國 立 交 通 大 學
電機與控制工程研究所

碩 士 論 文

# 利用訊號特徵及麥克風陣列之聲音監控系統

# Audio Surveillance Using Signal Characteristics and Microphone Array

研 究 生： 陳 俊 宇
指導教授： 胡 竹 生 博士
周 志 成 博士

中 華 民 國 九 十 七 年 九 月

# 利用訊號特徵及麥克風陣列

# 之聲音監控系統

# Audio Surveillance Using Signal Characteristics and Microphone Array


研 究 生：陳 俊 宇　　Student：Chun-Yu, Chen

指導教授：胡 竹 生 博士　　Advisor：Dr. Jwu-Sheng, Hu

周 志 成 博士　　Dr. Chi-Cheng, Jou



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

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of Master

in

Electrical and Control Engineering

September 2008

Hsinchu, Taiwan, Republic of China

中 華 民 國 九 十 七 年 九 月

# 利用訊號特徵及麥克風陣列之聲音監控系統

研究生：陳　俊　宇　　指導教授：胡　竹　生　博士

周　志　成　博士

國立交通大學電機與控制工程研究所碩士班

## 摘　要

本論文針對室內訊噪比(SNR)很低的吵雜環境提出辨識出雜訊和聲源的聲音監控系統。在傳統上單顆麥克風在訊噪比很低的情況下，無法辨識出雜訊和聲源。在此藉由麥克風陣列訊號擷取系統以擷取多通道聲音資訊，利用空間濾波器來抑制干擾源的影響可以提高訊噪比(SNR)降低偵測的錯誤率。並利用訊號特徵與高斯混合模型(Gaussian Mixture Model)的方法去建立聲音監控的背景模型，在此用期望值最大演算法(EM Algorithm)去估計模型參數。最後結合背景模型中聲場特徵的統計資訊，做背景聲音和聲源的判定。本研究以 USB1.1 介面、8 通道麥克風陣列訊號處理實驗平台進行研究，以此實驗平台錄製麥克風陣列語音樣本並供相關研究。

# Audio Surveillance Using Signal Characteristics and Microphone Array

Student： Chun-Yu, Chen Advisor：Dr. Jwu-Sheng, Hu
Dr. Chi-Cheng, Jou

Institute of Electrical and Control Engineering

## ABSTRACT

The purpose of this thesis is to construct an audio surveillance system to recognize the interference sources and sound sources in a low signal-noise ratio (SNR) noisy environment. Traditionally, single microphone can not recognize the noise and sound sources in a low signal-noise ratio (SNR). First, the time delay between microphones are estimated from the multiple-channel sound data acquired by the digital microphone array acquisition system. Second, we suppress the effect of the interference sources by spatial filter to promote signal-noise ratio (SNR) and to reduce the false detection ratio. Second, we create the background model of surveillance system by using the signal characteristic and the Gaussian Mixture Modeling method, and then to estimate the model parameters by utilizing the EM algorithm. Finally, we can recognize the sound sources from the interference sources through the statistic information of sound field characteristic. A microphone array of 8 microphones with USB 1.1 interface was made for the implementation platform.

# 誌 謝

在這兩年的研究所生涯令我受益良多，讓我發現生活的意義，也增長許多知識。也培養了自我尋找問題及解決問題之能力，更讓我學習如何在大環境下待人處世的方法。使我不僅在學術相關領域有所增長，更讓我個人的處事態度更成熟圓滑。在此感謝電控所中每一位老師，在課堂上的諄諄教誨。更要感謝恩師 胡竹生教授於課業與生活中的啟發與細心教導，讓我學習到研究之方法與為人處事等方面的道理。在研究所的日子塑造了全新的我，並改變了日後的想法，並使我對日後的人生有更明確的目標，對於老師兩年來之提攜與教誨，由衷感謝。

在這兩年的研究所生活中，感謝女友鳳呈對於我的支持與陪伴，並感謝實驗室成員的陪伴。特別感謝興哥在聲學方面的教導；感謝宗敏學長教導我如何管理實驗室網頁；感謝劉大帶我進去重訓的世界；感謝永融學長在課業或研究上的教導；感謝鏗元學姊教我最佳化理論；感謝楷祥學長在研究上的鼓勵；還有很會帶氣氛的弘齡，熬夜都不會累的法哥、聯誼一定要找的大師兄、研究能力很強的明唐、打架一定要帶的 Judo、認識超多人的 Dodo、有很多鬼點子的 papa、很喜歡唱歌的 gum、邏輯能力很強的啟揚在課業上的幫助、吃粥減肥的肉鬆、只吃金針菇的瓊文、實驗室最常拿書卷的阿吉，感謝你在論文上的指導、實驗室唯一敢嗆老師的 Lundy，另外還有我認識的朋友親人，因為有你們所以使我在研究所充滿了無限回憶。

最後，感謝我最深愛的父母親的養育之恩，辛辛苦苦將我拉拔長大，並在研究期間給予我許多支持、鼓勵與關懷，讓我順利完成研究所學業。將本論文獻給他們，願與他們共享這份成果與榮耀。

# 目錄

# 圖目錄

# 表目錄



# 第一章 緒論

## 1.1 研究背景與動機

為了能夠知道我們想要了解的環境狀況，一般都是使用影像資料當作環境的監控。但是一些特殊的狀況使用聲音來分別會是更容易的。特殊的情況如是否有語音的發生。因為一般影像的資料只能辨識出嘴唇的動作，並不能保證有語音的發生。因此若能設計一套聲音監控系統，針對環境中固定已經有的聲音進行訓練。根據麥克風收集到的聲音進行辨識，能夠區分不屬於環境中固定的聲音，就可以知道目前的環境有不屬於正常的狀況發生，藉由達到環境監控的目的。

一般的異音監測皆使用單一麥克風做為語音訊號的輸入，在安靜的環境下已有不錯的辨識成果，然而，當應用在噪音很大的室內裡，異音辨識的效果將大打折扣，因此，如何抑制噪音並加強異音訊號已成為聲音監控的關鍵性技術。

我們利用麥克風陣列的優勢，透過空間濾波器，可針對特定角度加強訊號。但可能因為環境的音量和說話的音量大小上差異不大，因此利用高斯混合模型針對環境固有的聲音當作背景聲音的模型。當有輕微的說話發生時，根據麥克風收到的聲音所發生的機率值來判斷是否發生與背景模型不相同的聲音出現。

使用高斯混合模型方法是因為大部分針對語音活動偵測的方法，都是當語音特徵明顯時，才會有較高的辨識結果。但是在背景環境音量和語音的差異不大時，使用語音活動偵測等方法也不容易辨識出。所以藉由環境聲音的統計資訊的幫助，判斷麥克風收到的聲音在統計資訊裡所發生的機率是否在我們所認定的範圍裡，使得達到聲音監控的目的。

## 1.2 研究目標

本論文目標將分為:

1. 利用麥克風陣列搭配高斯混合模型尋求異音方位的方法。
2. 利用麥克風陣列搭配高斯混合模型完成異音監控的方法。
3. 探討與比較不同語音活動偵測的演算法在不同增異和聲源在不同角度的實驗結果。

## 1.3 論文架構

整篇論文大致上可以分為三個部分，分別是第二章的系統簡介與理論說明第三章的聲音監控方法與最後的實驗比較。系統簡介與理論這章節會對系統所使用到或者是和系統互相比較的相關聲音演算法做介紹，聲音監控這章節則會針對異音監控整個系統的演算法和異音監控的建立流程做說明。最後的實驗比較會對目前可以對異音監控的演算法和本論文所提出的演算法互相比較和分析並對研究成果作個結論。

# 第二章 系統簡介與理論說明

## 2.1 系統之簡介

理論上，想要針對室內訊噪比很低的吵雜環境，提出辨識出原本屬於環境的背景聲音和非環境背景聲音的監控系統。在傳統上，單顆麥克風在訊噪比很低的情形下使用資料進行分析無法正確辨識出雜訊與聲源。所以在此藉由麥克風陣列擷取系統以擷取多通道聲音資訊。利用空間濾波器:Delay and Sum Beamformer 來抑制干擾源影響可以提高訊噪比降低偵測的錯誤率。之後利用麥克風陣列對多方向建立高斯混合模型[1][2]。對多方向進行聲音監控，當聲源發生在某一監控的角度上，聲源訊號會被放大並且根據高斯混合模型判斷出有發生異音和異音的方位。之後再利用高維度的高斯混合模型確認異音的發生。

論文中是以高斯混合模型(Gaussian mixture model，簡稱 GMM)描述出聲音之間頻率的分佈情形，用以決定一個環境聲音的統計資訊模型。論文中針對不同訊噪比、不同聲源角度所做的實驗分析，並且包括其他聲音監控所使用的演算法的差異比較。

## 2.2 陣列訊號處理

### 2.2.1 陣列式訊號處理簡介[22]

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

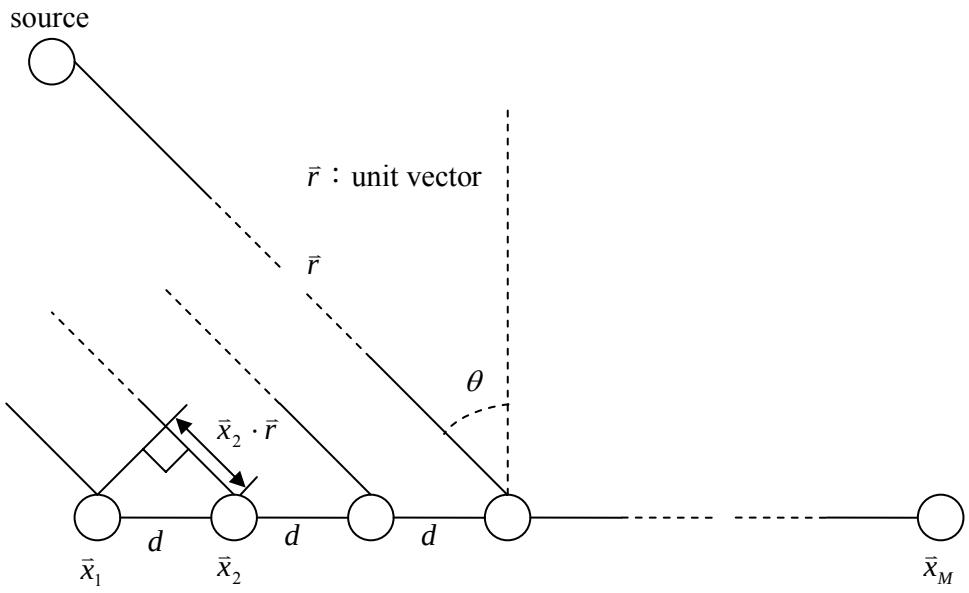

**圖 2-1 陣列模型**

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$\lambda_c$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。

