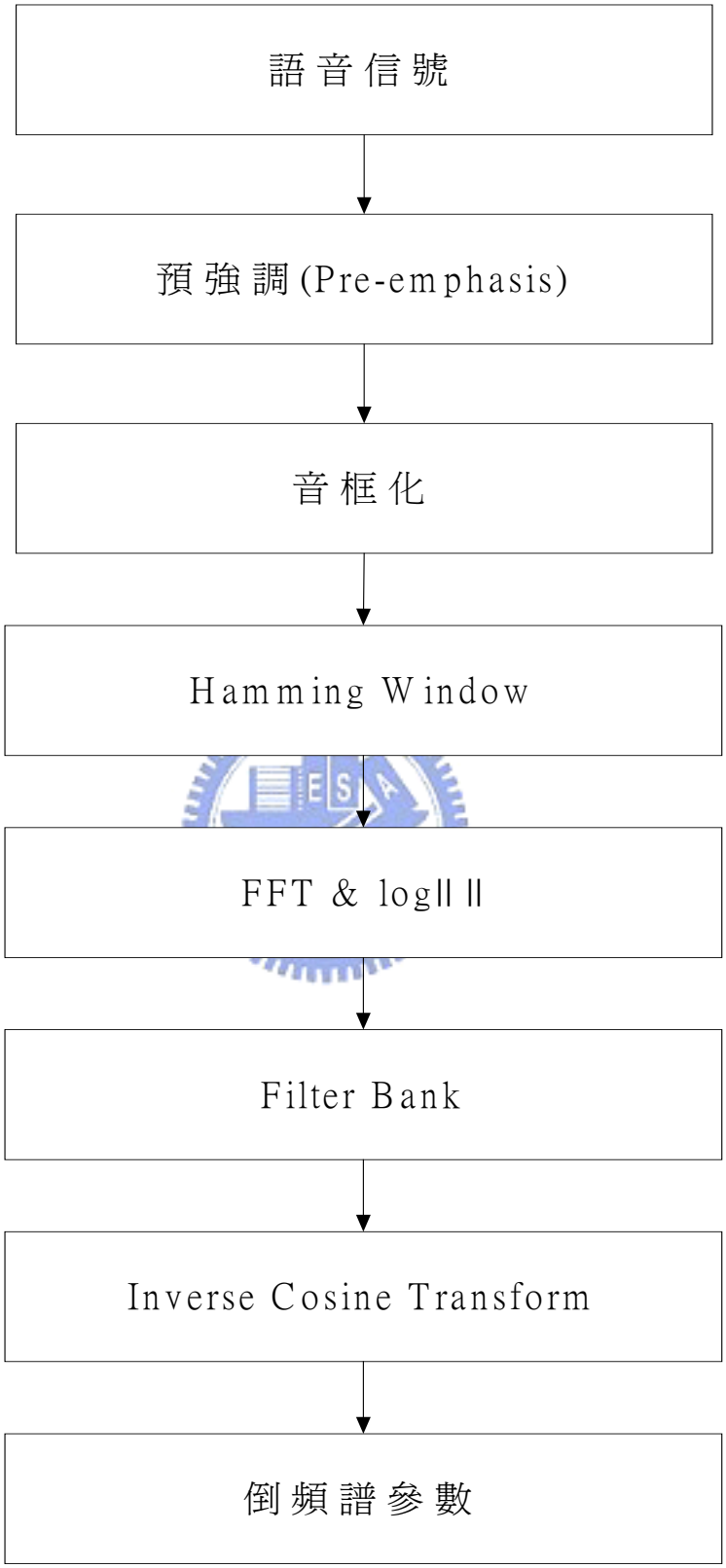

圖 3-1：語音特徵參數求取流程圖

$$w(n) = 0.54 - 0.46 \times \cos\left(\frac{2\pi n}{N-1}\right) \qquad (3\text{-}2)$$

其中的 N 值代表音框大小，其 Hamming Window 圖如圖 3-2 所示。

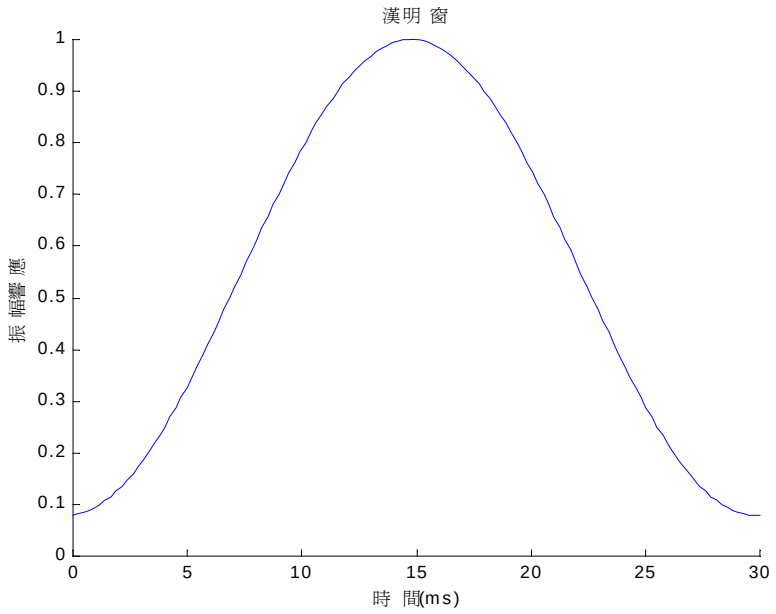

圖 3-2：Hamming Window

4. FFT ：

利用 FFT 將每個音框轉成頻譜。

5. Filter Bank：

Mel-Frequency Cepstrum Coefficients 則為一段音框的特徵代表，其 Filter Bank 的取法如（3-3）式與（3-4）式所示

$$H_m[k]=\begin{cases}0 & \\ \dfrac{2(k-f[m-1])}{(f[m+1]-f[m-1])(f[m]-f[m-1]} & k<f[m-1] \\ \dfrac{2(k-f[m-1])}{(f[m+1]-f[m-1])(f[m]-f[m-1]} & f[m-1]\le k\le f[m] \\ 0 & k>f[m+1]\end{cases} \quad f[m]\le k\le f[m+1]$$

（3-3）

$$f[m]=\left(\frac{N}{F_s}\right)B^{-1}\left(B(f_1)+m\frac{B(f_h)-B(f_l)}{M+1}\right)$$

$$B^{-1}(f)=7000(EXP(B/1125)-1)$$

（3-4）

（3-3）式為 filter bank 定義式，總共有 M 個 filter（m=1,2,…,M），而每個濾波形狀為三角形。而（3-4）式為將 Frequency-Scale 為 Mel-Scale。其中 f_h 和 f_l 代表 filter bank 中最高與最低 Hz，f_s 為取樣頻率，M 為 filter 數，N 為 FFT 點數。其 filter bank 圖如圖 3-3 所示。

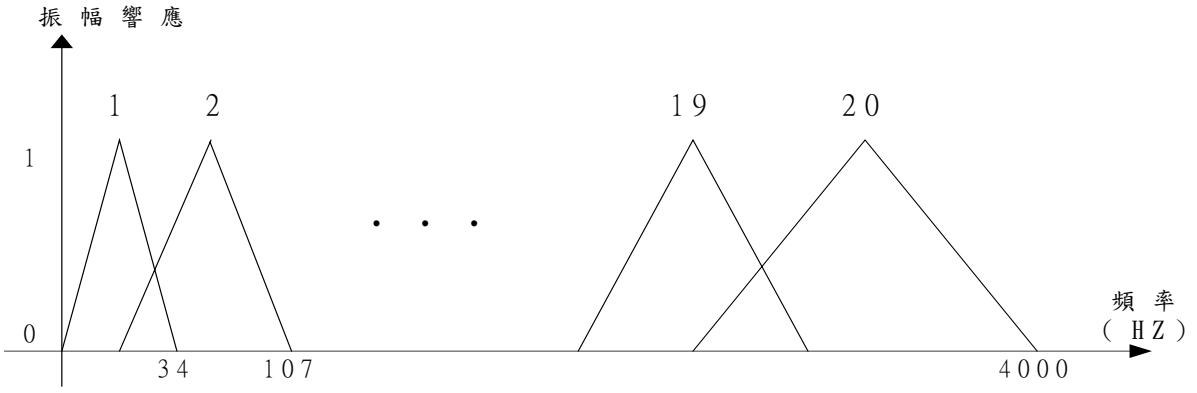

圖 3-3：用於計算 mel-cepstrum 之 filter bank

6. 反餘弦轉換：

將訊號經過反餘弦轉換得到 Mel-Frequency Cepstrum Coefficients（MFCC）。

3.2.2 建立語音辨識模型[22]

當語音特徵參數求取完成後，接著就必須建立語音模型，而目前最普遍使用的模型即是隱藏式馬可夫模型(Hidden Markov Model；HMM)，HMM的目的為以統計的方式來建立每個類別的（動態）機率模型，此種模型特別適用於長度不固定的輸入向量。

在整個 HMM 的型態，一般最常使用 left to right model(亦稱 bakis model)。如圖 3-4 所示，狀態序列只能由左至右，或停留在原處。其中 s1、s2、…、s6 代表狀態。

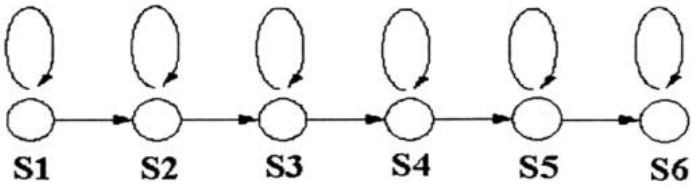

圖 3-4：HMM 狀態圖

而 HMM 相關參數定義如下：

(1)狀態數(states)

(2)狀態轉移機率a，即為從狀態i轉移至狀態j的機率值，$a_{ij}=P[q_{t+1}=j \mid q_t=i]$，其中 $1 \leqq i, j \leqq N$。

(3)狀態觀測機率b，某個狀態中，出現某個觀測值的機率值，$b_i(v_k)=P[o_t=v_k \mid q_t=i]$，其中 $1 \leqq k \leqq M$；$1 \leqq i \leqq N$。

(4)初始狀態π_i，初始狀態的機率$\pi_i=P[q_1=i]$，$1 \leqq i \leqq N$

HMM 相關參數中的 a,b 將會以矩陣形式來表示，而當語音模型建立完成後，當有辨識資料輸入時則利用 Viterbi Algorithm [22]的方法，計算輸入語句和每個模型的相似度，機率最高者，即為辨識結果。

3.3 新竹科學園區廠商名稱語音辨識器[28]

本論文將麥克風陣列與語音辨識器做結合，而語音辨識器為交通大學電信研究所，所開發的 API 介面軟體，其使用者介面如圖 3-5 所示，當啟動辨識後，語音辨識器會透過麥克風接收語音信號並將其波形顯示出來，最後將顯示最有可能的辨識結果。

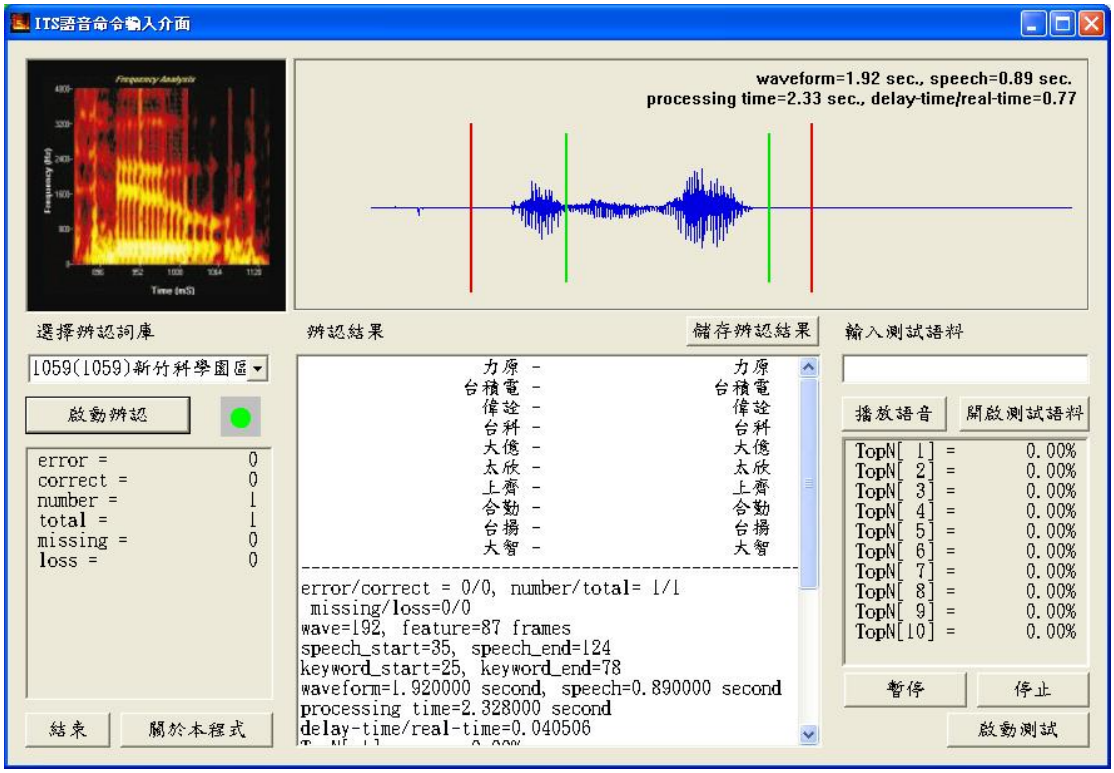

圖 3-5：語音辨識器使用者介面

3.4 IBM ViaVoice[23]

論文中，麥克風陣列也與 IBM ViaVoice 做結合，ViaVoice 為 IBM 所開發的語音軟體，其最大的功用為用語音來聽寫文字並控制電腦。其產品功能如下：

- 連續聽寫

  可建立、編輯及修改文件並支援大多數 Widows 應用程式，包括 Microsoft Word 等。可讓 Via Voice 分析已輸入的文章，讓電腦了解使用者的習慣用語。

- 語音命令

  可使用聲音來告訴系統要執行的動作，啟動程式、切換程式、視窗移動、最大化、最小化、更正聽寫錯誤，甚至是控制滑鼠的動作(語音滑鼠)。

- 語音合成

  Via Voice 可以朗讀電腦內的中英文字，如:文字檔、E-mail、網頁的內容等，同時還可以讓選擇不同的腔調如男生與女生聲音等。

- 網上瀏覽

  Via Voice 可以用語音來瀏覽網頁，只要唸出想要選的超文字連結(Hyper-link)，就可漫遊網際網路，並可用語音來控制瀏覽器，例如首頁、我的最愛、重新整理等等。

而麥克風陣列與 Via Voice 結合最大的用意即在吵雜的環境下也能用語音來控制電腦，例如，使用者在用喇叭放音樂時，但使用者依然能用語音來瀏覽網路、撰寫 Word 等等。