## 2.2.2 陣列型態:均勻線性陣列

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$\theta$，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。

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

### 2.2.3 均勻陣列空間響應

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：

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

$$=e^{j\frac{k_c(M-1)d}{2}\sin\theta}\frac{\sin\left(\frac{k_cMd}{2}\sin\theta\right)}{\sin\left(\frac{k_cd}{2}\sin\theta\right)} \tag{2-4}$$

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。

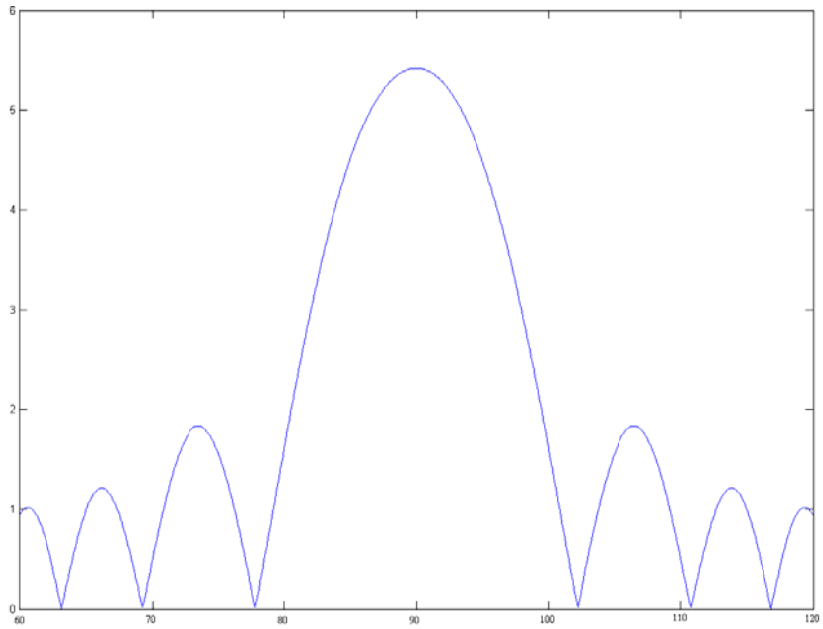

**圖 2-2 均勻線性陣列之空間響應（M=8，frequency=100Hz，d=10）**

### 2.2.4 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。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 sinθ 是一週期為 $\lambda_c/d$ 的週期性的函式，關係圖如(圖 2-3)所示。

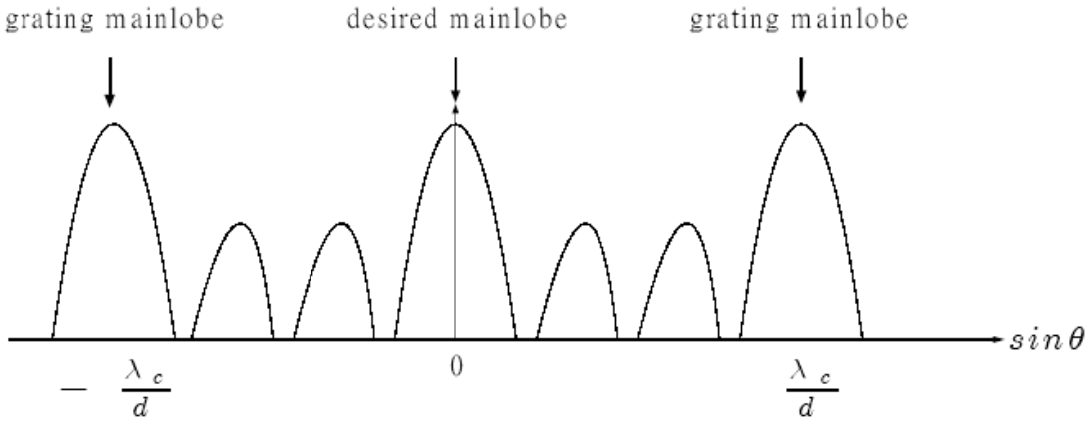

**圖 2-3 Grating Lobe 示意圖**

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 2-3)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。

## 2.2.5 Delay-and-Sum Beamformer[3]

假設有一包含$M$個麥克風的麥克風陣列，每一組相鄰的麥克風的距離為 d，今有一語音訊號(假設為平面波)從我們假設方向$\theta_s$傳播過來，麥克風的輸出為$x_t^i$，$1 \le i \le M$，則在時間 t 的時間，當第$i$支麥克風收到平面波的訊號，第$i+1$支麥克風則需要等到聲波再前進距離$R(R = dcos\theta_s)$方可收到訊號如(圖 2-4)。

若聲波的速度為$C$，則第$i+1$個麥克風延遲的時間

$$\tau = \frac{R}{C} = \frac{dcos\theta_s}{C} \qquad (2\text{-}6)$$

所以$x_t^i = x_{t+\tau}^{i+1}$，因此我們可以估算第 i 個麥克風與第 1 個麥克風的關係如下:

$$x_t^i = x_{t+(i-1)\tau}^{i+1} \qquad (2\text{-}7)$$

而整個 Delay-and-Sum Beamformer 的輸出$\hat{x}_t$，就是將每個麥克風間的時間研延遲作補償後合成再取平均而得:

$$\hat{x}_t = \frac{1}{M}\sum_{i=1}^{M} x_{t+(i-1)\tau}^{i} \qquad (2\text{-}8)$$

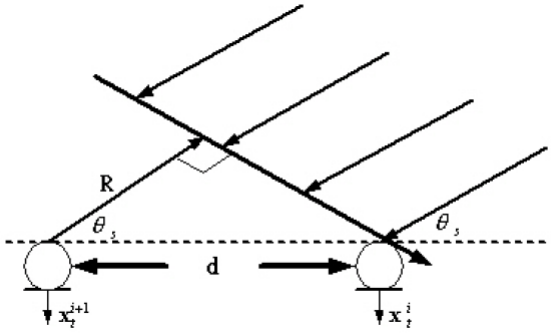


**圖 2-4 相鄰麥克風的時間延遲**

## 2.3 語音活動偵測

### 2.3.1 音框的選取

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越零率，雜訊及氣音的越零率都很高，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[4][5]。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、頻譜失真和越零率四種參數來判定[6][7]。

對於一段語音的離散時間訊號，使用一個固定長度的視窗(window)套上去，只看視窗內的訊號，對這些訊號做演算，求出在這視窗內的語音特徵。套在視窗上的這一段語音即稱為音框[25]。移動視窗到下一個時間點，就得出下一個音框。根據每一個音框計算語音的特徵參數。例如對於一個取樣頻率為 16kHz 的語音訊號，它的取樣間距是 62.5 微秒，視窗長度取 512 點，換算成時間就是 32 ms，這也就是一個音框的長度。我們讓前後音框重疊 1/2 個音框長度，則每次向前移動視窗的距離就是 16 ms，也就是

256 個取樣點。

對於一定長度的語音訊號，視窗的移動距離越短，得出的音框數目就越多，語音特徵也就越多。當語音訊號變化快的時候，如果視窗越多，可以看出語音特徵在短時間內改變的情形。

### 2.3.2 語音端點偵測

由於量化後的語音資料，包含實際語音訊號與雜訊及靜音部份。所以我們要利用語音的特徵偵測發生語音的判斷。根據語音訊號端點偵測判斷時所採用的參數，大致上可區分為三大類型(1)時域端點偵測法;(2)頻率端點偵測法;(3)混合參數端點偵測法等;其中時域端點偵測法最簡單也最常被應用的一種方法[26]，但是抗雜訊能力較低;而頻域端點偵測法以及混合參數端點偵測法，兩者的抗雜訊能力較高，但所需的計算量較複雜;所以本文僅針對時域端點偵測法做說明。

在時域端點偵測中，用來判斷端點偵測判斷依據的參數:能量參數(Energy)以及越零率參數(Zero-Crossing Rate：ZCR)[26]。可藉由這兩個參數配合臨界值的設定，即可找出錄音訊號中，發生語音的端點所在之處。在時域端點偵測法中，因為語音訊號的能量一般都比背景雜訊大，所以偵測有效語音訊號最直接的方法，就是依據訊號區段的能量大小，來做為判斷的依據。然而，有些字音在字首或字尾有子音或磨擦音的存在，因其能量太小不易被偵測出來，因此在端點偵測時，除了依據能量參數外，也常使用訊號的越零率參數，來找出精確的語音訊號端點。

#### 2.3.2.1 能量偵測

一個音框內的能量測量為

$$E(n) = \log\{\sum_{n=m-N+1}^{m}[x(n)]^2\} \quad (2\text{-}8)$$

把一個位於 n 時刻的音框內 N 個訊號，取平方再相加，就得到音框之能量 E(n)。視窗長度會影響計算出的音框能量，長度越長，含蓋的取樣點愈多，相當於是越多的值被平均掉，因此得到的曲線越平滑。反之，音框越短則音框能量隨著訊號改變的現象越明顯，曲線就較不平滑。

### 2.3.2.2 越零率測量

語音訊號波形上可以找到一條零線，振幅在其上為正，在其下為負，沿著時間看波形的擺動，當振幅從負變正，或是從正到負，都是越過零線。單位時間內越過零線的次數多，表示波形擺動劇烈。我們計算一個音框內越過零線的次數，可以得出其越零率，計算公式如下:

$$Z_x(m) = \frac{1}{N}\sum_{n=m-N+1}^{m}\frac{1}{2}|sgn[x(n)] - sgn[x(n-1)]| \quad (2\text{-}9)$$

上式中的函數 sgn[.]是符號函數。

利用能量與越零率，我們可以做語音的端點偵測。最簡單的想法就是從能量上做判斷，能量大於某個門檻就是語音，否則就不是語音。但這種做法不太精確，因為聲帶不振動的輔音通常能量也很小，有時候不說話而背景噪音很大，反而會看到大的能量，一個比較可靠的做法是同時看能量的變化與越零率的變化。

在語音未開始之前，會取到一段背景噪音的訊號，在語音結束之後也會取到一段背景噪音。語音端點偵測就是要找出語音從第幾個音框開始，到第幾個音框結束。假設開始有一段背景噪音，利用這段噪音的訊息來訂定能量與越零率的門檻值，作為判斷語音端點的依據[25]。

(1) 將取到的一段聲音以固定音框長度及固定音框間距，轉成一序列的音框。然後計算每一音框的能量與越零率。

(2) 取最前面的若干個音框視為噪音部份，求其能量與越零率之分佈，即計算其平均值與變異量。

$$\bar{E}_{bn}=\sum_{l=1}^{N_{bn}} E_{x,l} \quad (2\text{-}10)$$

$$\bar{Z}_{bn}=\sum_{l=1}^{N_{bn}} E_{x,l} \quad (2\text{-}11)$$

$$\sigma_{E_{bn}}^2=\frac{1}{N_{bn}}\sum_{l=1}^{N_{bn}}(E_l-\bar{E}_{bn})^2 \quad (2\text{-}12)$$

$$\sigma_{Z_{bn}}^2=\frac{1}{N_{bn}}\sum_{l=1}^{N_{bn}}(Z_l-\bar{Z}_{bn})^2 \quad (2\text{-}13)$$

(3) 訂兩個能量門檻及一個越零率門檻值。

$$T_{EL}=\bar{E}_{bn}+\alpha_1\sigma_{E_{bn}} \quad (2\text{-}14)$$

$$T_{EU}=\bar{E}_{bn}+\alpha_2\sigma_{E_{bn}}\ ,\alpha_1<\alpha_2 \quad (2\text{-}15)$$

$$T_Z=\bar{Z}_{bn}+\alpha_3\sigma_{Z_{bn}} \quad (2\text{-}16)$$

(4) 沿著音框序列，標注第一個能量超過$T_{EL}$的音框，注記為$N_V$。如果其後連續的$B$個音框，其能量大於$T_{EL}$，而且$B$個音框之後，能量更是大於$T_{EU}$，則$N_V$視為可能的語音起點。反之，在$N_V$之後的$B$個音框內，有小於$T_{EL}$的，或是$B$個音框之後不會大於$T_{EU}$，則不是語音起點，可能只是短暫的噪音造成的現象。因此放棄此$N_V$點，繼續往下找。

(5) 找到$N_V$之後，往回檢查，看其前個音框越零率，是否大於$T_Z$，若是就繼續往回找，直到越零率小於$T_Z$為止。這時候音框視為真正語音的起點，將此音框訂為$N_O$。原因是元音之前若有輔音，不容易從能量看出來，可以藉助越零率來判斷。如果在$N_V$之前，$C$個音框內沒有越零率大於$T_Z$者，就將$N_V$作為真正的語音起點，這表示沒有低能量的輔音在前面。

(6) 從$N_V$之後應該是元音，以後的音框能量大於$T_{EL}$，就是語音存在，一直到能量小於$T_{EL}$，就視為語音結束，語音終點的音框標注為$N_E$。

(7) 從$N_O$或$N_V$到$N_E$之間，就是語音存在的區域。

## 2.4 Support Vector Machine 統計學習原理

### 2.4.1 簡介

本文中所使用到支持向量機(Support Vector Machine)是眾多統計學習機器中的一種方法，其基礎理論由 Vapnik 所發展，目前較常拿來應用於分類。在有限個已知的兩群資料，透過 SVM 的訓練我們可以得到一個新的支持向量機，當有新的資料輸入時，SVM 就可以對其做分類。下面將詳細介紹 SVM 演算法[24]。

### 2.4.2 支持向量機分類法(SVC)

在分類的方法中，主要可分為兩種，一種是將資料分類為數個類別，而另一個則是將資料一分為二，支持向量機的分類則是屬於後者。在兩群已知的資料中，可以使用一個函數將其區隔開來，這個函數我們稱他為分類器，理論上來看此種分類器將會有無限多個，但是擁有最大 margin(兩群中分別最靠近分類器的點的距離)的分類器只有一個，如(圖 2-5)所示。

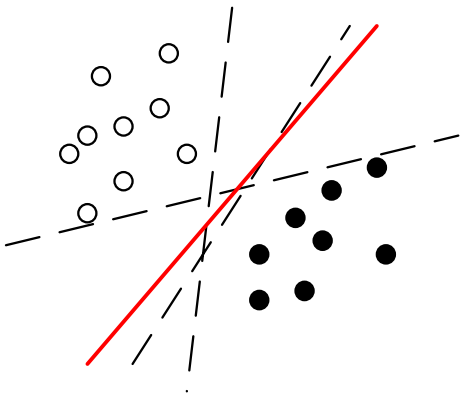

**圖 2-5 最佳化超平面(optimal separating hyperplane)**

我們稱此分類器為最佳化超平面，而 SVM 的訓練過程，也就是找出

此超平面，再使用這個超平面來對新的、未知的資料做分類。

### 2.4.2.1 最佳化超平面(Optimal Separating Hyperplane)

假設我們有一群訓練資料

$$D=\{(x_1,y_1),\ldots,(x_l,y_l)\} \tag{2-17}$$

可以被下式的超平面分為兩類

$$(w\cdot x_i)-b=0 \tag{2-18}$$

當此超平面能夠完全無誤並且使最接近平面的向量離平面的距離為最大，我們就稱之為最佳化超平面。有了最佳化超平面限制了$w$與$b$，接著定義支持向量(support vector)與標準超平面(canonical hyperplane)如下：

$$\begin{cases}(w\cdot x_i)-b=1\\(w\cdot x_i)-b=-1\end{cases} \tag{2-19}$$

在我們的資料群裡符合式子(2-19)的$x_i$，稱為支持向量，而式子(2-19)即代表標準超平面，其幾何概念如(圖 2-6)所示。

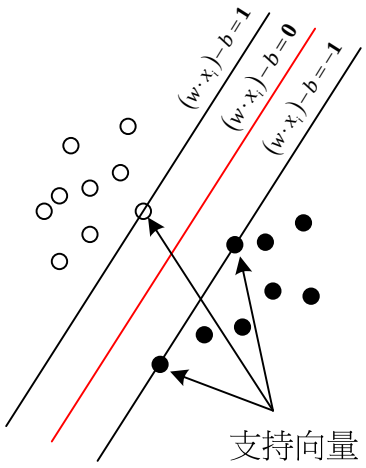


**圖 2-6 標準超平面與支持向量示意圖**

很容易的可以得到

$$\begin{aligned}(w\cdot x_i)-b\geq 1 \quad & \text{if } y_i=1\\ (w\cdot x_i)-b\leq -1 \quad & \text{if } y_i=-1\end{aligned} \tag{2-20}$$

將式(2-20)重新整理導出式(2-21)

$$y_i\left[(w\cdot x_i)-b\right]\geq 1,\ \ i=1,...,l \tag{2-21}$$

從支持向量到最佳化超平面的距離可以表示為式(2-22)

$$d=\frac{\left|(w\cdot x_i)-b\right|}{\|w\|} \tag{2-22}$$

根據式(2-19)與式(2-22)，margin $\rho$就等於

$$\begin{aligned}\rho(w,b)&=2\times d\\ &=\frac{2\times\left|(w\cdot x_i)-b\right|}{\|w\|}\\ &=\frac{2}{\|w\|}\end{aligned} \tag{2-23}$$

所以如果想要得到最大的 margin，就必須對$\|w\|$做最小化處理，此為二次項規劃最優化問題，也就是說將$\|w\|$最小化等於對$\frac{1}{2}\|w\|^2$做最小化處理。因此在式(2-21)的限制條件下，對$\frac{1}{2}\|w\|^2$最小化處理就可以得到最佳化超平面。

引進拉格朗日乘數(Lagrange multipliers)來解決這個極值問題，可以列出式(2-24)

$$L(w,b,\alpha)=\frac{1}{2}(w\cdot w)-\sum_{i=1}^{l}\alpha_i\{[(w\cdot x_i)-b]y_i-1\} \tag{2-24}$$

其中$\alpha_i$為拉格朗日乘數。在極值點將會有一組最佳解$w_0$、$b_0$、$\alpha^0$滿足下列式子

$$\frac{\partial L(w_0,b_0,\alpha^0)}{\partial b}=0 \quad\rightarrow\quad \sum_{i=1}^{1}\alpha_i^0 y_i=0,\quad \alpha_i^0\geq 0,\ \ i=1,...,l \tag{2-25}$$

$$\frac{\partial L(w_0, b_0, \alpha^0)}{\partial w} = 0 \quad \rightarrow \quad w_0 = \sum_{i=1}^{l} y_i \alpha_i^0 x_i, \quad \alpha_i^0 \geq 0, \ i = 1,...,l \tag{2-26}$$