第四章 軟硬體設計與實現

4.1 實驗平台架構

本章節將介紹實現第二章理論的實驗平台，平台架構圖如圖 4-1 所示

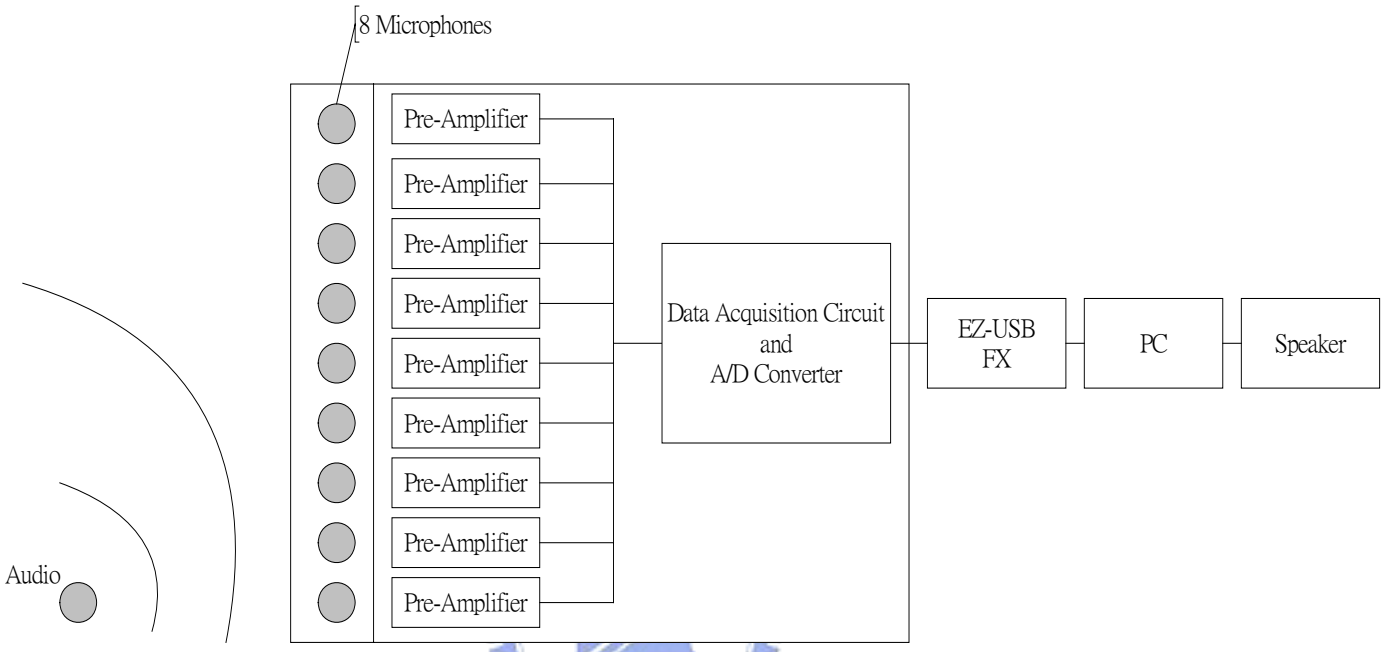

圖 4-1：語音純化系統架構圖

首先，聲音訊號經由 8 通道麥克風陣列所擷取，接著通過訊號放大電路與濾波器並將聲音訊號透過 A/D（Analog to Digital）轉換器轉換為數位形式，數位形式的資料經由 EZ-USB FX 平台以 USB 為傳輸介面傳到 PC 端，最後資料在 PC 端上作演算法處理，並即時由喇叭播放出演算法處理完的資料。

實驗平台可分為四部分來介紹：

1. 聲音訊號放大及濾波電路
2. 類比訊號擷取及轉換電路
3. USB 傳輸裝置
4. PC 端演算法處理

4.2 聲音訊號放大及濾波電路

聲音訊號放大及濾波電路的目的是用來放大麥克風所收到的訊號並濾掉高頻及低頻的雜訊，每顆麥克風都有各自的放大及濾波電路，8 組放大及濾波電路構造及功能皆相同，放大及濾波電路架構圖如圖 4-2 所示：

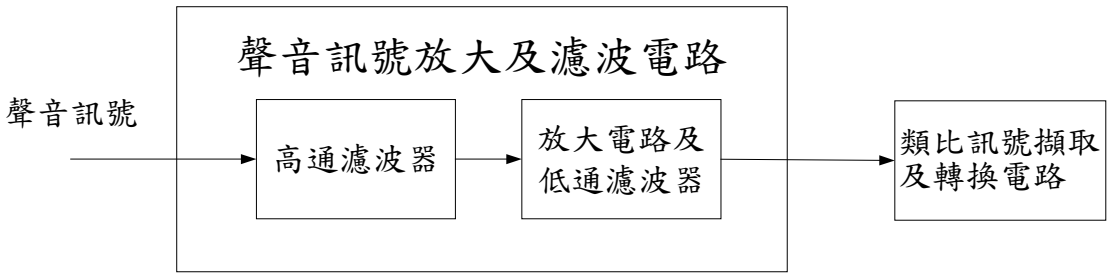

圖 4-2：聲音訊號放大及濾波電路架構圖

聲音訊號經由電容式麥克風轉為電壓訊號後，必須先經過高通濾波器，以濾掉低頻雜訊及直流訊號，而高通濾波器的 3 dB 點設於 80Hz 的地方。訊號經過高通濾波器後還是一個非常小的電壓，因此必須經過一放大電路來放大電壓訊號，以供後端的 A./D 來取樣，而本電路的取樣頻率為 16 k Hz，所以必須將訊號通過低通濾波器來避免 Aliasing 問題，而電路中低通濾波器的 3 dB 點設定於 6 k Hz。聲音訊號放大及濾波電路圖展示於圖 4-3，本放大電路為一兩級的 OP 放大器，工作電壓介於 5V 和-5V 之間，並採用負回授的形式，圖 4-3 的放大倍率為 60dB。而電路的頻率響應圖由 P-SPICE 所模擬如圖 4-4 所示。

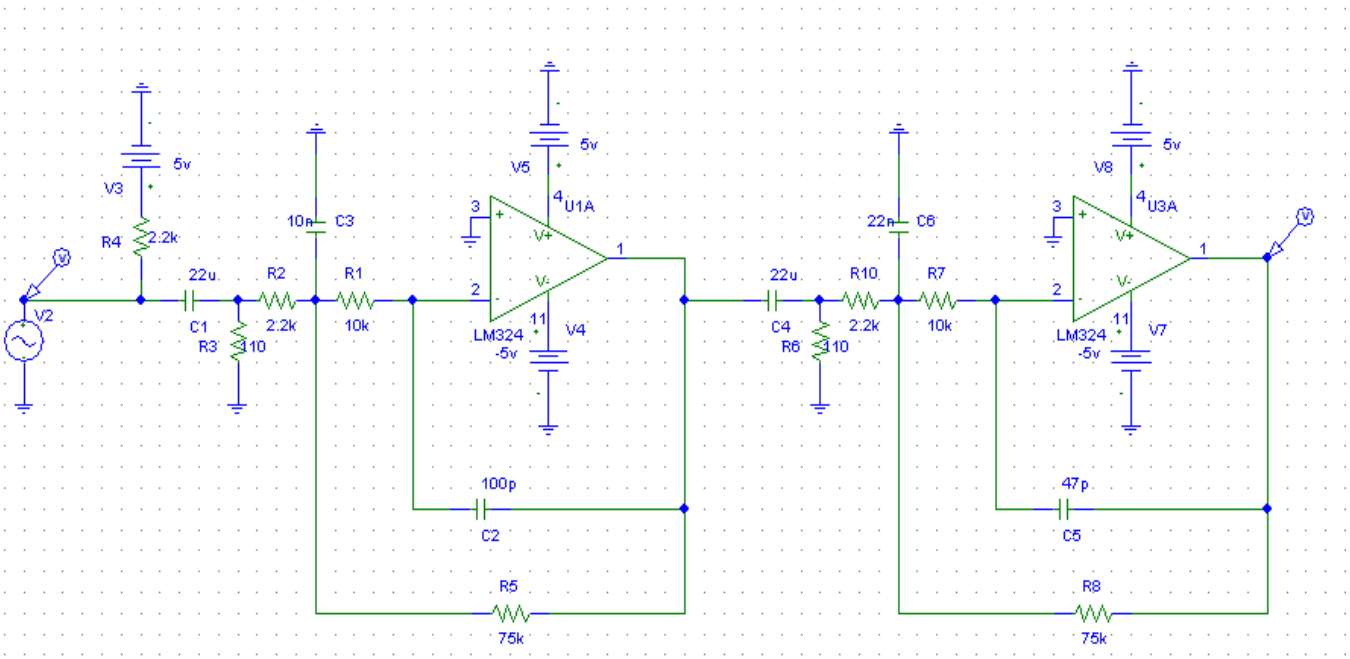

圖 4-3：聲音訊號放大及濾波電路圖

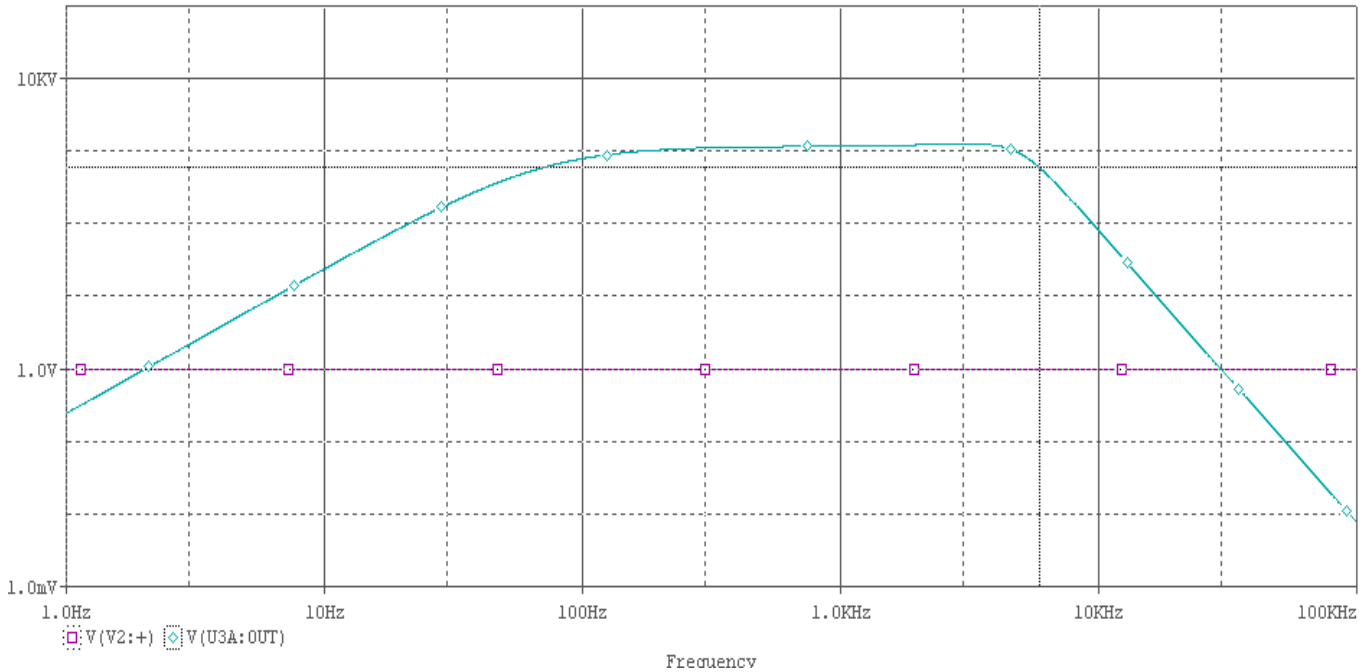

圖 4-4：聲音訊號放大及濾波電路之頻率響應圖

4.3 類比訊號擷取及轉換電路

類比訊號擷取及轉換電路的目的是將放大倍率後的類比聲音訊號，轉換為數位訊號（16 位元），電路架構圖如 4-5 所示，為了節省功率消耗，電路只用了一個 A/D 轉換器，因此，8 通道的類比聲音訊號和 A/D 轉換器之間需要一個切換器，將 8 通道的類比聲音訊號輪流切給 A/D 做轉換。

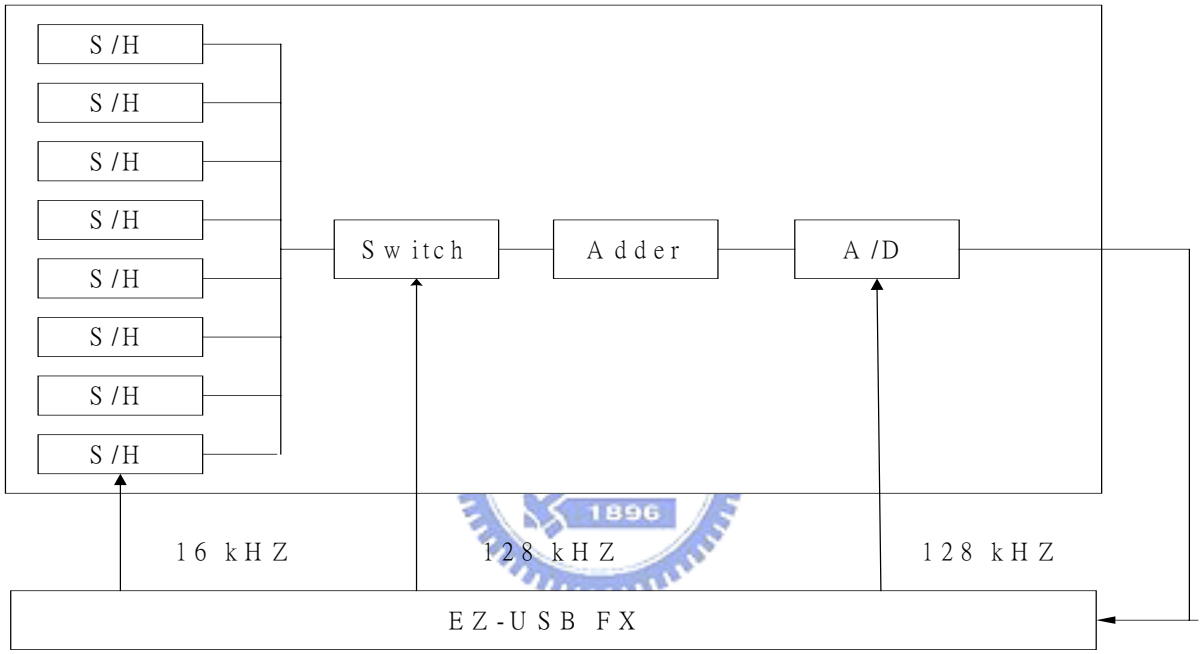

圖 4-5：類比訊號擷取及轉換電路架構圖（S/H：Sample and Hold）

其中 A/D 轉換器只能供給正電壓，但 S/H 輸出會含有負電壓，因此 Switch 和 A/D 之間必須加一個加法器，讓 A/D 的輸入皆在 0 伏以上。

而 S/H、Switch 和 A/D 的時序則由 EZ-USB FX 平台所控制，本系統的取樣頻率為 16 kHz，因此 S/H 的工作頻率為 16 kHz，因為總共有 8 個通道，所以 Switch 和 A/D 的工作頻率為 16x8 kHz，最後 A/D 所轉換出來的 16 位元資料，由 EZ-USB FX 平台所接收。