將式(2-25)、式(2-26)代入式(2-24)，可以得到新的拉格朗日函數

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \tag{2-27}$$

在式(2-25)的限制條件下，求式(2-27)的解，就可以得到一組 $\alpha_0 = (\alpha_1^0, \alpha_2^0, ..., \alpha_l^0)$，而根據式(2-26)，此組拉格朗日乘數變決定了最佳化超平面，如下列式子

$$\begin{aligned} w_0 &= \sum_{i=1}^{l} y_i \alpha_i^0 x_i \\ b_0 &= \frac{1}{2}\left[(w_0 \cdot x^*(1)) + (w_0 \cdot x^*(-1))\right] \end{aligned} \tag{2-28}$$

其中 $x^*(1)$ 代表第一個類別裡的任一個支持向量，而 $x^*(-1)$ 代表另一個類別裡的任一個支持向量。

得到了 $w_0$ 與 $b_0$ 便決定了我們的最佳化超平面為 $(w_0 \cdot x) - b_0 = 0$，當我們獲得新的一筆資料，對他做下列的運算

$$f(x) = \text{sign}[(w_0 \cdot x) - b_0] \tag{2-29}$$

便可以知道此筆新資料屬於哪個類別。

# 第三章 聲音監控方法

## 3.1 高斯混合模型

### 3.1.1 高斯混合模型簡介

高斯混合模型(GMM)是在聲音監控(Audio Surveillance)研究上一種常用來建立背景聲音模型的方法[17][21][22]。

以數學的觀點來看，對任一具有多類似的樣本(Pattern)而言，高斯混合模型具有極佳的近似能力，與傳統的單一高斯分佈(Single Gaussian Mixture)及向量量化(Vector Quantization)兩種模型比較，單一高斯分佈模型，僅能用一個平均值向量來代表一堆樣本在向量空間的中心位置，用共變異矩陣來近似這些樣本在空間中分佈的形狀，其效果當然不好。而向量量化的模型，是用幾個重要的位置來代表整個向量空間，但模型本身並沒有把這些樣本在空間中的分佈大小，形狀描述出來，因此此種方法也不理想。而高斯混合模型使用多個高斯來代表特徵向量的分佈，以數學的觀點來看，它不但精準地記錄樣本的各種類別、在向量空間中的位置，也能描述出這些類別在空間中的大小及形狀。因此，高斯混合模型適合描述特徵向量在聲音空間的分佈。

在採用高斯混合模型時，有一點要注意，假設我們所求取的特徵向量的每一個維度在統計上是互相獨立(Statically Independent)的關係，即 8 顆麥克風的分佈是各自獨立的，所以全共變異矩陣(Full Covariance Matrix)是不需要的，對角共變異矩陣(Diagonal Covariance Matrix)的高斯分佈的線性組合，就具有描述特徵向量維度間的相關能力；做此假設的另一個原因，是可以降低計算時的複雜度，因此在本論文中，高斯混合模型的共變異矩陣皆是對角矩陣。

### 3.1.2 模型描述

一個高斯混合模型具有三個參數，分別是混合加權值(mixture weights)、平均值向量(mean vector)以及共變異矩陣(covariance matrix)，將這這些參數集合起來並賦予新的符號，並稱之為成分密度，如下所示：

$$\lambda = \{w_i, \mu_i, \Sigma_i\}, i = 1, 2, ..., M \tag{3-1}$$

其中 $w_i$ 表示混合加權值、$\mu_i$ 表示平均值向量以及 $\Sigma_i$ 表示共變異矩陣，$M$ 則是高斯分佈的個數。若我們的資料 $X_N = \{x_1, x_2, \cdots, x_n\}$ 為在 d 維空間中的分佈，則高斯函數可如下所示：

$$p(x_N \mid \lambda) = \sum_{i=1}^{M} w_i g_i(x_N) \tag{3-2}$$

$$g_i(x) = \frac{1}{(2\pi)^{d/2} \left|\Sigma_i\right|^{1/2}} \cdot \exp[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)]$$

其中 $g(x; \mu_i, \Sigma_i)$ 為第 i 個高斯分佈函數，而混合加權值也必須滿足 $\sum_{i=1}^{M} w_i = 1$ 的條件。我們可以將高斯混合模型的架構用(圖 3-1)來表示。

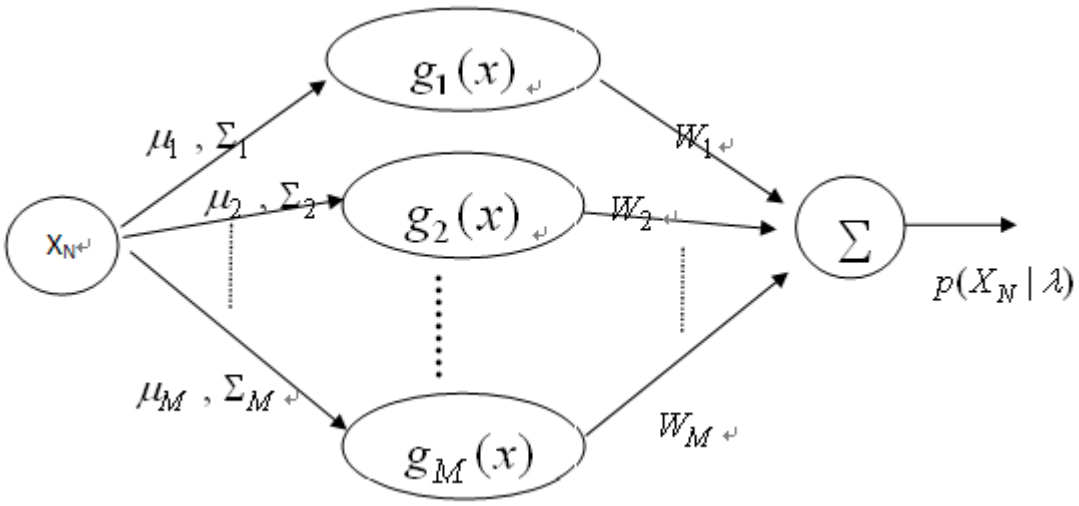


**圖 3-1 高斯混合模型架構圖**

### 3.1.3 模型參數的初始化

現在想找出高斯混合模型的最佳參數，使得系統有最佳的表現，則在尋找最佳參數之前必須對參數作初始化的動作。向量量化(VQ)是一項運用非常廣泛的技術能將一堆特徵向量的資料，濃縮成幾個具代表性的類別(class)或群集(cluster)，所以這裡我們先採用 VQ 的技術，將我們得到聲音頻率分佈，作初步的分群，得到高斯混合模型參數的初始化值(群的中心)。向量量化的方法有很多種，在此採用 K 平均值分類法(K-means Cluster)，其流程如(圖所示，詳細的步驟說明如下：

0、 收集資料：

經過一段時間的收集，獲得 N 個欲做訓練的特徵向量。

1、 初始化：

設一開始的群數是 K，並隨機地取 K 個向量當成每群的中心點。

2、 以新的群中心來分群：

其它(N-K)個向量對這個 K 個群中心作距離量測，以距離作為分群的依據，每個向量被分類到距離最短的中心。

3、 更新群中心：

接著對每一群算出新的向量平均值，以此作為新的群中心。

4、 判斷分群是否收斂：

將新的群中心與舊的群中心作比較，如果不再有變動，表示已收斂。則做步驟 5；反之，則重複步驟 2、3。

5、 得到初始化的參數：

將最後分群的中心，當作高斯混合模型的初始參數

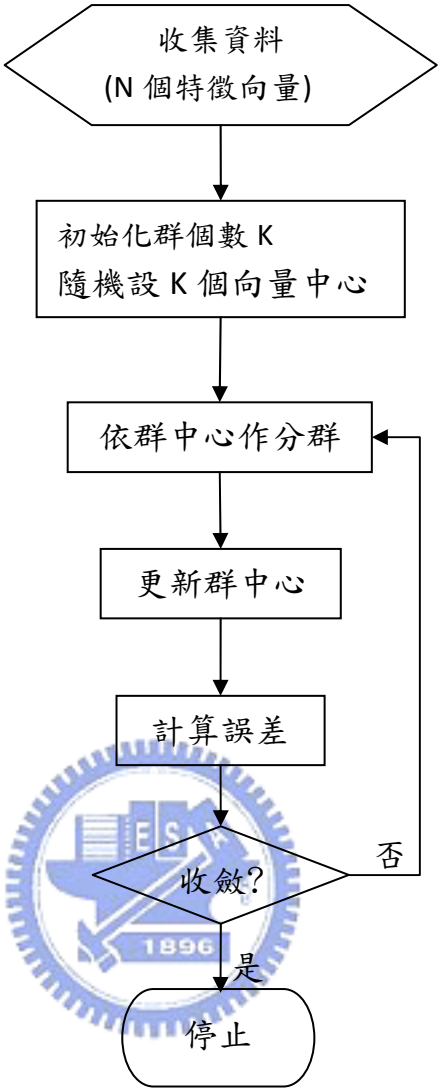


**圖 3-2 K-means 流程圖**

## 3.1.4 期望值最大演算法(Expectation Maximization, EM)[18]

我們在做模型訓練時，最終的目的是估測高斯混合模型參數，使得頻率的分佈與模型參數估測出來的分佈有相當的相似度，估測最佳參數的方法有很多種，但最受歡迎且適合的方法是最大相似性估測法( Maximum

Likelihood Estimation, MLE)。

在高斯密度函數的假設下，當 $x = x_i$ 時，其機率密度為 $P(x_i \mid \lambda)$ ，如果 $x_i, i = 1 \cdots n$ 之間是互相獨立的事件，則發生 $X = \{x_1, x_2, \cdots, x_n\}$ 的機率密度相似函數(likelihood function)可以表示成：

$$P(X \mid \lambda) = \prod_{i=1}^{n} P(x_i \mid \lambda) \qquad (3\text{-}3)$$

由於 $X$ 是確定的，因此 MLE 主要就是找出使得高斯混合模型的相似函數值為最大時的參數 $\lambda'$，也就是 $\lambda' = \arg\max_{\lambda} P(X \mid \lambda)$ ，但(3-3)式對 $\lambda$ 而言是一個非線性的方程式，無法直接最大化相似函數，所以我們採用期望值最大演算法(Expectation Maximization Algorithm)，利用疊代的方式找出 MLE 的估測參數 $\lambda'$。

EM 演算法的基本作法是由 K-means 分類法找出的初始化的參數，再利用 EM 估計出新的參數 $\bar{\lambda}$ ，使得滿足 $P(X \mid \bar{\lambda}) \geq P(X \mid \lambda)$ ，再另 $\lambda = \bar{\lambda}$ 重新疊代估計新的 $\bar{\lambda}$ ，直到 $P(X \mid \lambda)$ 收斂或是達到某個門檻值才停止。EM 演算法主要分成兩部分，分別是與 likelihood 函數有關的 E-Step，以及更新參數方程式的 M-Step：

**E-Step**

目的是測試我們所求的 likelihood 函數值，是否達到我們的要求，若符合要求，EM 演算法就停止，反之就繼續執行 EM 演算法。這裡為了數學推導的方便，假設我們的模型是由三個高斯分佈函數所構成，則其密度函數可表示成：

$$P(x) = w_1 g(x; \mu_1, \Sigma_1) + w_2 g(x; \mu_2, \Sigma_2) + w_3 g(x; \mu_3, \Sigma_3) \qquad (3\text{-}4)$$

其中共變異矩陣 $\Sigma_i$，因為假設每個維度彼此獨立，所以只剩對角有值，P(x)

的參數 $\lambda=[w_1,w_2,w_3,\mu_1,\mu_2,\mu_3,\Sigma_1,\Sigma_2,\Sigma_3]$，參數個數為 (1+1+1+d+d+d+d+d+d)=3+6d 個，依前述 MLE 原則，求出 likelihood 的最大值：

$$E(\lambda)=\ln\left(\prod_{i=1}^{n}P(x_i)\right)$$
$$=\sum_{i=1}^{n}\ln\left(P(x_i)\right)=\sum_{i=1}^{n}\ln\left[w_1g(x;\mu_1,\Sigma_1)+w_2g(x;\mu_2,\Sigma_2)+w_3g(x;\mu_3,\Sigma_3)\right] \quad (3\text{-}5)$$

為了簡化討論，再引進另一個數學符號稱為事後機率(posterior probability)：

$$\beta_j(x)=p(j\mid x)=\frac{p(j\cap x)}{p(x)}=\frac{p(j)p(x\mid j)}{p(x)}$$
$$=\frac{p(j)p(x\mid j)}{p(1)p(x\mid 1)+p(2)p(x\mid 2)+p(3)p(x\mid 3)}$$
$$=\frac{w_jg(x;\mu_j,\Sigma_j)}{w_1g(x;\mu_1,\Sigma_1)+w_2g(x;\mu_2,\Sigma_2)+w_3g(x;\mu_3,\Sigma_3)} \quad (3\text{-}6)$$

**M-Step**

主要目的是為了要找到使 likelihood 函數最大化的參數，因此我們分別對 $w_i$、$\mu_i$、$\Sigma_i$ 做偏微分，再做後續的運算，於是我們可以得到所求的參數，接著返回 E-Step 繼續做。

假設初始參數是 $\lambda_{old}$，我們希望找出新的 $\lambda$ 值，滿足 $E(\lambda)>E(\lambda_{old})$，因為根據 $\ln(\frac{a}{b})=\ln(a)-\ln(b)$，$E(\lambda)-E(\lambda_{old})$ 可以延伸成下式：

$$
\begin{aligned}
&E(\lambda)-E(\lambda_{old}) \\
&=\sum_{i=1}^{n}\ln\left[\frac{w_1 g(x_i;\mu_1,\Sigma_1)+w_2 g(x_i;\mu_2,\Sigma_2)+w_3 g(x_i;\mu_3,\Sigma_3)}{w_{1,old}\, g(x_i;\mu_{1,old},\Sigma_{1,old})+w_{2,old}\, g(x_i;\mu_{2,old},\Sigma_{2,old})+w_{3,old}\, g(x_i;\mu_{3,old},\Sigma_{3,old})}\right] \\
&=\sum_{i=1}^{n}\ln\left[\frac{w_1 g(x_i;\mu_1,\Sigma_1)}{D(\lambda_{old})}\frac{\beta_1(x_i)}{\beta_1(x_i)}+\frac{w_2 g(x_i;\mu_2,\Sigma_2)}{D(\lambda_{old})}\frac{\beta_2(x_i)}{\beta_2(x_i)}+\frac{w_3 g(x_i;\mu_3,\Sigma_3)}{D(\lambda_{old})}\frac{\beta_3(x_i)}{\beta_3(x_i)}\right] \\
&\geq\sum_{i=1}^{n}\left[\beta_1(x_i)\ln\frac{w_1 g(x_i;\mu_1,\Sigma_1)}{D(\lambda_{old})\beta_1(x_i)}+\beta_2(x_i)\ln\frac{w_2 g(x_i;\mu_2,\Sigma_2)}{D(\lambda_{old})\beta_2(x_i)}+\beta_3(x_i)\ln\frac{w_3 g(x_i;\mu_3,\Sigma_3)}{D(\lambda_{old})\beta_3(x_i)}\right] \\
&=Q(\lambda)
\end{aligned}
$$

(3-7)

上式中，因為 $\ln(x)$ 是一個凸函數(Convex Function)，滿足下列不等式：

$$\ln\left[\alpha x_1+(1-\alpha)x_2\right]\geq\alpha\ln(x_1)+(1-\alpha)\ln(x_2) \tag{3-8}$$

推廣上式到「傑森不等式」(Jensen Inequality)：

$$\ln\left(\sum_{i=1}^{n}\alpha_i x_i\right)\geq\sum_{i=1}^{n}\alpha_i\ln(x_i),\sum_{i=1}^{n}\alpha_i=1 \tag{3-9}$$

因為 $\sum_{j=1}^{3}\beta_j\ln(x_i)=1$，所以可以將傑森不等式套用在(3-7)式，最後得到下式：

$$E(\lambda)\geq E(\lambda_{old})+Q(\lambda) \tag{3-10}$$

只要 $Q(\lambda)>0$，必滿足 $E(\lambda)>E(\lambda_{old})$，但我們通常希望 $E(\lambda)$ 越大越好，最直接的方式就是找出使得 $Q(\lambda)$ 最大的 $\lambda$ 值，那 $E(\lambda)$ 也會跟著變大。

$Q(\lambda)$ 是 $\lambda$ 的函數，將一些與 $\lambda$ 不相關的部份並入常數項，並重新整理 $Q(\lambda)$ 成下式：

$$
\begin{aligned}
Q(\lambda)&=\sum_{i=1}^{n}\sum_{j=1}^{3}\beta_j(x_i)\left[\ln w_j+\ln g\left(x_i;\mu_j,\sum\nolimits_j\right)\right]+c1 \\
&=\sum_{i=1}^{n}\sum_{j=1}^{3}\beta_j(x_i)\left\{\ln w_j+\ln\left[\frac{1}{(2\pi)^{d/2}\left[\det\sum_j\right]^{1/2}}\exp\left(-\frac{(x_i-\mu_j)\sum_j^{-1}(x_i-\mu_j)^T}{2}\right)\right]\right\}+c1
\end{aligned}
$$

對 $\mu_j$ 偏微分，$\partial_{\mu_j} Q = 0 \Rightarrow \mu_j = \dfrac{\sum_{i=1}^{n} \beta_j(x_i) x_i}{\sum_{i=1}^{n} \beta_j(x_i)}$

(3-11)

對 $\sum_j$ 偏微分，$\partial_{\sum_j} Q = 0 \Rightarrow \sum_j = \dfrac{\sum_{i=1}^{n} \beta_j(x_i)(x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^{n} \beta_j(x_i)}$

(3-12)

欲得到最佳之 $w_j$ 值，需將 $w_j$ 的總和為 1 的條件加入，引進 Lagrange Multiplier，並定義新的目標函數(object function)為：

$$E_{new}(\lambda) = E(\lambda) + \alpha(w_1 + w_2 + w_3 - 1) \tag{3-13}$$

將 $E_{new}$ 對 3 個 weighting 做偏微分，可得到下面方程式：

$$\frac{\partial E_{new}}{\partial w_j} = -\frac{1}{w_j}\sum_{i=1}^{n} \beta_j(x_i) + \alpha = 0, j = 1,2,3 \tag{3-14}$$

最後將(15)的 3 個不同 $j$ 的式子相加，可得到：

$$(w_1 + w_2 + w_3)\alpha = -\sum_{i=1}^{n}\left[\beta_1(x_i) + \beta_2(x_i) + \beta_3(x_i)\right]$$

$$\text{and let } \alpha = -\sum_{i=1}^{n} 1 = -n \tag{3-15}$$

$$\Rightarrow w_j = \frac{1}{n}\sum_{i=1}^{n} \beta_j(x_i), j = 1,2,3$$

## 3.1.5 GMM 建立的流程

綜合前面各小節的說明，GMM 建立的流程如圖所示，先將 N 個準備

訓練的資料點，經過 K-means 分類後得到初始的參數，再由 EM 演算法得到的三個方程式：

$$\mu_j = \frac{\sum_{i=1}^{n} \beta_j(x_i) x_i}{\sum_{i=1}^{n} \beta_j(x_i)}, \qquad \Sigma_j = \frac{\sum_{i=1}^{n} \beta_j(x_i)(x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^{n} \beta_j(x_i)}, \qquad w_j = \frac{1}{n}\sum_{i=1}^{n} \beta_j(x_i)$$