4.4 系統電路板

電路實作設計時，將聲音訊號放大及濾波電路和類比訊號擷取及轉換電路結合在一起，用 Protel 軟體佈局出其電路圖，其印刷電路板實際照片如圖 4-6 所示：

圖 4-6：麥克風訊號濾波器與數位/類比轉換電路板

此印刷板電路為一四層板架構，長x寬為 27 公分x10 公分，工作電壓為 5 伏特，其麥克風放大倍率和 A/D 取樣範圍皆為可調。

4.5 EZ-USB FX 平台[24]

EZ-USB FX 平台是由 Cypress 半導體公司所推出，將 EZ-USB 晶片與 USB 週邊介面所需的各種功能包裝成一個精簡的整合電路，其微處理機是一個增強的 8051 核心。而 EZ-USB FX 平台在整個系統中有三項目的：

1. 控制 S/H、Switch 和 A/D 的時序
2. 接收 A/D 轉換器的數位資料輸出
3. 將數位資料傳送給 PC 端

架構圖如圖 4-7 所示

4.5.1 控制 S/H、Switch 和 A/D 時序

S/H、Switch 和 A/D 的控制時序由圖 4-7 中 8051 的 PORT A 來傳達控制指令，控制時序由 8051 等時中斷來完成，將中斷時間設為 $\frac{1}{16k}$ 秒，每次中斷發生後，就同時啟動 8 通道的 S/H，使 8 組 S/H 在同一時間完成取樣的動作，接著控制 Switch 將 8 通道 S/H 的輸出輪流切給 A/D 做轉換，換句話說，要再下一次等時中斷發生時，將 8 通道的 A/D 全處理完成。

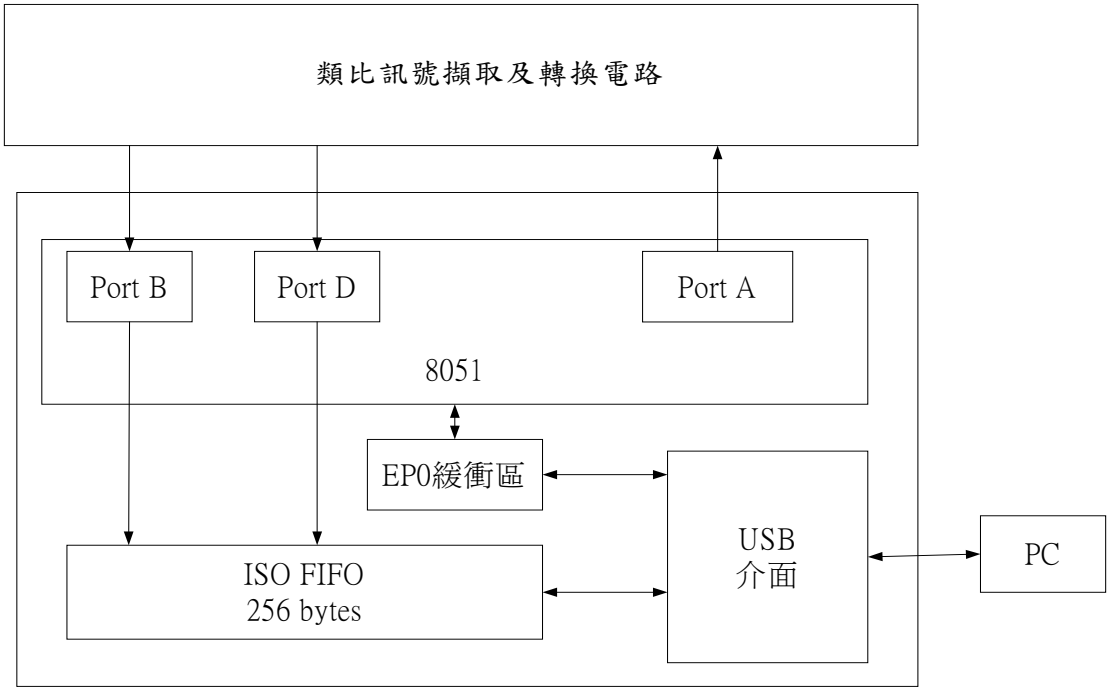

圖 4-7：EZ-USB FX 平台架構圖

4.5.2 接收 A/D 轉換器數位資料輸出

A/D 轉換器的輸出資料由 8051 的 Port B 和 Port D 所接收，A/D 輸出為 16 位元，而 Port B 和 Port D 各接收 8 位元，每次計時中斷發生時，Port B 和 Port D 會將其所接收到的資料存於等時傳輸 FIFO 中，EZ-USB FX 的等時傳輸 FIFO 總共有四個，每個容量為 256 bytes，在系統中，我們只用了一個等時傳輸 FIFO，也就是每 1ms，等時傳輸 FIFO 會被填滿。

4.5.3 USB 傳輸[25]

等時傳輸 FIFO 會由等時中斷的方式,每 1ms 透過 USB 傳輸給 PC 端,其 USB 提供了四種不同的傳輸模式：

1、 巨量傳輸(Bulk)：突發性的傳輸模式。資料封包大小為 8、16、32、64 位元組。除了資料封包之外另有交握封包(Hand-Shake Package)，及自動錯誤資料檢核機制(CRC)，如資料傳送錯誤，可要求裝置重送封包，確保資料的正確性。

2、 中斷傳輸(Interrupt)：類似巨量傳輸，資料封包大小為 1~64 位元組。高速的裝置中。需經由主機規則固定間隔詢問。

3、 等時傳輸(Isochronous)：在固定的時間傳出封包，主要使用在音頻與影像等資料流中。為了確保封包可以在固定的時間送出，無 Hand-Shaking 封包，僅具有 CRC 錯誤檢核，資料傳輸錯誤亦不再重送封包。時間是最重要的要求條件。

4、 控制傳輸(Control)：用來配置及傳送命令給裝置，確認裝置要求。

而在實驗平臺的架設中，只用到控制傳輸及等時傳輸。

系統中，USB 傳輸模式我們選擇等時傳輸的方式，因此每 1 ms 等時中斷發生，等時傳輸 FIFO 中的資料將透過 USB 傳輸給主機。表 4-1 為四種傳輸模式的比較。

	封包大小(Byte)	時間	資料檢查	應用
巨量傳輸	8、16、32、64	盡快完成	有	儲存裝置、印表機
中斷傳輸	1～64	1～255ms	有	滑鼠、鍵盤
等時傳輸	1～1024	1ms	沒有	影像、聲音
控制傳輸		盡快完成	有	命令

表 4-1. USB 四種傳輸模式比較

在 8051 中，系統啟動了兩種中斷模式，一種為 4.5.1 所介紹的計時中斷，另一種為本節所介紹的等時中斷，計時中斷的時間為 $\frac{1}{16k}$ s，而等時中斷的時間為 1ms。因此一次完整的等時傳輸資料由 1 次等時中斷和 15 次計時中斷所構成，其中斷同步說明圖如圖 4-8 所示

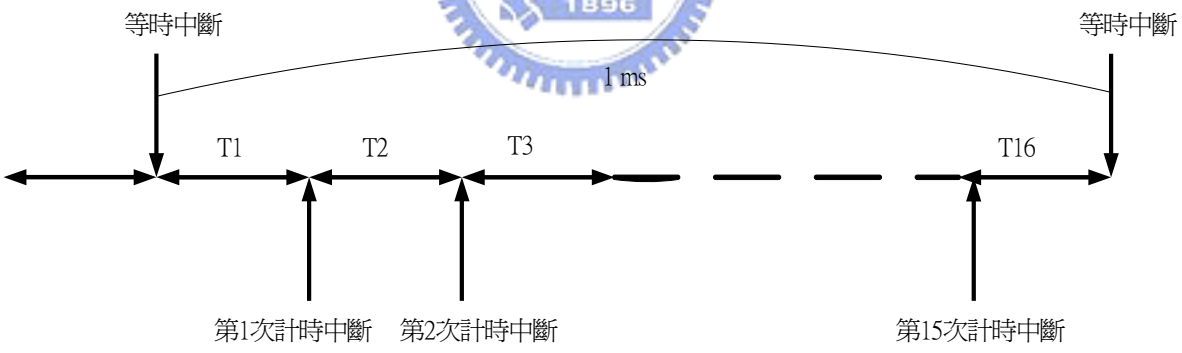

圖 4-8：USB 韌體中計時中斷與等時中斷同步說明

從圖中可看出，每 1ms 會發生 15 次計時中斷和 1 次等時中斷，T2－T15 為計時中斷的間隔，差不多為 $\frac{1}{16k}$ s，第 15 次計時中斷發生後的 T16 s，會發生等時中斷，為了確保中斷的同步，T1 會由等時中斷的副程式來校正。而 USB 裝置韌體流程圖如圖 4-9 所示。

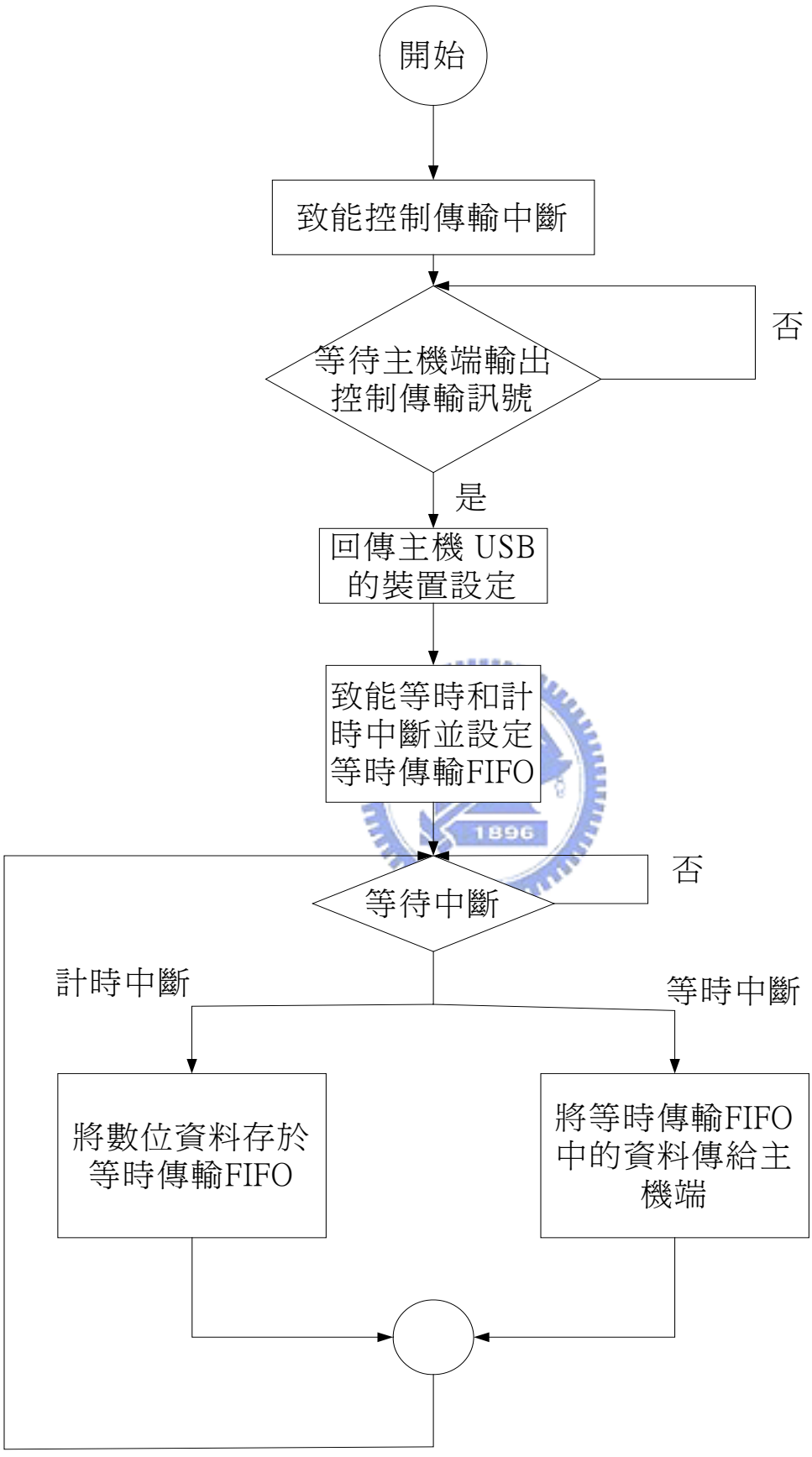

圖 4-9：USB 裝置韌體流程圖

4.6 主機端程式設計

4.6.1 整體架構

主機端程式目的是用來處理語音訊號，使其經過VAD和Dahl's Algorithm的處理，並即時性的用喇叭播出處理結果，軟體架構圖如 2.5.3 節所述。

而主機端軟體編譯環境為Microsoft Visual C++，並將使用者介面寫成視窗形式以方便使用者操作，而程式中VAD和Dahl's Algorithm的結合是使用multithread方式來處理。在實做上，因為考慮到有限的資源達到最好的效能，因此架構採取 2.5.3 節所述架構，音訊信號經由麥克風陣列擷取後會直接通過Lower Beamformer，並再通過VAD的判定，若為真人語音，訊號會直接由喇叭播出。若為非真人語音，系統會將非真人語音的原始訊號（未通過Lower Beamformer）做適應性的調整，並將調整後的係數傳遞給 Lower Beamformer。其流程圖如圖 4-10 所示。而VAD判定為非真人語音有兩種情況，第一為真人語音字尾後，第二為音訊能量小於某個臨界值，而 Lower Beamformer率波係數則是在第二種情況下會去更新。