進行參數的更新，並計算新的函數值，如此一一的疊代，更新模型的參數，直到相似函數的值與前次相似函數的值差異小於某個 threshold 才停止疊代。

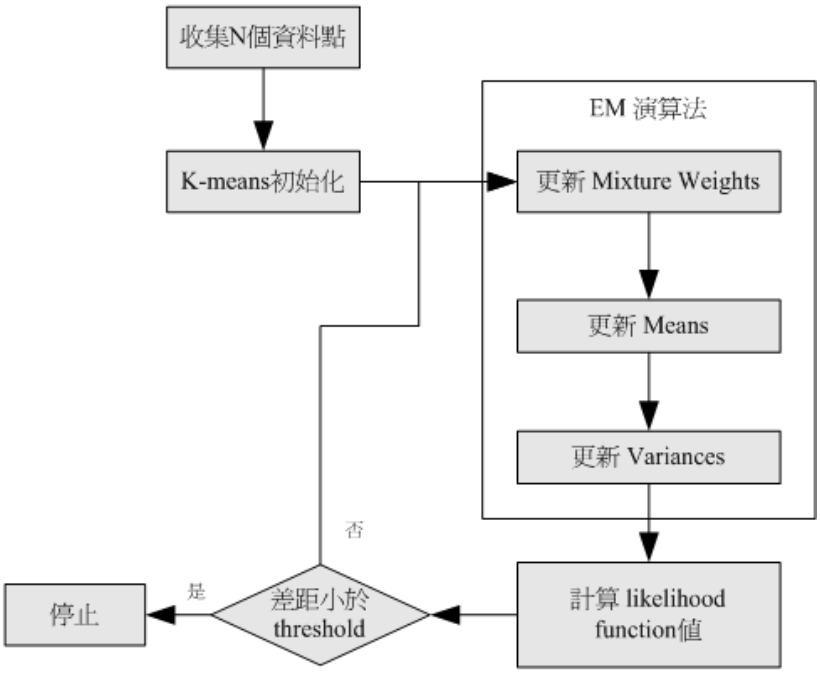


**圖 3-3 高斯混合模型建立流程圖**

## 3.2 現行語音活動偵測所面臨問題

一般傳統所使用的語音活動偵測使用越零率和能量當作門檻[25][26]，將收錄到的聲音切成一序列的音框，之後最每一音框取越零率和能量大小，接著判斷越零率和能量的值是否大於所設定的值，如果大於門檻就判斷發生語音，如果沒有大於門檻的值就判斷沒有發生語音。根據實驗結果，在訊噪比大的情況下可以準確的判斷發生語音的時間點。但是，當訊噪比低的情況下會出會有語音的部份沒有被判斷為語音，或者是屬於環境背景的音框被判斷為語音。

## 3.3 建立聲音監控模型

### 3.3.1 語音前處理和特徵值擷取

在本論文是使用一次使用 8 顆麥克風接收聲音訊號。當訊噪比很低時，我們所監控的聲音和其環境背景聲音在人耳聽起來都不能輕易辨識。先使用空間濾波器對我們所監控的聲音方向做聲源訊號的放大。首先我們假設對想要針對多方向做聲音監控，所以使用麥克風陣列得到多方向的 Beamformer。現在假設經由麥克風陣列產生出 n 個 Beamformer，之後對每個 Beamformer 取越零率。越零率因為是計算聲音通過零點的個數，在論文中每 512 點為一個音框。實驗上就是求在每一個音框內該聲音的頻率。每個 Beamformer 經過特徵的擷取也就是計算完越零率後，利用每個 Beamformer 所得到的統計資訊利用高斯混合模型理論建立高斯混合模型。

### 3.3.2 異音監測方法架構[15]

本論文對於異音監測可以分為兩部分。第一部分是假設異音出現在麥克風陣列所監控的角度上。第二部分是異音不在我們所假設的角度上。首先(圖 3-4)表示異音出現在所監控的角度上:

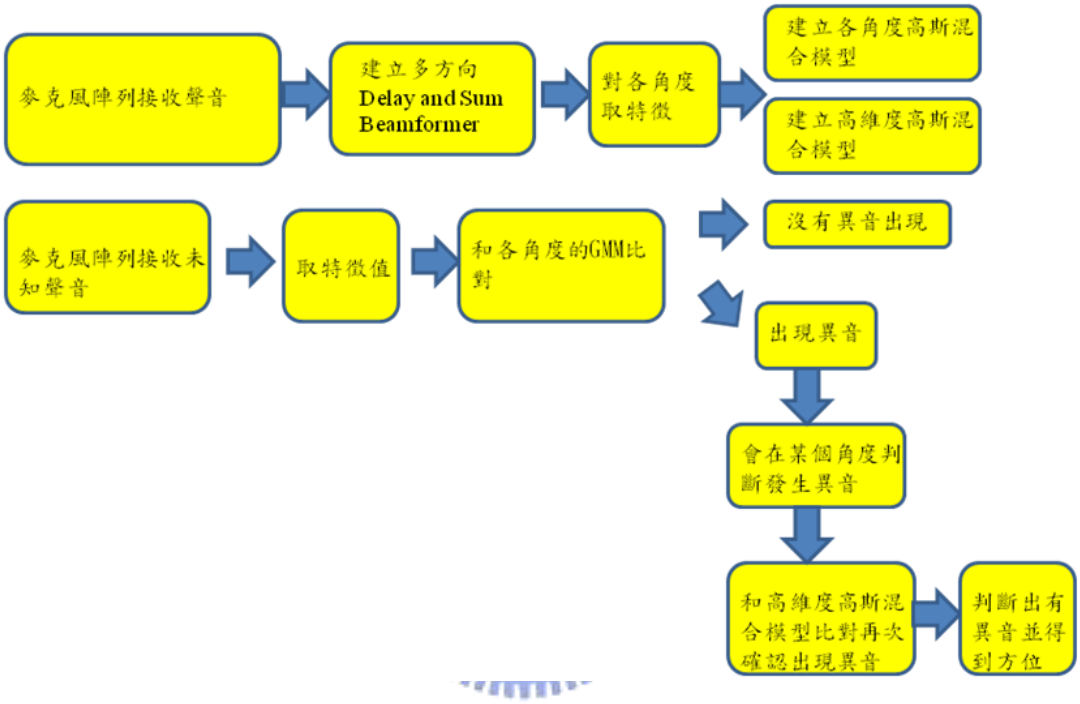


**圖 3-4 系統流程一**

假設麥克風陣列的麥克風個數是 8，可以根據(2-8)式對某一方向做聲音的監控。在本篇論文中是假設以麥克風陣列前方 180 度，以每 22.5 度依序是(22.5 度、45 度、67.5 度、90 度、112.5 度、135 度、157.5 度、180 度)8 個角度做聲音監控。以 90 度為例，因為是平面波假設。根據(2-6)可以得知 8 顆麥克風之間的延遲時間為零，所以 8 顆麥克風是同時收到訊號。所以對 90 度方向的 Beamformer 就是將 8 顆麥克風收到的訊號相加之後再取平均。這是(系統流程圖一)中建立多方向 Delay and Sum Beamformer 的部分。

建立好 Delay and Sum Beamformer 之後，會得到 8 組方向的聲音資料。

一樣還是以 90 方向的 Beamformer 為例，可以根據(2-19)對 90 度的 Beamformer 聲音資料當作$x(n)$，以每 512 點為一音框所以是計算音框內的通過零點的次數。所以 90 方向的 Beamformer 的資料計算完越零率之後會得到一組數據，這筆數據代表在這段時間內每 512 點聲音的頻率變化。

想要對這筆數據建立高斯混合模型可以利用 K-means 分群的方法得到這筆數據可以分的群數和群中心點的值加快其收斂速度。利用 k-means 得到初始值，再由期望值最大演算法求得代表 90 方向的 Beamformer 的聲音頻率分佈情形。

論文中監控 8 個角度，依序求得 8 個監控方向各自的聲音頻率分佈情況，所以就有 8 個高斯混合模型。現在訓練好環境的背景模型後，依照(系統流程圖一)。現在假設異音出現在 90 度的方向，當麥克風陣列再次收到一組聲音後，根據 Delay and Sum Beamformer 和(2-19)會求得 8 組方向的聲音資料，之後代入各自監控角度的高斯混合模型。現在假設異音出現在 90 度的方向，所以 Delay and Sum Beamformer 會將 90 度異音的訊號放大。放大之後對於高斯混合模型就會有較好的辨識能力。判斷是否為異音判斷的方法是收到的資料需落在高斯混合模型的每一個高斯分佈的 2.5 倍的標準差以內。根據統計推論，發生在 2.5 倍的標準差之外的機率值發生的不到百分之五。所以認定此聲音不屬於此分佈，也就不是環境的背景聲音。其他角度因為異音不在所監控的角度上，所以異音會被抑制，辨識的效就不佳。

根據(表 3-1)當有異音出現在 90 度方向時，表中的 8 個監控角度各自計算出來的機率值，其中有 6 個角度的事後機率值都大於 2.5 倍標準差的機率值，所以都把異音判斷為環境的背景聲音。只有 90 度和 135 度判斷為異音，所以就知道異音出現在大約 90 度到 135 度的方向。

(表 3-2)是將 8 個監控角度建立一個高維的高斯混合模型，判斷是否有

發生異音的方法和上述一樣。由實驗可得知，將監控方向一起建立高維度高斯混合模型只要發生的異音，利用麥克風陣列所收到的聲音，即使微量的頻率變化能可辨識出來。但高維度的高斯混合模型就無法知道方向，但可以由(表 3-1)的方法得知方向。

| | 機率值 | 判斷的機率值 |
|---|---|---|
| 22.5 度 | 0.0122944608409649 | 0.00118979577102284 |
| 45 度 | 0.00578747663259773 | 0.00108626085249277 |
| 67.5 度 | 0.00326971304910512 | 0.00137091506324151 |
| 90 度 | 0.000296154795649129 | 0.0014976867597792 |
| 112.5 度 | 0.00277923363716396 | 0.00117718514408275 |
| 135 度 | 0.000678437091262902 | 0.00171458881773066 |
| 157.5 度 | 0.0104770571235841 | 0.00151854383892775 |
| 180 度 | 0.00331716029086139 | 0.00137528519985491 |