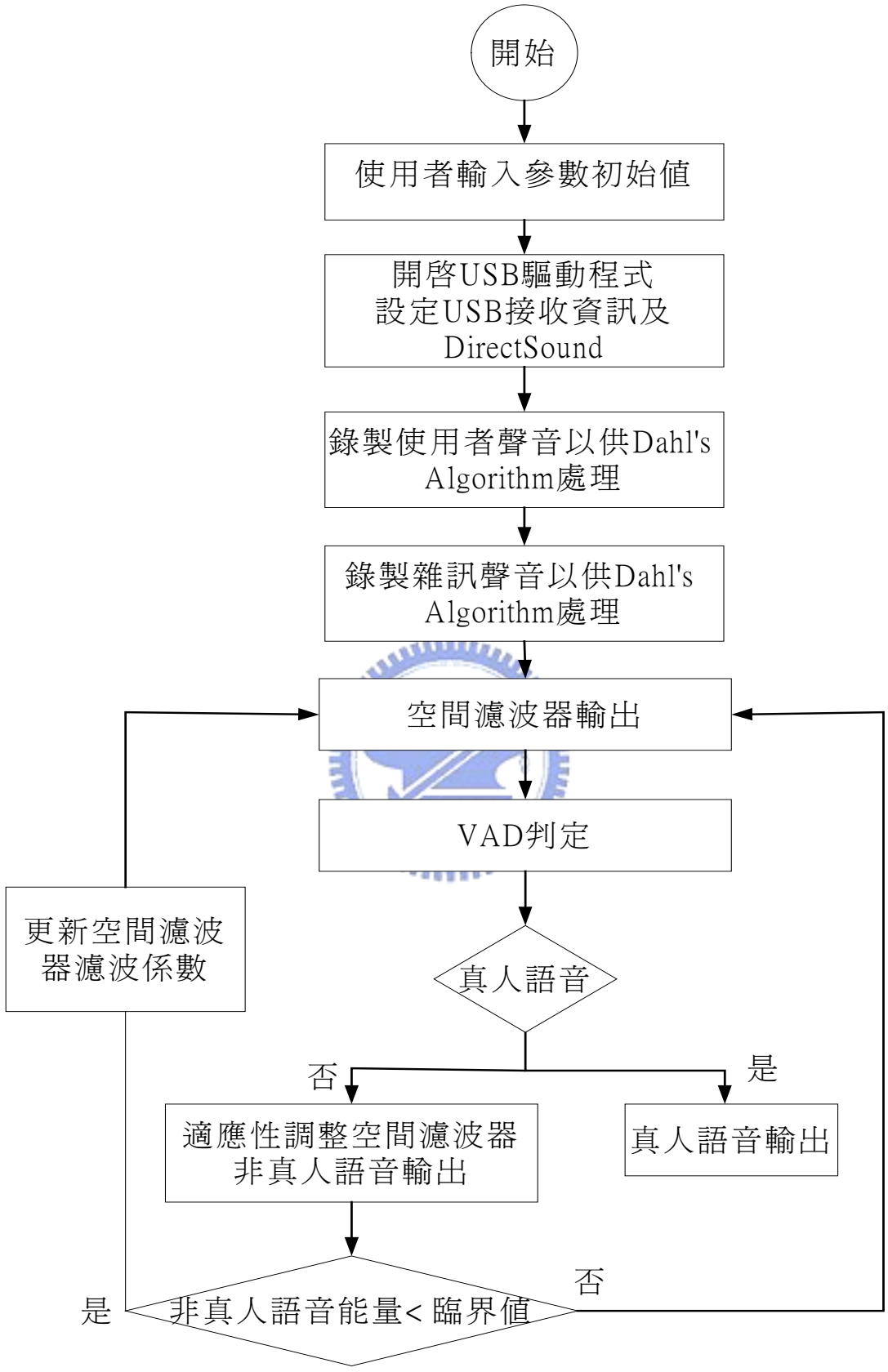
開始
使用者輸入參數初始值
開啓USB驅動程式
設定USB接收資訊及
DirectSound
錄製使用者聲音以供Dahl's
Algorithm處理
錄製雜訊聲音以供Dahl's
Algorithm處理
空間濾波器輸出
VAD判定
更新空間濾波
器濾波係數
真人語音
否
是
適應性調整空間濾波器
非真人語音輸出
真人語音輸出
是
非真人語音能量< 臨界值
否

圖 4-10：主機端軟體流程圖

4.6.2 Intel Math Kernel Library（MKL）[26]

Intel Math Kernel Library（MKL）為一套專門對 Intel CPU 平台系列作數學式子最佳化加速動作的函式庫，如果電腦有多顆處理器，則可以啟動 OpenMP 平行計算如此更可加快運算速度。其 MKL 可用於 C 語言和 Fortran 介面，並包含包括線性代數運算、FFT 等最佳化的函式。

在主機端程式中，VAD 或高階 LMS 演算法皆需要龐大的計算量，為了使系統達到即時的效果，主機端程式會加入 MKL 所提供的函式，圖 4-11 為作 Normalize LMS 演算法加入 MKL 函式和沒加入 MKL 函式所花的時間比較圖。

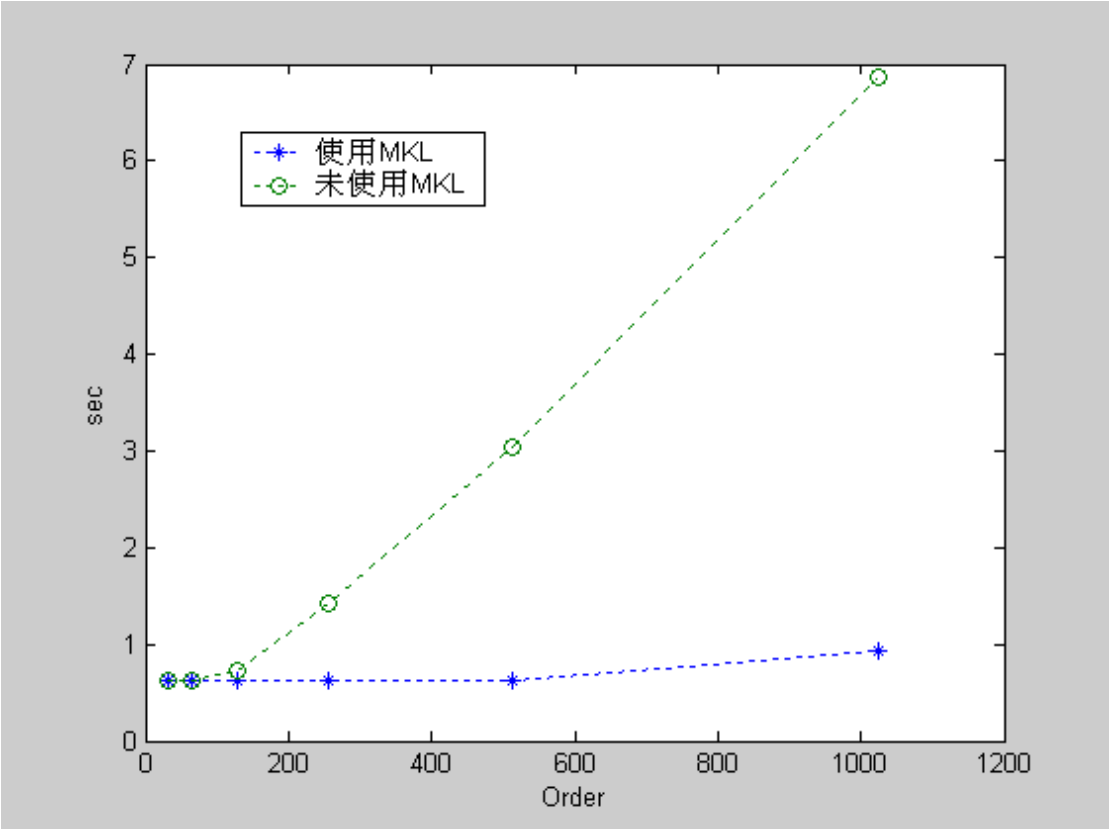

圖 4-11：計算 Normalize LMS 使用與未使用 MKL 比較圖

圖 4-11 中顯示了作 10000 次的 Normalize LMS，濾波器階數與所花時間的比較圖，橫軸為 Normalize LMS 階數，縱軸為所花時間，從圖中可發現當

階數在 100 階時使用 MKL 和未使用 MKL 所花時間差距不大,但當階數愈高時差距就愈大，尤其當 1024 階時，所花時間就差了 5.936 s，因此若以及時運算為目標，那麼使用 MKL 效果會好於未使用 MKL。

4.6.3 Direct X[27]

Microsoft DirectX 是提供給遊戲開發者一個應用程式選項，遊戲開發者可選擇一個強大並具有完整說明的平台，並在此平台建立一個遊戲或多媒體軟體。而為了讓空間濾波器處理完的資料能即時性地播出，我們選擇了 DirectSound 這項介面來達到即時聲音播出的功能。DirectSound 是 DirectX 中，提供 Wave 類型的音效處理介面，音效處理的觀念如下：

1、 裝置物件(Device Object)：DirectSound 裝置物件是在程式中呼叫 DirectX API 的函數來建立，它是一個代表程式所使用的音效輸出入裝置的物件，在物件裝置之後才能進一步設定裝置的協調層級，建立主緩衝區與次緩衝區。

2、 協調層級(Cooperative Level)：Windows 是一個多工作業系統，在同一時間可能會多程式用到相同的硬體資源，因此當音效裝置物件建立之後，還要設定程式對於裝置的使用權限，也就是協調層級。

3、 主緩衝區(Primary Buffer)：主緩衝區是用來儲存要播放聲音資料的記憶體區域。當裝置物件建立時，主緩衝區便會自動產生。

4、 次緩衝區(Secondary Buffer)：次緩衝區是儲存聲音資料記憶體區域，可以有很多個，其中的聲音資料可被放置到主緩衝區來播放。依播放格式設定次緩衝區的主要參數，如取樣頻率、每筆資料大小、一秒鐘的資料量。

當播放命令啟動後，必須先將聲音資料複製到次緩衝區，而聲音資料會不斷地由次緩衝區複製到主緩衝區，需要同步進行資料的傳遞與緩衝區中現

行播放位置的通知，如此載入資料到次緩衝區後，即可馬上被複製到主緩衝區，將聲音播放出來。而要將聲音載入到次緩衝區中，必須進行三個動作：鎖定次緩衝區、載入資料到次緩衝區、解除鎖定次緩區。在此過程中鎖定次緩衝區的用意，是為了防止資料尚未複製到次緩衝區之前的任何不當的存取動作。圖 4-12 為聲音播放的架構圖

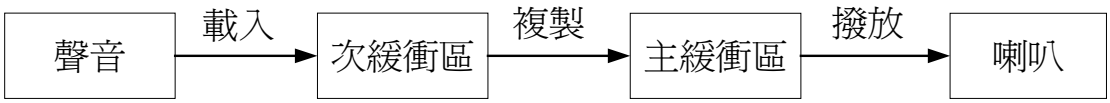

圖 4-12：聲音播放架構圖

圖 4-12 中的緩衝區為一循環緩衝區，圖 4-13 顯示出順著箭頭播放與寫入位置的循環緩衝區觀念。只要把資料寫在亮區而非暗區的話，無論寫在哪都是安全的，但在實作上，應讓資料在現行寫入位置到達之前寫入，這是為了容許資料緩衝區所產生的所有延遲。

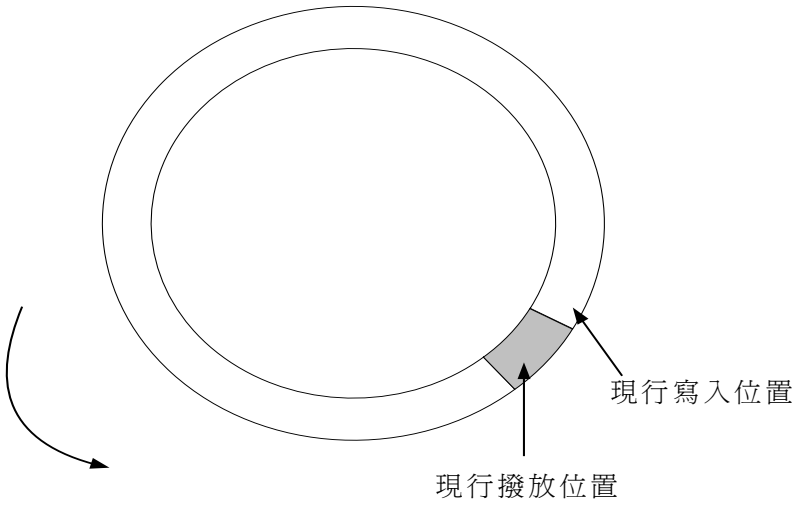

圖 4-13：循環緩衝區之現行寫入與播放位置

4.7 實驗平台實際照片

圖 4-14 為實驗平台實際照片，包含了：

1. 麥克風陣列（間距為 6cm）
2. 系統電路板
3. EZ-USB FX 平台
4. PC

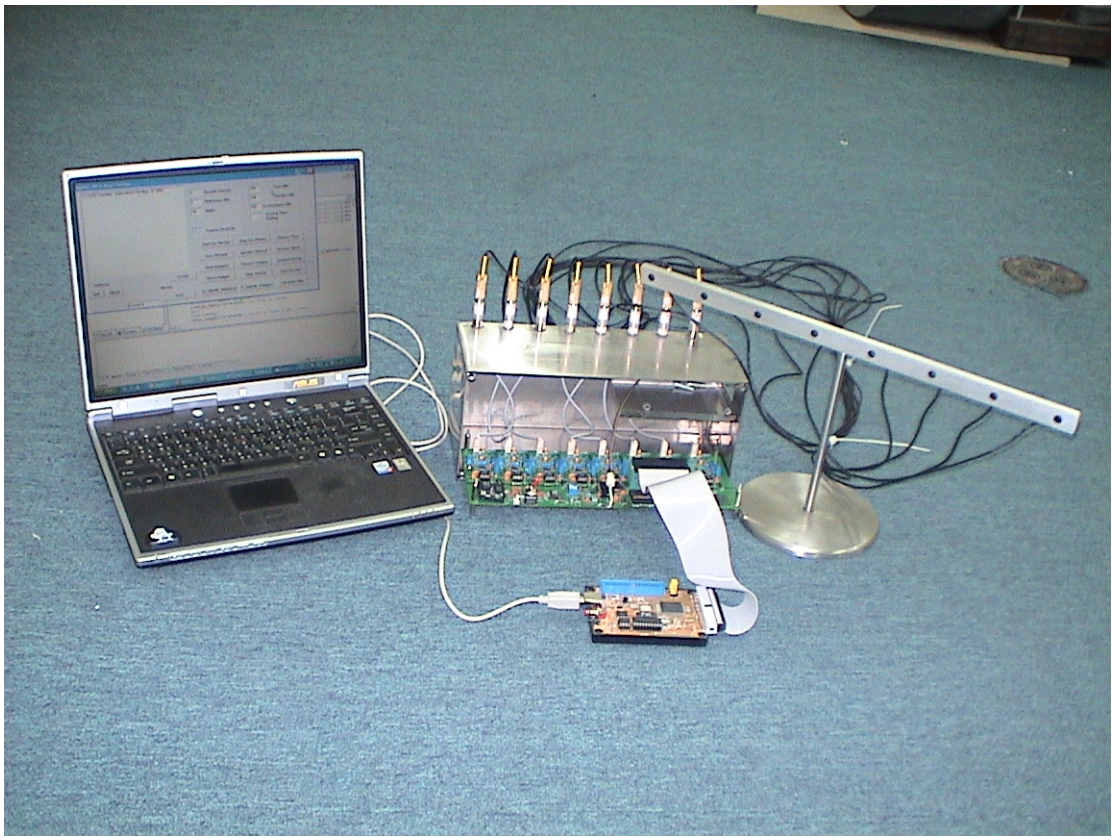

圖 4-14：實驗平台實際照片

第五章 實驗結果與分析

本章節將介紹將麥克風陣列平台於不同環境下測試的結果，而實驗環境分為下列兩種：

1. 室內環境
2. 車內環境

在室內環境中，麥克風陣列系統將與語音辨識器作結合，並探討空間濾波器與語音辨識器辨識率的關係，另外也將會探討 VAD 與空間濾波器結合對語音辨識率的影響。

在車內環境中，將探討車內噪音的來源、車內錄音設備和空間濾波器與 SNR 的關係。

5.1 麥克風陣列於室內環境

5.1.1 空間濾波器與語音辨識率關係

本章節將探討語音訊號通過空間濾波器前後對語音辨識率的改善關係。 此語音辨識器為一辨識新竹科學園區廠商名稱辨識器，其字彙庫大小為 1339。圖 5-1 為實驗環境的實際照片，實驗環境中有兩個喇叭，一個喇叭用來播放園區廠商名稱，另一個為播放音樂聲，而圖中兩台電腦一為用來將訊號通過空間濾波器並即時的輸出給另一台電腦作辨識。