**表 3-1 異音位於監控方向的判斷**

| 機率值 | 判斷的機率值 |
|---|---|
| 1.65713387434465e-094 | 2.44685719600539e-054 |

**表 3-2 高維度高斯混合模型對異音的判斷**

(圖 3-5)另外一種情況表示異音沒有在所監控的角度上:

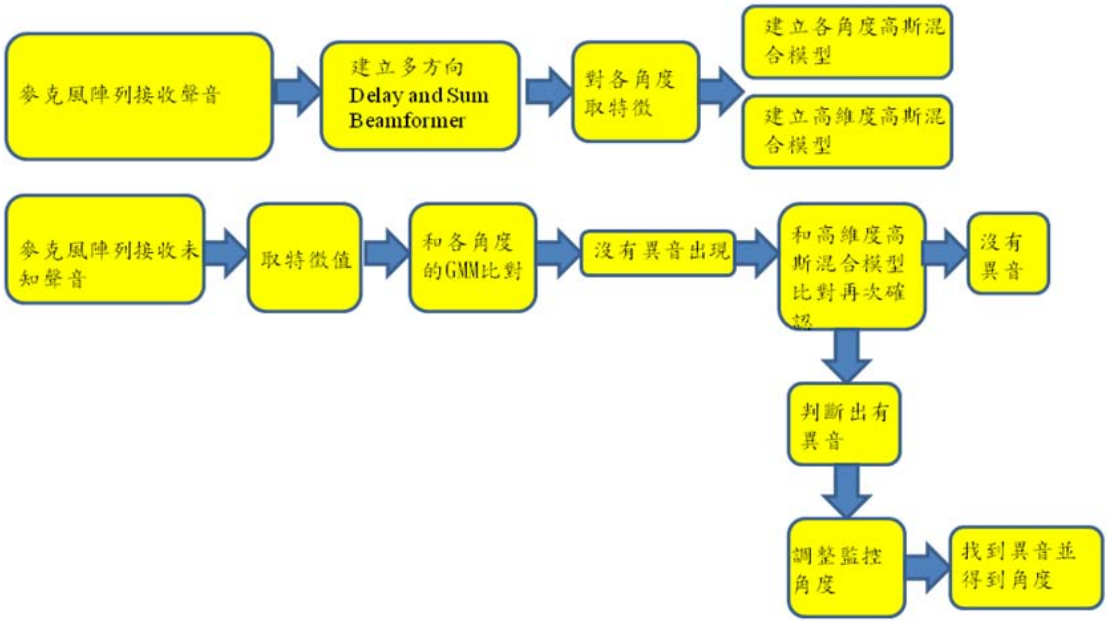


**圖 3-5 系統流程二**

系統流程二說明的是假設聲源一開始不在所監控的角度上。當麥克風陣列收到聲音後，一樣會各自和一開始所監控的高斯混合模型比對。但是異音不在監控的角度上所以在監控角度上的高斯模型都無法判別出來，由(表3-3)可看出當有異音的資料進入到8個監控角度時，都判斷環境背景聲音。

| | 機率值 | 判斷的機率值 |
|---|---|---|
| 22.5 度 | 0.0165733731334948 | 0.00118979577102284 |
| 45 度 | 0.0133653890710188 | 0.00108626085249277 |
| 67.5 度 | 0.0136452102442473 | 0.00137091506324151 |
| 90 度 | 0.0147645098870454 | 0.0014976867597792 |
| 112.5 度 | 0.0159753304010597 | 0.00117718514408275 |
| 135 度 | 0.0127393977987231 | 0.00171458881773066 |
| 157.5 度 | 0.0116954210058476 | 0.00151854383892775 |
| 180 度 | 0.01447470927028 | 0.00137528519985491 |

**表 3-3 異音不在監控方向的判斷**

但是由(表 3-4)判斷的結果可以看出是有異音出現的。

| 機率值 | 判斷的機率值 |
|---|---|
| 6.69184101280151e-218 | 2.44685719600539e-054 |

**表 3-4 高維度高斯混合模型對不在監控方向的異音判斷**

所以在系統流程二中如果異音沒有出現在監控的角度上，但是高維度的高斯混合模型仍可以判斷出來有異音的出現。如果再移動監控的角度，重覆判斷直接找到異音大約的方位。

# 第四章 實驗結果與分析

本系統的實際成品圖如(圖 4-1)所示，硬體上包括了:

1. 麥克風陣列
2. FPGA
3. 數位式麥克風陣列連接 IO 板
4. EZ-USB FX 平台

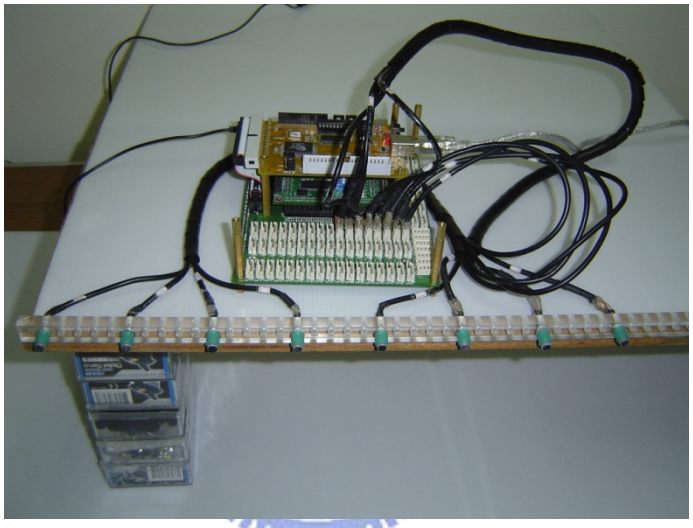

**圖 4-1 實驗平台實施照片**

## 4.1 實驗平台架構簡介

數位式麥克風陣列聲音訊號擷取系統主要是由數位式麥克風、壓克力線性陣列、IO 板與 FPGA 共四部分硬體所構成，以下分別介紹:

■ 數位式麥克風

數位式麥克風為實驗室自行研發設計，交由音賜公司量產。(圖 4-2)中可見數位麥克風有 4 根腳位，其中時脈使用 1.2MHz，輸出 1-bit 的數位訊號。

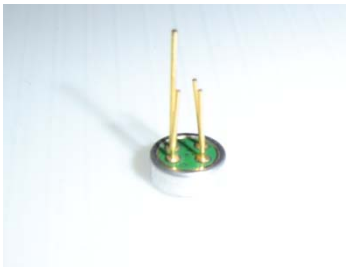

**圖 4-2 數位麥克風實際成品圖**

■ 壓克力線性陣列

圖 4-3 為壓克力線性陣列的實體(圖 4-3)。

**圖 4-3 線性陣列實體圖**

■ 數位式麥克風陣列連接 IO 板

實驗室自行設計一套 I/O 板供數位麥克風陣列作應用。此板提供 76 個數位式麥克風插槽。I/O 板負責連接 FPGA 單板，透過 FPGA 程式撰寫，可與不同介面平台作溝通。I/O 單板提供三種 regulator 的介面，將 5V 轉成 1.8V 供數位式麥克風陣列使用(圖 4-4)。

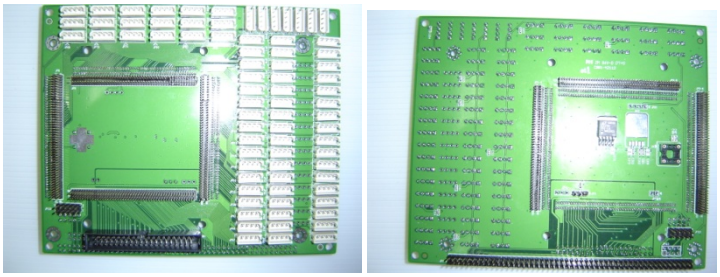

圖 4-4 I/O 板實際成品

■ FPGA 硬體

FPGA 使用的是 ALTERA Cyclone II 系列的 EP2C35F484C6N 晶片，由茂綸公司所開發的實驗板。尺寸為 11cm X 8cm(圖 4-5)。