首先，先用真人語音錄製一百組新竹科學園區廠商名稱，並在下列三種情況下播放測試其辨識效果：

1. 安靜的環境下
2. 播放音樂的情況下並用單一麥克風作即時輸出
3. 播放音樂的情況下並將訊號通過空間濾波器

圖 5-2 為實驗環境的平面關係圖。

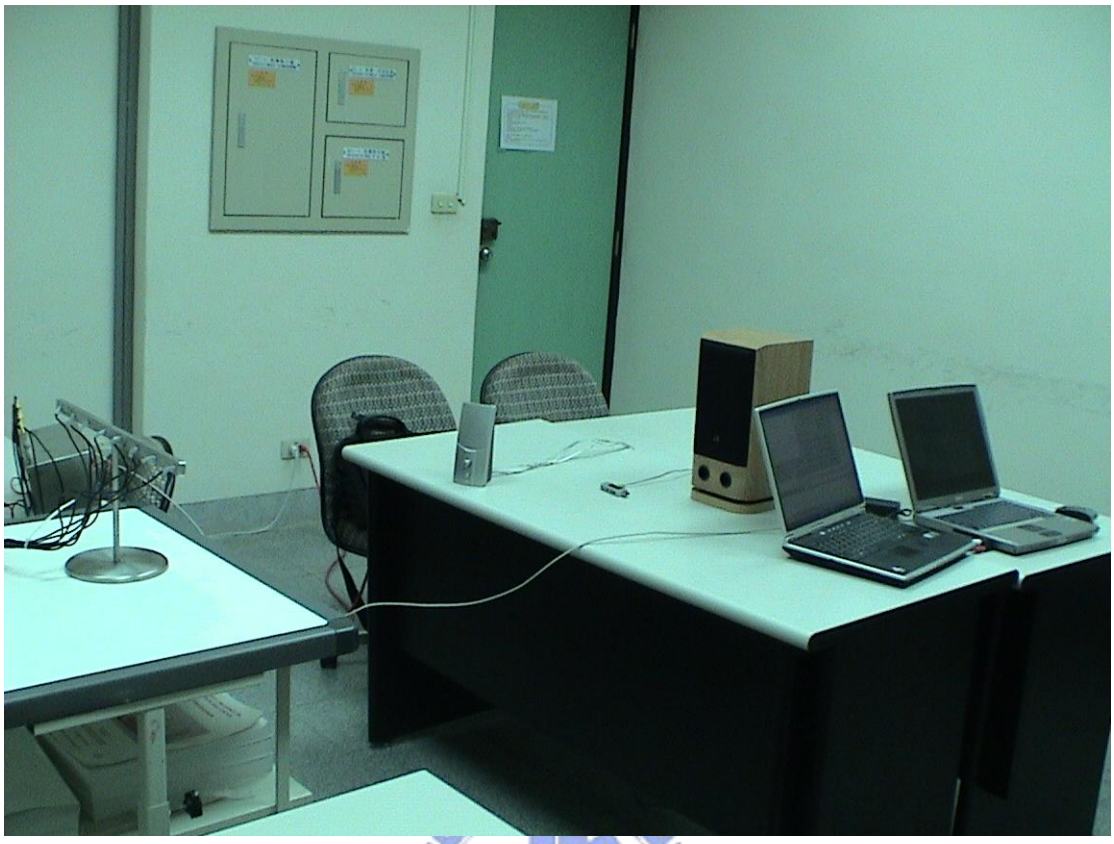

圖 5-1：實驗環境實際照片

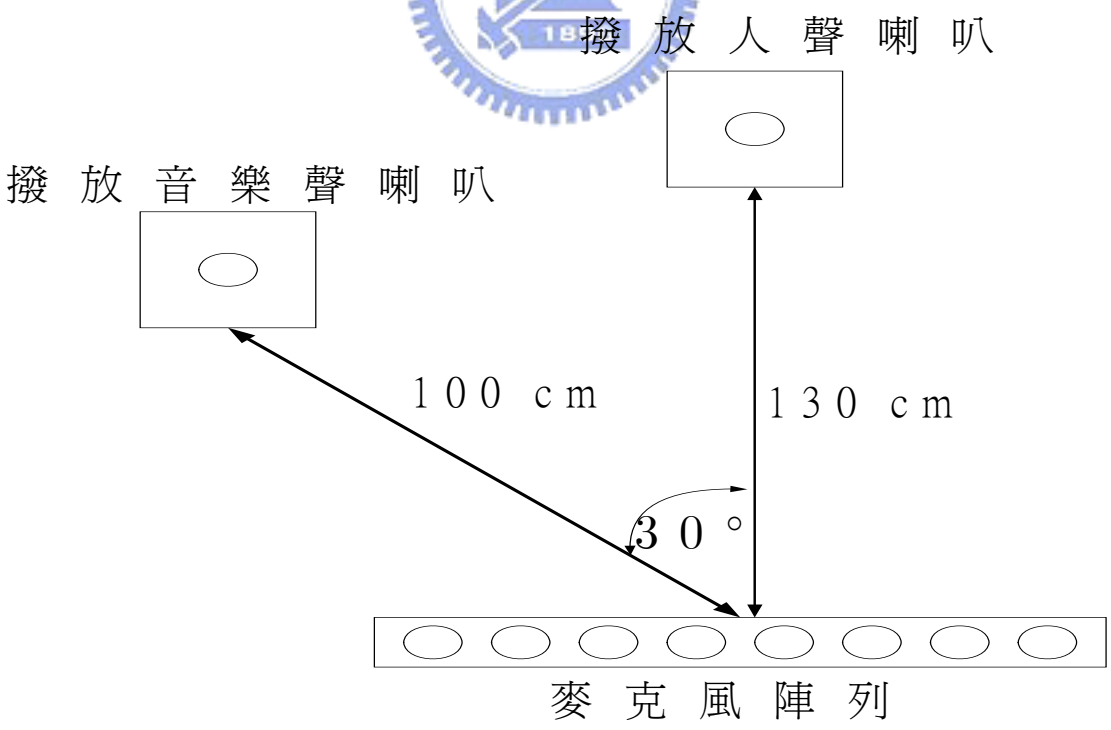

圖 5-2：實驗環境平面關係圖

實驗中 SNR 的計算方式如下：

$$10\log\left(\frac{\sum_{i=M}^{N}x^2(i)}{N-M+1}\right) \qquad (5\text{-}1)$$

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

$$10\log\left(\frac{\sum_{i=M_2}^{N_2}x^2(i)}{N_2-M_2+1}\right)-10\log\left(\frac{\sum_{i=M_1}^{N_1}x^2(i)}{N_1-M_1+1}\right)\ \text{dB} \qquad (5\text{-}2)$$

■ **測試一：真人語音「聯發科」＋音樂聲 空間濾波器濾波階數=256**

圖 5-3 為真人語音「聯發科」與音樂聲之混合訊號（流行歌曲：孫燕姿-奔），用單一麥克風錄到情形：

圖 5-3：真人語音「聯發科」與音樂聲混合訊號

圖 5-3 中音樂聲能量為-33.91 dB，而真人語音「聯發科」與音樂聲混合部分的能量為-24.4 dB，因此 SNR=9.51 dB。

圖 5-4 為測試一經過 256 階空間濾波器的處理結果：

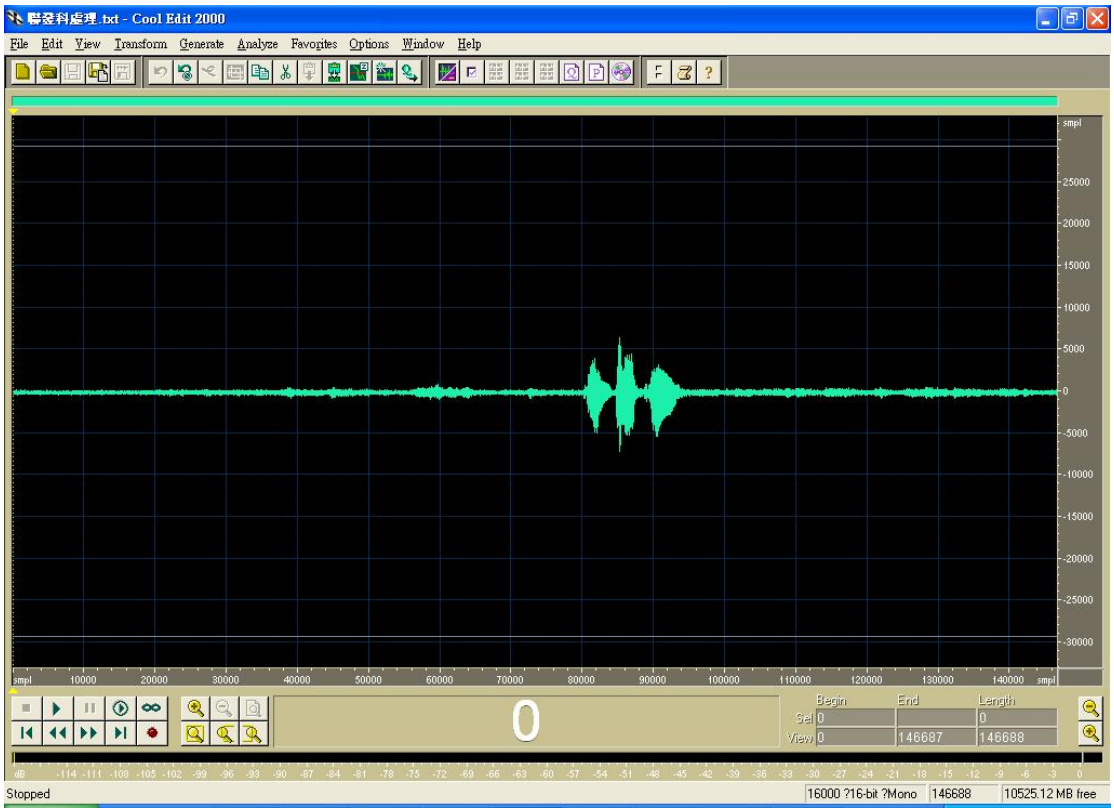

圖 5-4：測試一通過空間濾波器處理結果 （濾波器階數=256）

圖 5-4 中，音樂聲能量為-46.71 dB，而真人語音「聯發科」與音樂聲混合部分的能量為-25.27　dB，因此 SNR=21.44 dB。

測試一總結：

通過空間濾波階數為 256 的濾波作用，SNR 由原生的 9.51 dB 提升到 21.44 dB，其 SNR 增加了 11.93 dB。

■ **測試二：真人語音「聯發科」＋音樂聲　空間濾波器濾波階數=512**

圖 5-5 為真人語音「聯發科」與音樂聲之混合訊號（流行歌曲：孫燕姿-奔），用單一麥克風錄到情形：

圖 5-5：真人語音「聯發科」與音樂聲混合訊號

圖 5-5 中音樂聲能量為-34.73 dB，而真人語音「聯發科」與音樂聲混合能量部分為-25.11 dB，因此 SNR=9.62 dB。

圖 5-6 為經過 512 階空間濾波器的處理結果，在圖 5-6 中，音樂聲為-47.36 dB，而真人語音「聯發科」與音樂聲混合部分的能量為-24.71 dB，因此 SNR=22.65 dB。

測試二總結：

通過空間濾波階數為 512 的濾波作用，SNR 由原生的 9.62 dB 提升到 22.65 dB，其 SNR 增加了 13.03 dB。

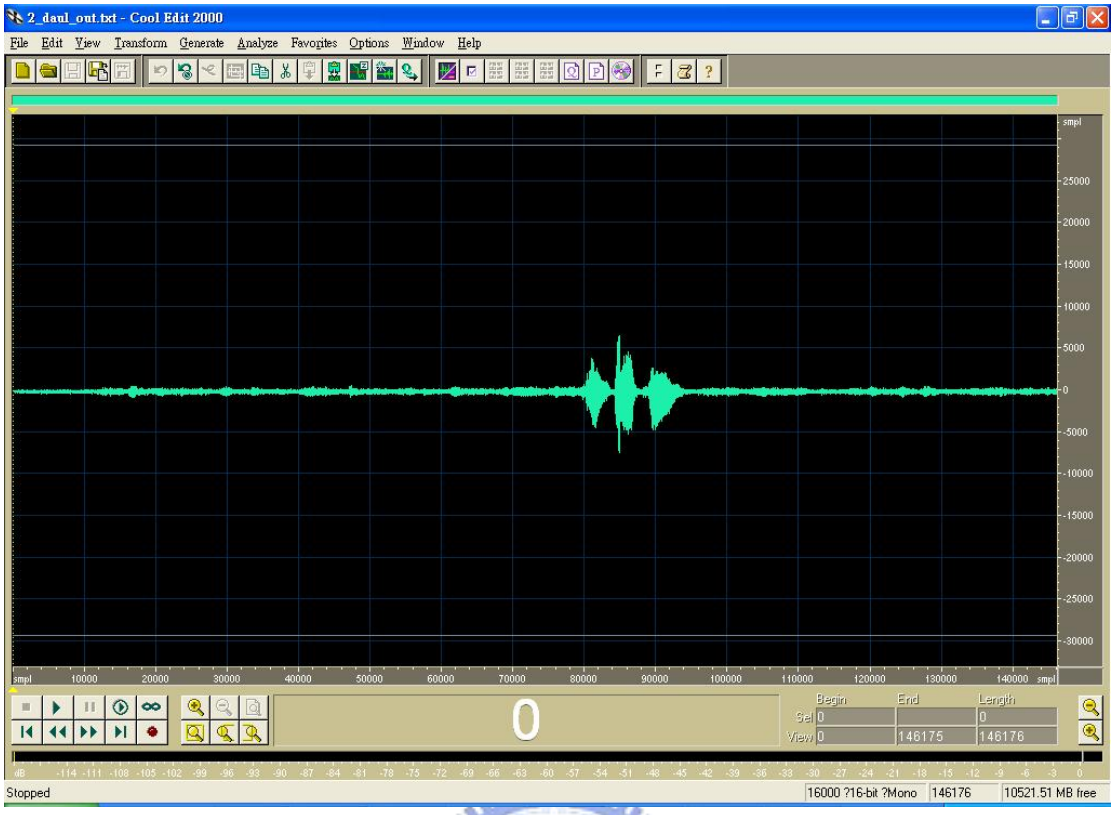

圖 5-6：測試二通過空間濾波器處理結果 （濾波器階數=512）

經過大量的統計，可發現空間濾波器階數的提高可增加其 SNR，但其運算量也相對的增大許多。

表 5-1 顯示出了在三種情況下的語音辨識情況，而在這三種情況下喇叭的音量皆為固定，而音樂聲的平均能量為-33.15dB，並且單一麥克風的增益與麥克風陣列的增益是相等的， 先錄製好一百組新竹科學園區廠商名稱，每組皆播放三次，因此每一種情況下會有三百種結果。

	正確次數	錯誤次數	正確率
安靜環境下使用單一麥克風	288	12	96%
吵雜環境下使用單一麥克風	101	199	33.6%
吵雜環境下使用麥克風陣列	231	69	77%

表 5-1：辨識率比較

5.1.2 VAD 結合空間濾波器與語音辨識率關係

本章節將探討 VAD 結合空間濾波器對語音辨識率的影響，其實驗環境與 5.1.1 節所介紹相等。

■ **測試一：真人語音「台積電」＋音樂聲 空間濾波器濾波階數=10**

圖 5-7 為真人語音「台積電」與音樂聲之混合訊號（流行歌曲：孫燕姿-奔），用單一麥克風錄到情形：

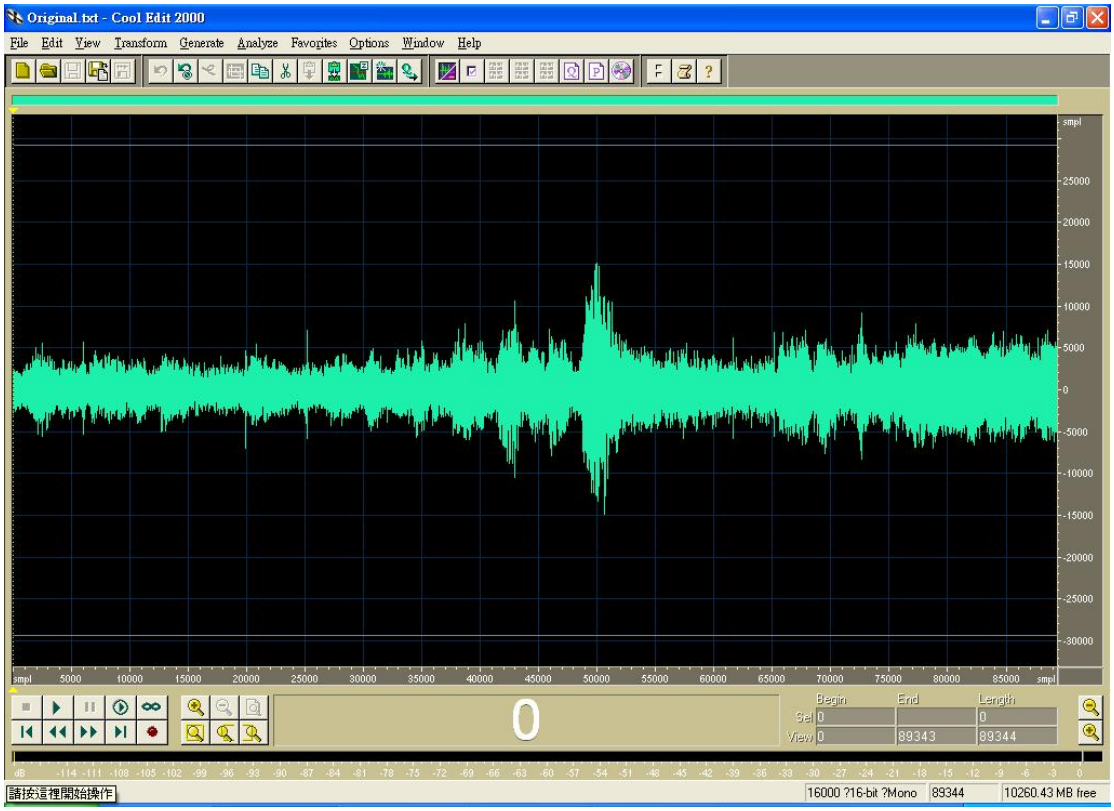

圖 5-7：真人語音「台積電」與音樂聲混合訊號

圖 5-7 中音樂聲能量為-24.52 dB，而真人語音「台積電」與音樂聲混合部分的能量為-20.75 dB，因此 SNR=3.77dB。

圖 5-8 為混合訊號經過 10 階空間濾波器的處理結果，在圖 5-8 中，音樂聲能量為-35.9 dB，而真人語音「聯發科」與音樂聲混合部分的能量為-23.76 dB，因此 SNR=12.14 dB。圖 5-9 為混合訊號經過 VAD 與 10 階空間濾波器的處理結果。

圖 5-8：測試一通過空間濾波器處理結果 （濾波器階數=10）

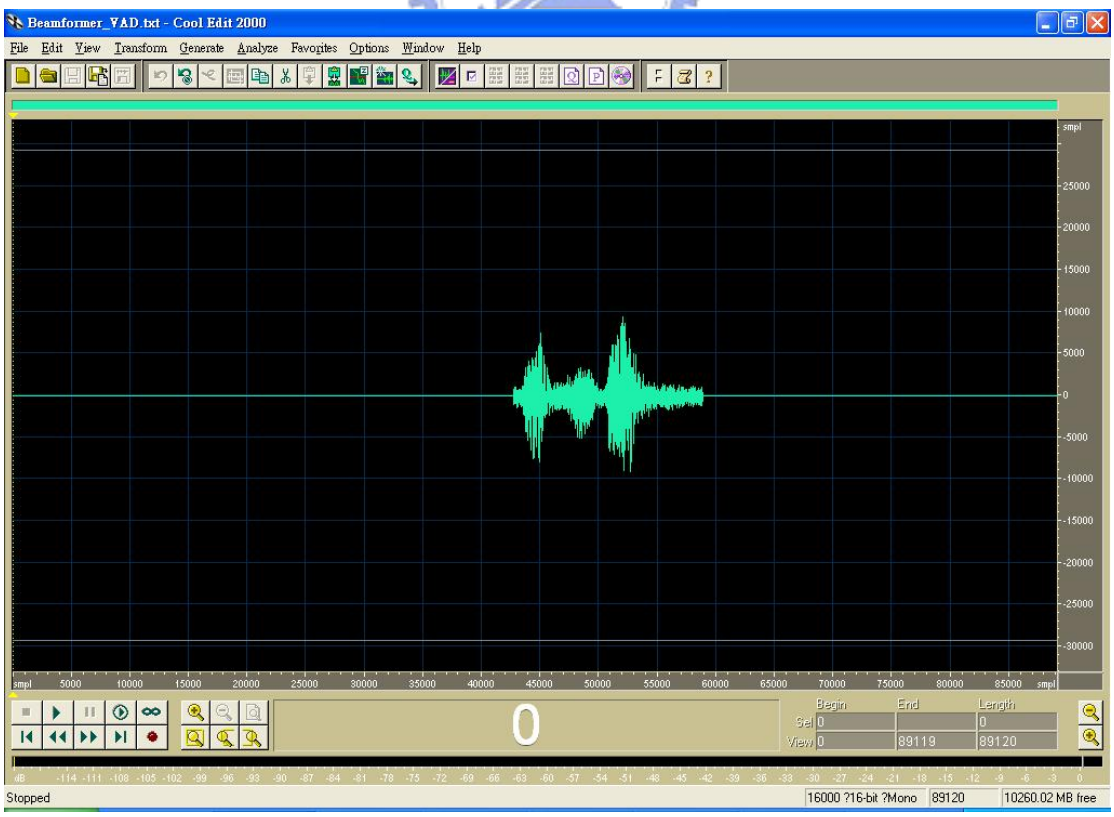

圖 5-9：測試一通過 VAD 與空間濾波器處理結果 （濾波器階數=10）

從圖 5-9 可觀察出，經過 VAD 與空間濾波器的處理，最後輸出只留下真人語音與音樂混合的部分，並且音樂聲有被壓制的效果，而圖 5-7 單純只有音樂聲的部分在圖 5-9 中已完全為零。

■ **測試二：真人語音「台積電」＋更高音樂聲　空間濾波器濾波階數=10**

圖 5-10 為真人語音「台積電」與音樂聲之混合訊號（流行歌曲：孫燕姿-奔），用單一麥克風錄到情形：

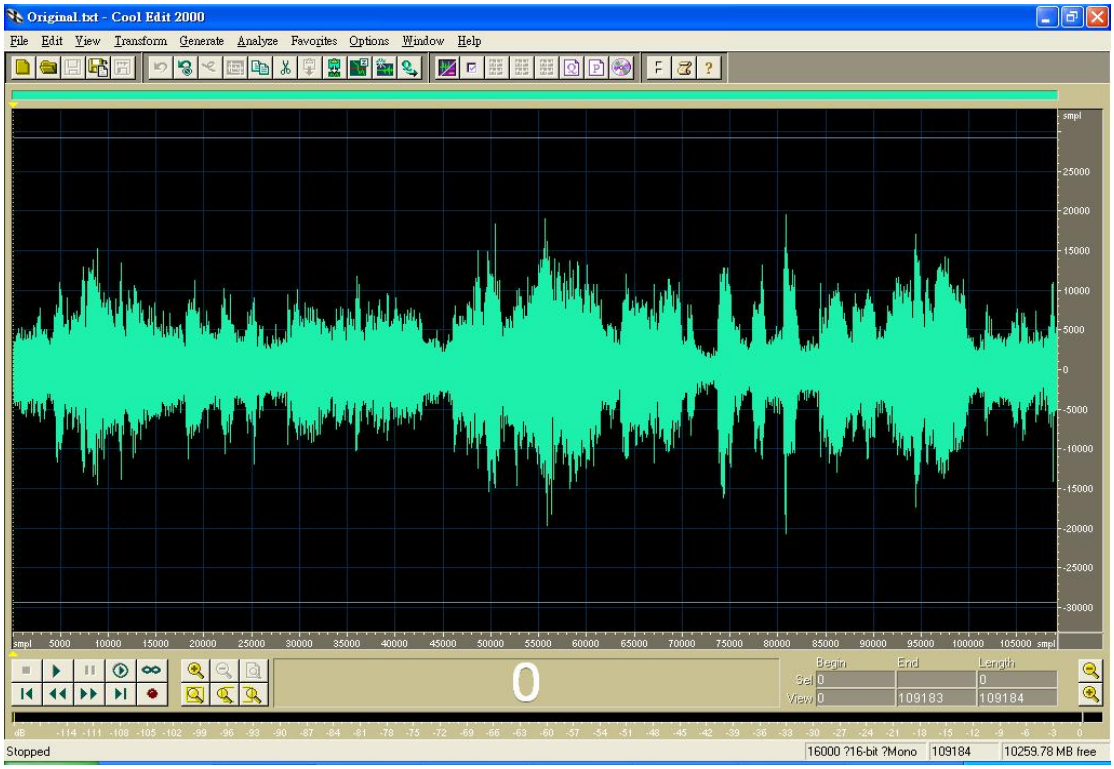

圖 5-10：真人語音「台積電」與音樂聲混合訊號

圖 5-10 中音樂聲為-20.32　dB，而真人語音「台積電」與音樂聲混合能量為-17.71 dB，因此 SNR=2.61dB。

圖 5-11 為混合訊號經過 10 階空間濾波器的處理結果，在圖 5-11 中，音樂聲為-33.84 dB，而真人語音「台積電」與音樂聲混合部分為-23.34 dB，因此 SNR=10.51 dB。圖 5-12 為混合訊號經過 VAD 與 10 階空間濾波器的處理結果。

圖 5-11：測試二通過空間濾波器處理結果 （濾波器階數=10）

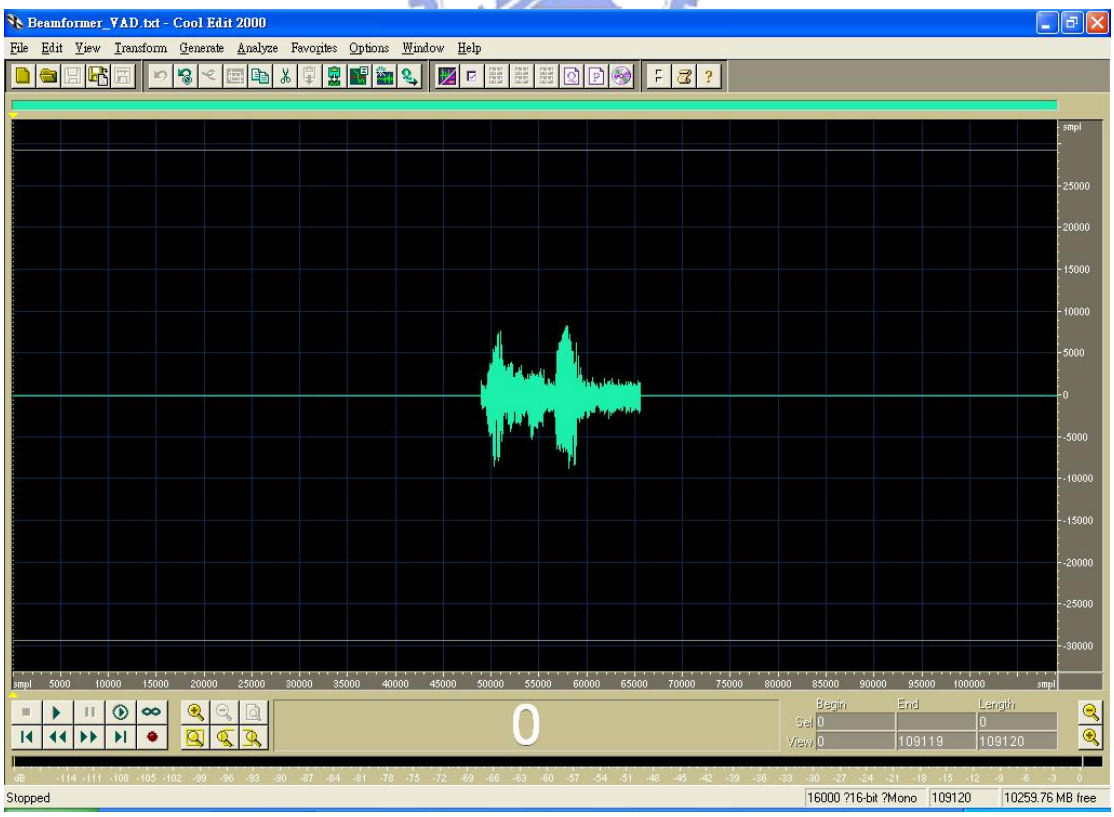

圖 5-12：測試二通過 VAD 與空間濾波器處理結果 （濾波器階數=10）

從測試二中可發現，只要空間濾波器效果良好，就算真人語音與音樂聲能量差不多，VAD 與空間濾波器處理過後，就可將真人語音過濾出。

表 5-2 顯示出了在四種情況下的語音辨識情況，而在這四種情況下喇叭的音量皆為固定，而音樂聲的平均能量為-31.06dB，並且單一麥克風的增益與麥克風陣列的增益是相等的， 先錄製好一百組新竹科學園區廠商名稱，每組皆播放三次，因此每一種情況下會有三百種結果。

	正確次數	錯誤次數	正確率
安靜環境下使用單一麥克風	289	11	96.3%
吵雜環境下使用單一麥克風	49	251	16.3%
吵雜環境下使用空間濾波器	221	79	73.7%
吵雜環境下使用 VAD 與空間濾波器	263	37	87.7%

表 5-2：辨識率比較

5.1.3 VAD 結合空間濾波器用於噪音源變動環境

本章節將探討 VAD 結合空間濾波器用於噪音源變動環境的效果，實驗環境如圖 5-13 所示，環境中有兩個噪音源位置，分別為第一噪音源與第二噪音源。實驗中將用兩種系統架構，分別為空間濾波器與 VAD 結合空間濾波器，兩種系統架構置於兩種噪音源位置測試其辨識效果。空間濾波器的係數為事先適應性調整適合於第一噪音源，而 VAD 結合空間濾波器係數為系統依據當時環境自動適應性調整。實驗條件與 5.1.2 節相同，先錄製好一百組新竹科學園區廠商名稱，每組皆播放三次，因此每一種情況

下會有三百種結果，實驗結果如表 5-3 所示。

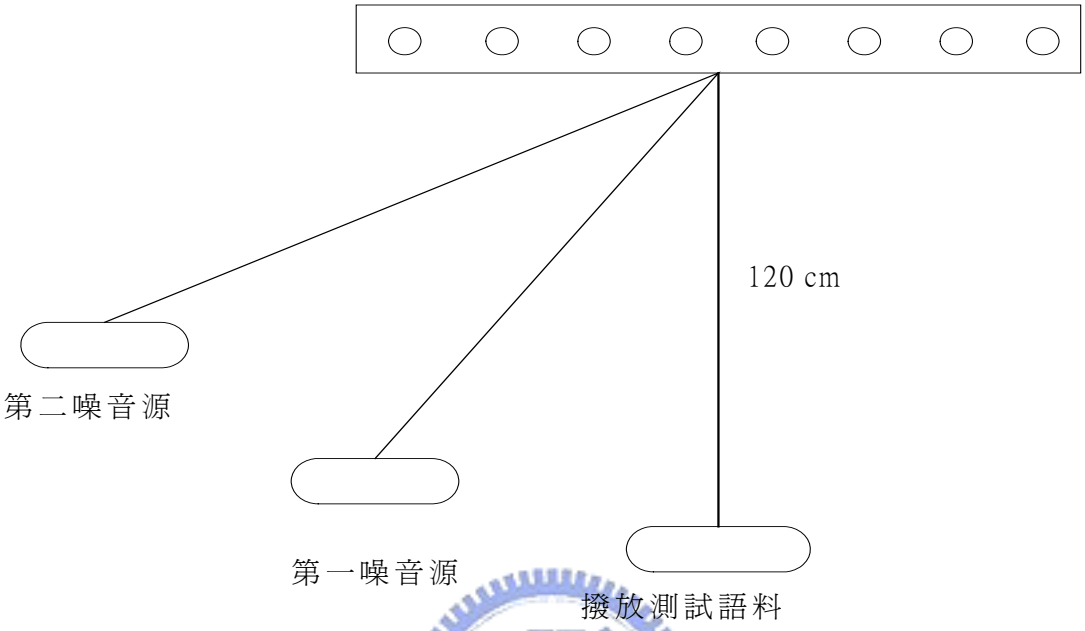

圖 5-13：實驗環境圖

實驗環境 / 系統架構	第一噪音源			第二噪音源		
	正確	錯誤	正確率	正確	錯誤	正確率
使用空間濾波器	212	88	70.7%	156	144	52%
使用 VAD 與空間濾波器	252	48	84%	248	52	82.7%

表 5-3：辨識率比較

從表 5-3 中可發現，VAD 結合空間濾波器用於不同位置的噪音源其辨識率改變不大，這是因為系統會自動適應性整空間濾波器係數，若只使用單一空間濾波器，因為系統並不會自動做適應性調整，因此當改變噪音源位置時，其辨識率變動就會較大。

5.2 麥克風陣列於車內環境

本章節將介紹麥克風陣列用於車內環境的效果，圖 5-14 為車內環境照片，車型為 Savrin 2.0，附駕駛座頭部有一喇叭模仿真人語音，而喇叭與麥克風陣列距離為 100 cm。

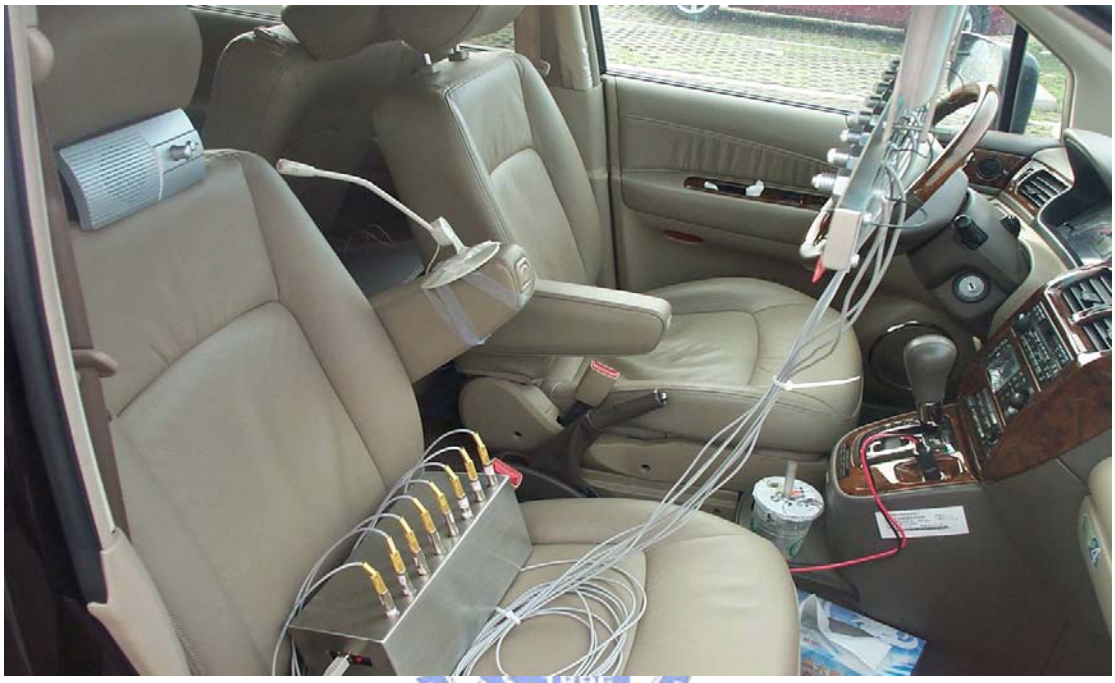

圖 5-14：麥克風陣列於車內環境實照

在處理 Normalize LMS 演算法前，必須先錄製一固定噪音源，而於車內環境中，固定噪音源的情況為：

1. 發動引擎
2. 不踩油門
3. 啟動冷氣
4. 窗戶關起
5. 音響關閉

圖 5-15 為在上述五種情況下錄製到的噪音，其噪音為一低頻噪音，頻率介於 20Hz 到 30Hz 之間。而圖 5-16 為將車子駕駛於高速公路上，時速 90 到 100 公里，喇叭播放真人語音所錄製到的結果。

- **測試一：麥克風增益為 20log（150）dB**

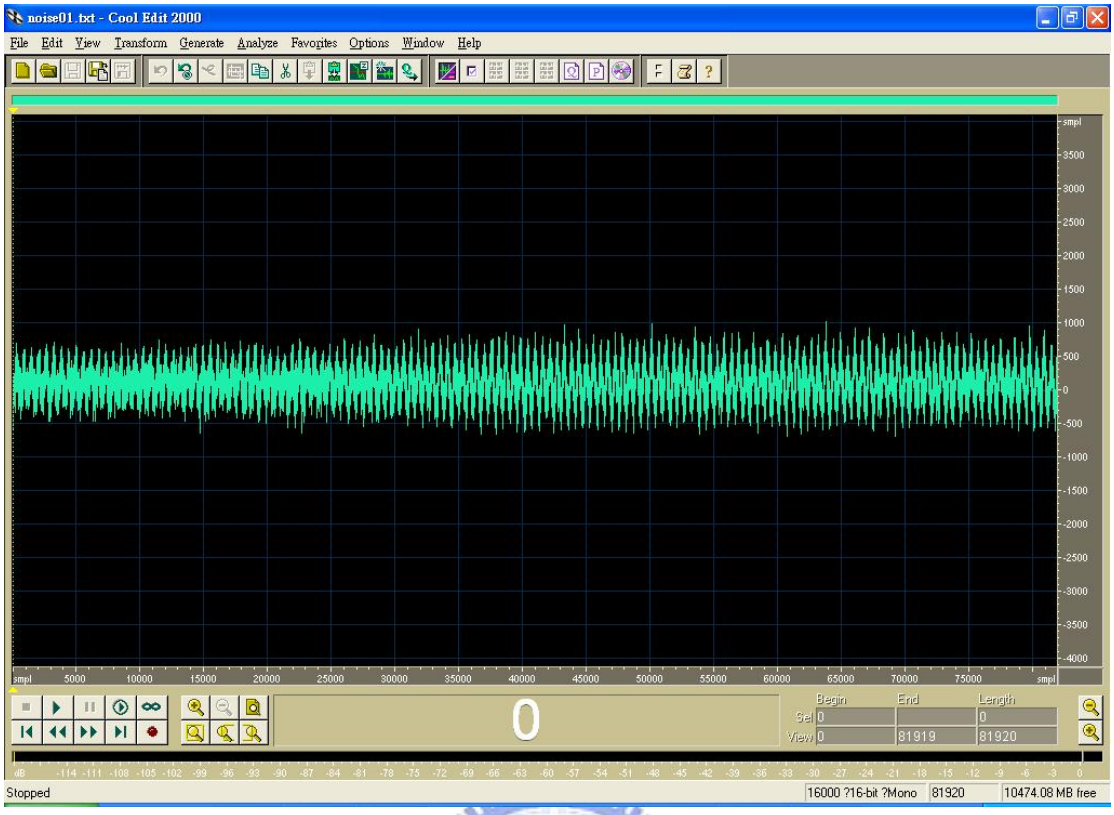

圖 5-15：車內固定噪音源

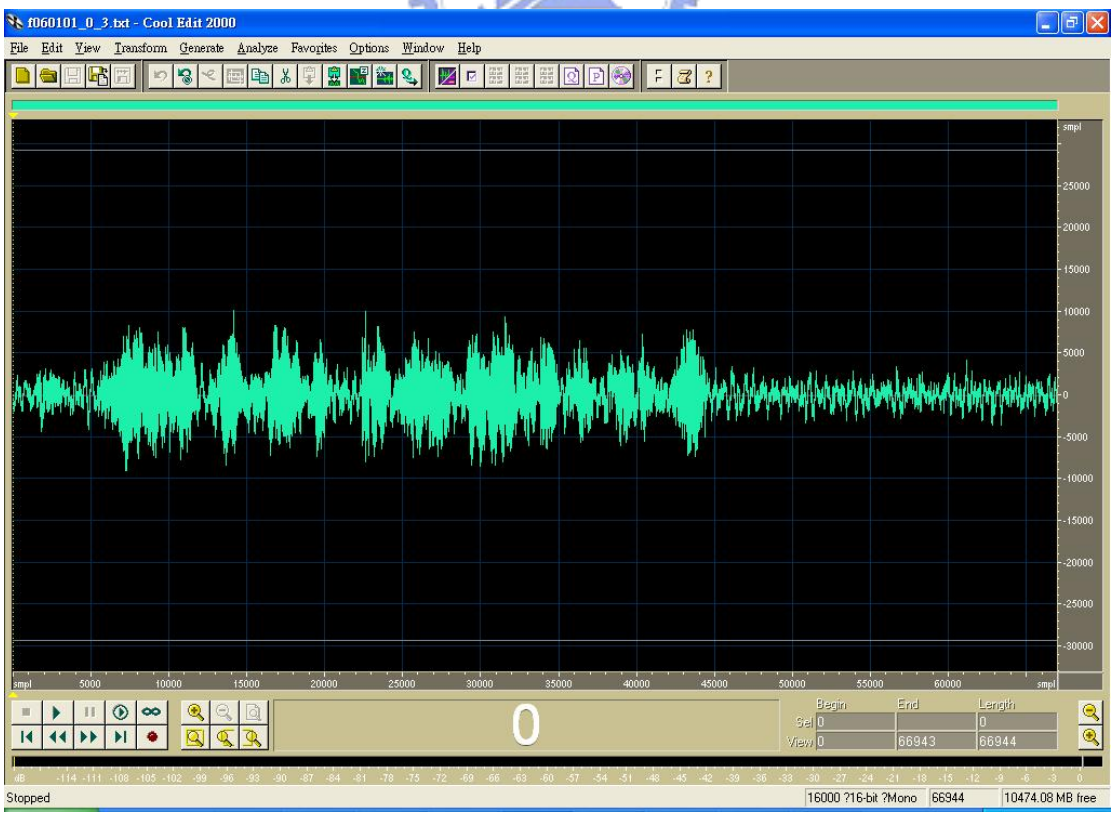

圖 5-16：車內真人語音與車內噪音混合之訊號

圖 5-17：車內真人語音與車內噪音混合訊號經過空間濾波器處理結果

在圖 5-16 中噪音為-29.51 dB，而真人語音與車內噪音混合部分的能量為-22.1 dB，因此 SNR=7.41 dB。

圖 5-17 為經過 256 階空間濾波器的處理結果，在圖 5-17 中，噪音為-51.42 dB，而真人語音與車內噪音混合部分的能量為-26.8 dB，因此 SNR=24.62 dB。

混合訊號經過空間濾波器的處理後，SNR 可提高 17.21dB。

測試二：麥克風增益為 20log（1000）dB

圖 5-18 為在第 65 頁所述五種情況下錄製到的噪音，其噪音為一低頻噪音，頻率介於 20Hz 到 30Hz 之間。而圖 5-19 為將車子駕駛於高速公路上，時速 90 到 100 公里，喇叭播放真人語音所錄製到的結果。

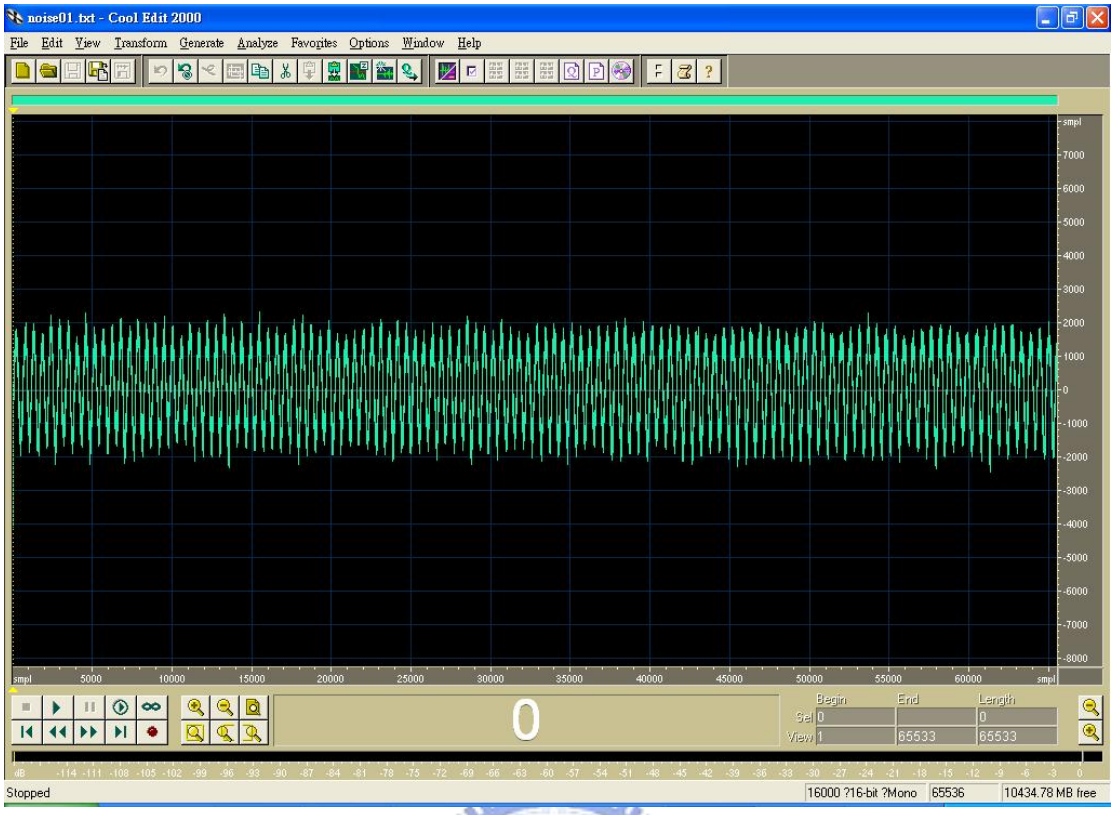

圖 5-18：車內固定噪音源

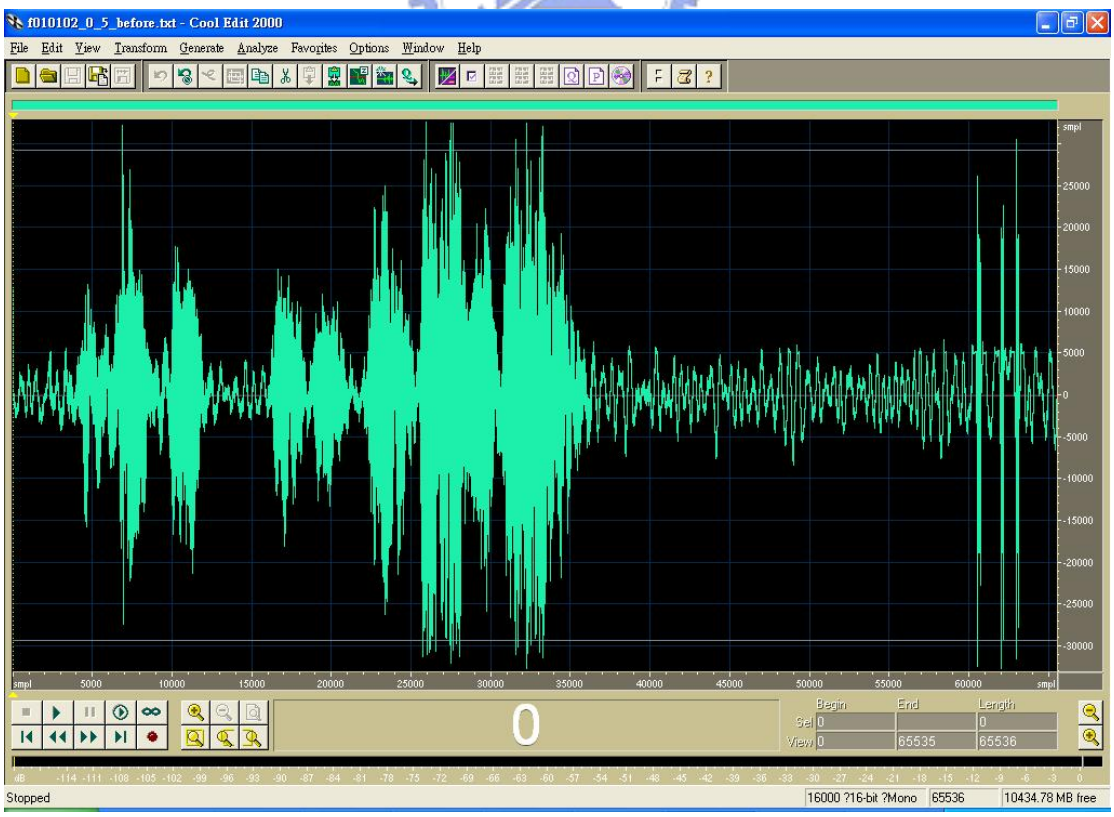

圖 5-19：車內真人語音與車內噪音混合之訊號

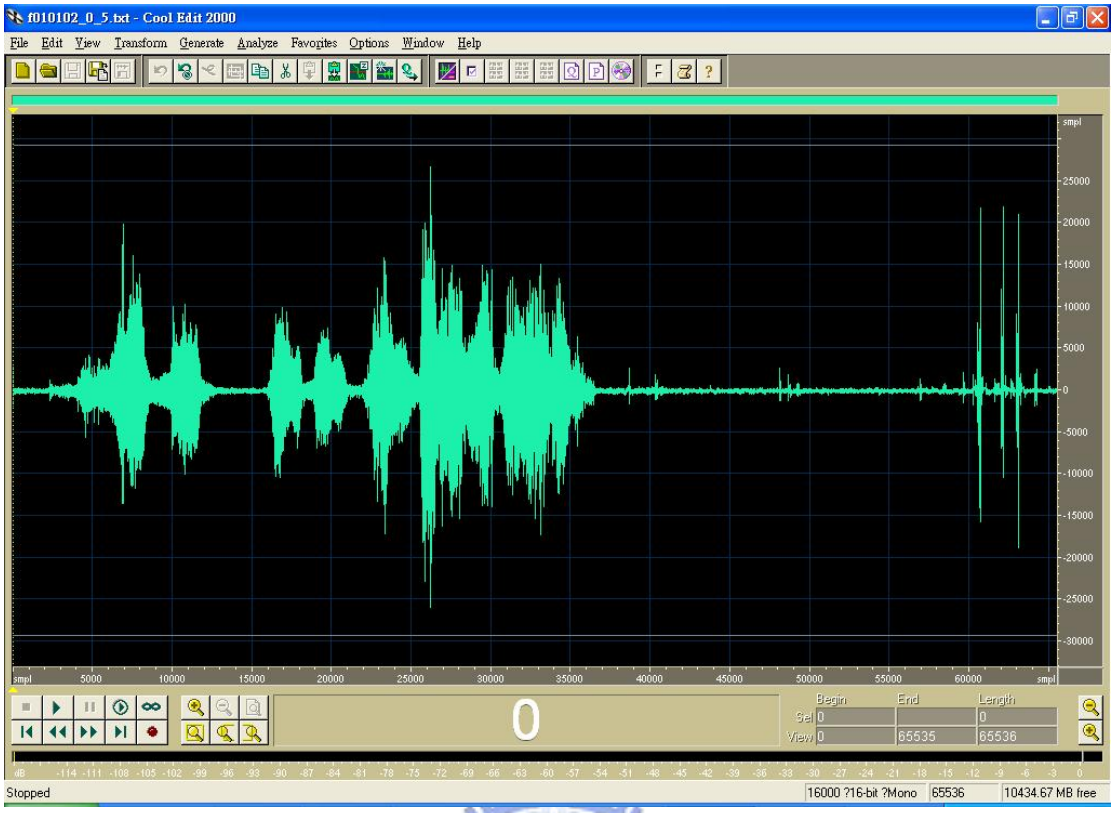

圖 5-20：車內真人語音與車內噪音混合訊號經過空間濾波器處理結果

在圖 5-19 中噪音為-20.01 dB，而真人語音與車內噪音混合部分的能量為-10.43 dB，因此 SNR=9.58 dB。

圖 5-20 為經過 30 階空間濾波器的處理結果，在圖 5-20 中，噪音為-34.69 dB，而真人語音與車內噪音混合部分的能量為-15.77 dB，因此 SNR=18.92 dB。混合訊號經過空間濾波器的處理後，SNR 可提高 9.34dB

從測試一與測試二可發現，當麥克風增益大時，路面顛簸的聲音皆會被錄進來（圖 5-19 語音後段的凸波），雖然 SNR 皆有提高，但空間濾波器濾除不掉路面顛簸的影響，因此 VAD 可考慮用進車內環境來消除路面顛簸的影響，因為路面顛簸為瞬間性的聲音，經過 VAD 的判別，可將此顛簸聲判定為非真人語音。圖 5-21 即為測試二之語音通過 VAD 的結果。

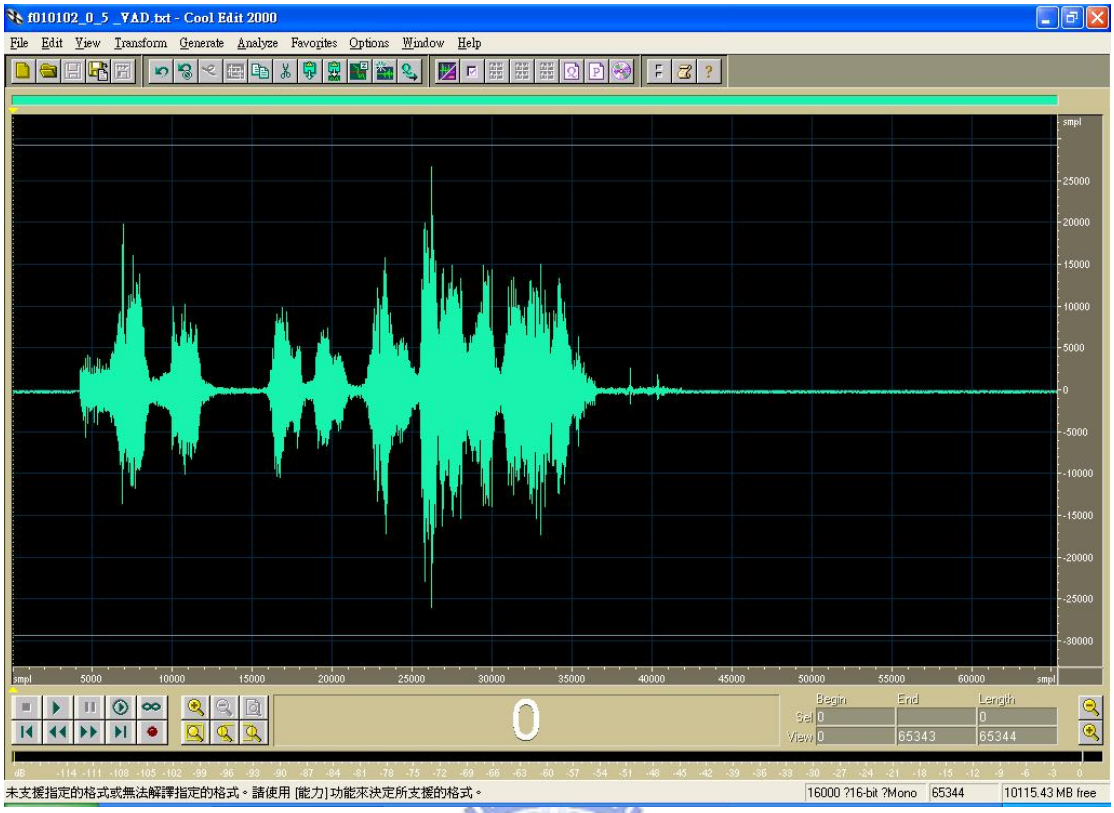

圖 5-21：車內真人語音與車內噪音混合訊號經過空間濾波器與 VAD 處理結果

從圖 5-21 中可發現，路面顛簸聲經過 VAD 後會判定為非真人語音，因此輸出為語音資料庫中的 silence。另外從實驗中我們也得之路面顛簸的噪音是透過聲波而傳遞近麥克風而非固體震動傳遞。

第六章 結論

6.1 研究成果

本論文已實作完成一以 USB1.1 為傳輸介面之 8 通道及時性麥克風陣列訊號處理平台，並在此平台實作完成真人語音偵測（VAD）與適應性空間濾波器整合，可使用於室內環境或車內環境，測試結論為：

1. 於吵雜環境下使用麥克風陣列，其語音辨識率較使用單一麥克風高。
2. 空間濾波器階數的提高可提升其 SNR。
3. VAD 與空間濾波器的結合可應用於噪音源位置變動的環境，若只使用空間濾波器則噪音源位置改變，則須人為去重新啟動 LMS，使其重新調整空間濾波器係數，若加入 VAD 則會當沒真人語音時，系統自動調整空間濾波器係數。
4. VAD 與空間濾波器的結合其語音辨識率高於只使用空間濾波器。
5. VAD 結合空間濾波器適用於噪音源變動的環境。
6. 在車內環境中，引擎一啟動就有一低頻雜訊，可透過空間濾波器將此低頻雜訊濾除乾淨，但若麥克風增益大時，空間濾波器無法消除路面顛簸的影響，但可透過 VAD 消去路面點跛的影響。

6.2 未來展望

在做 LMS 演算法時，必須先針對雜訊作出 zero-mean 及 Gaussian 的假設，若雜訊本身並非 zero-mean 及 Gaussian，則 Normalize LMS 處理效果會有限，但現實生活中並非所有接收到的雜訊皆為 zero-mean 及 Gaussian，因此未來日子可考慮其他適應性訊號處理演算法，像 H∞等。

Reference

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

[2] J. L. Flanagan , L. Landgraf , D. J. McLean , ”Matched-filter processing of hydrophone array” , J. Acousr. Soc. Am. 42, 1165 (A)(1967).

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

[4] 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

[5] 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

[6] 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

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

[8] 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

[9] 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

[10] 黃佑霖,”應用延遲濾波器之麥克風陣列訊號處理”,交大電控碩士論文,Jun 1999

[11] 陳界全,”即時聲源追蹤與空間濾波器設計”,交大電控碩士論文,Jun 2000

[12] 康創閔,”應用於個人電腦環境之即時語音純化系統設計”,交大電控碩士論文,Jul 2004

[13] Ta-Sung Lee,Array Signal Processing,(class note)

[14] 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

[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] Ali H. Sayed, Fundamentals of Adaptive Filtering, pp. 214-229.

[20] 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

[21] 來源網站：http://andrew.csie.ncyu.edu.tw

[22] Xuedong Huang,Alex Acero and Hsiao-Wuen Hon,Spoken Language Processing.

[23] 來源網站： http://www-306.ibm.com/software/voice/viavoice/

[24] Cypress Semiconductor, The EZ-USB FX Technical Reference Manual, Version 1.2.Cypress Semiconductor Corporation, 2000.

[25] 許永和,8051 微處理機程式設計,pp3-11 – 3-15.

[26] 來源網站： http://www.intel.com/software/products/mkl/

[27] Bradley Bargen and Peter Donnelly, *Inside DirectX*, 1999, pp. 203-246.

[28] 來源網站： http://speech.cm.nctu.edu.tw/