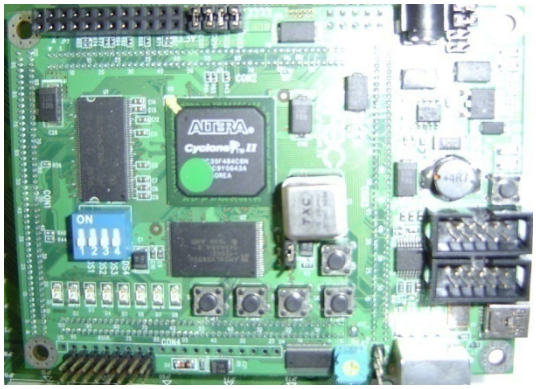

**圖 4-5 GFEC Cyclone II Strarter Kit**

實驗中 SNR 的計算方式如下：

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

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

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

## 4.2 實驗結果

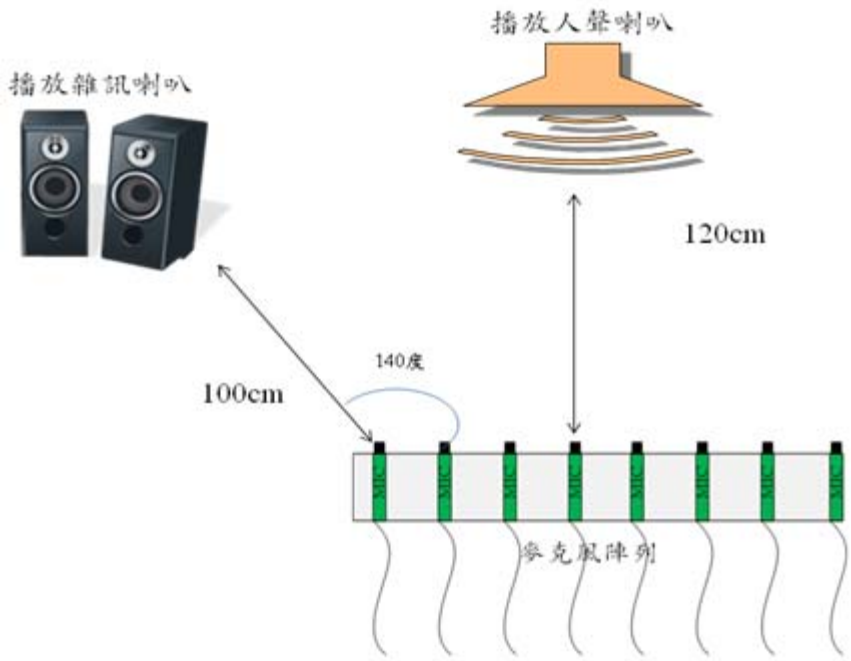


**圖 4-6 實驗環境平面關係圖**

(圖 4-6)實驗環境為在麥克風陣列左邊 100 公分 140 度播放白雜訊喇叭，距離麥克風陣列 120 公分播放語音。

### 4.2.1 空間濾波器的實驗結果

利用麥克風陣列建立空間濾波器，加強語音的特徵。現在就是實驗在各種角度 SNR 提升的程度。

**測試一:語音位於麥克風陣列 90 度方向播放「交通大學工五館 905」與雜訊位於麥克風陣列 140 度方向**

語音播放「交通大學工五館 905」重覆出現三次與雜訊之混合訊號，用單一麥克風錄到情形(圖 4-7):

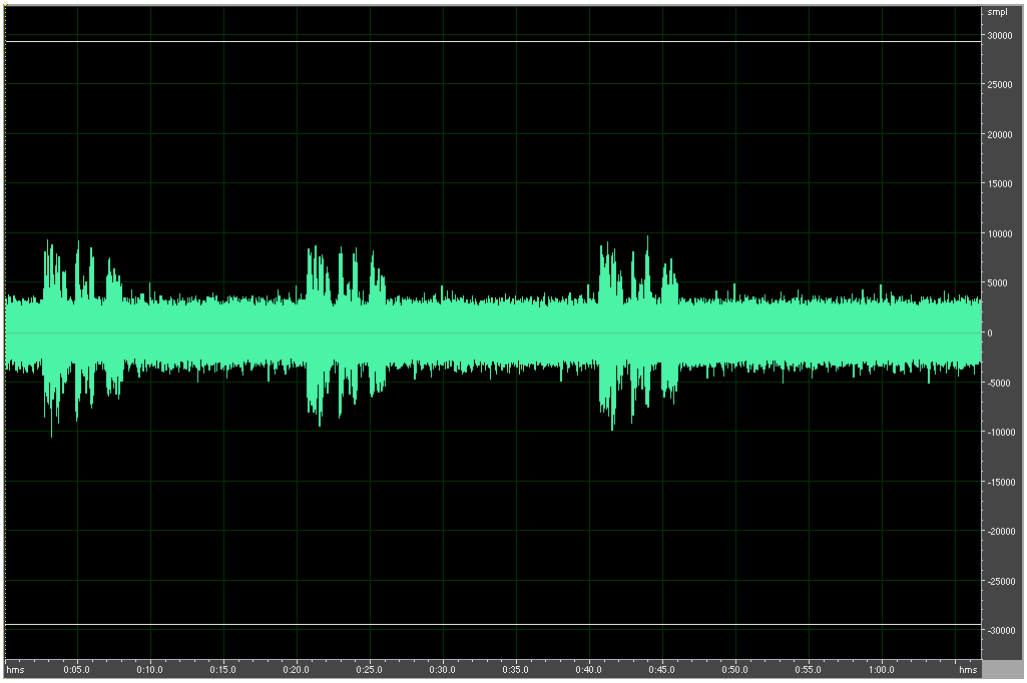


**圖 4-7 麥克風於測試一實驗接收的聲音訊號**

白雜訊能量為-30.08dB，而語音「交通大學工五館 905」與雜訊混合部份的能量為-21.67dB，SNR=8.41dB。

經過空間濾波器對 90 度處理的結果(圖 4-8):

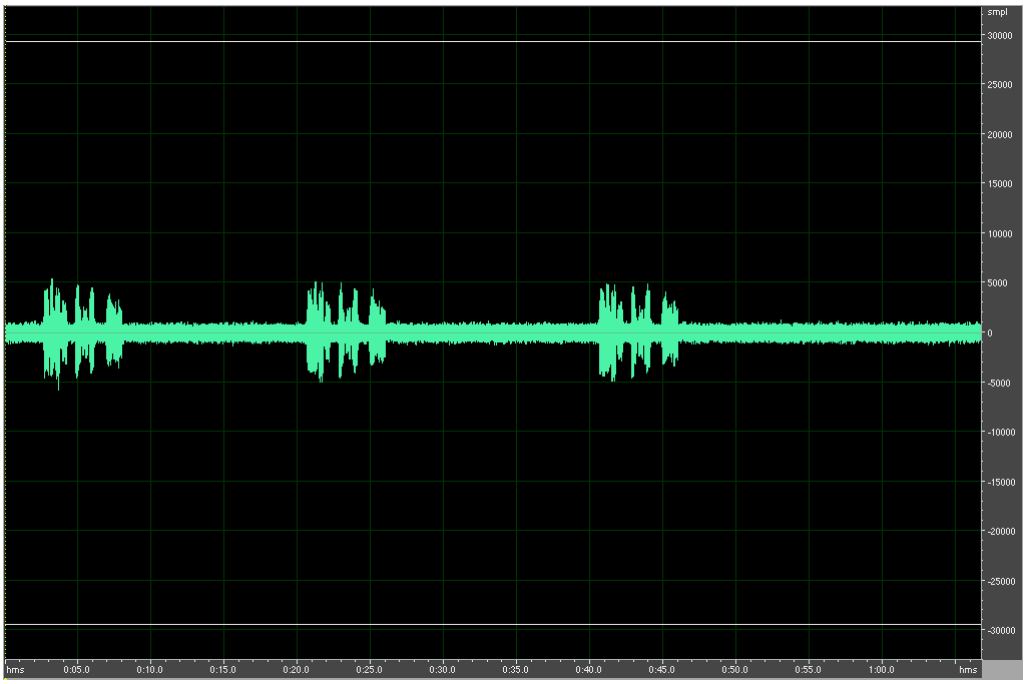


**圖 4-8 通過空間濾波器的處理結果**

(圖 4-8)中，白雜訊能量為-41.39Db，而語音「交通大學工五館 905」與雜訊混合部份的能量為-25.84dB，因此 SNR=15.55dB。

**測試一總結**:

通過空間濾波器的濾波作用，SNR 由原本的 8.41dB 提升到 15.55dB，其 SNR 增加了 7.14dB。

**測試二:語音位於麥克風陣列 45 度方向播放「交通大學工五館 905」與雜訊位於麥克風陣列 140 度方向**

語音播放「交通大學工五館 905」重覆出現三次與雜訊之混合訊號，用單一麥克風錄到情形(圖 4-9):

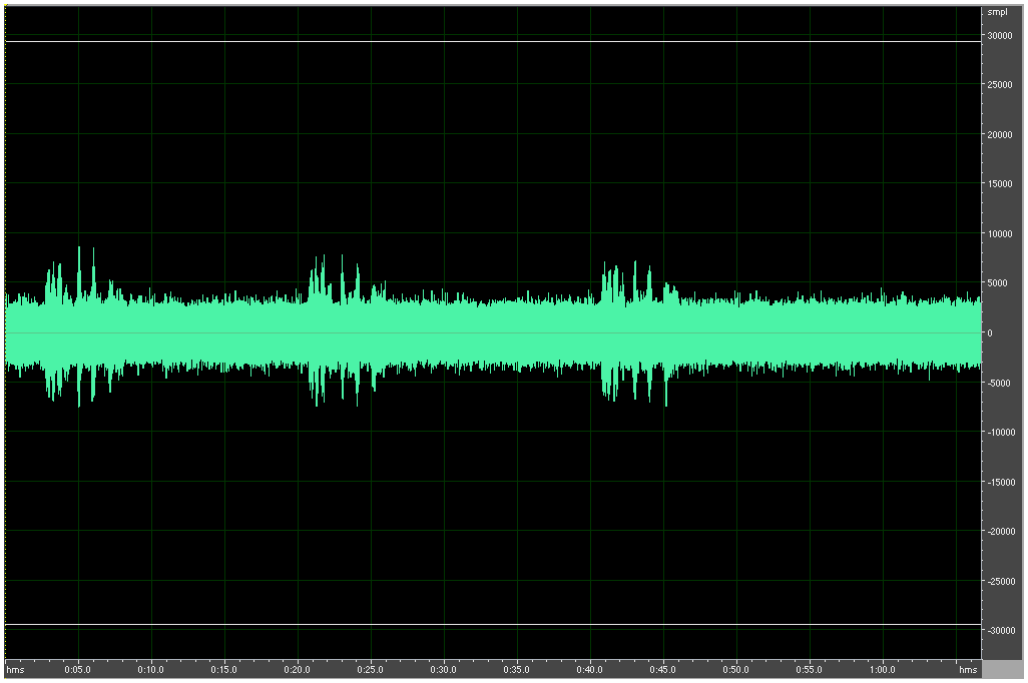


**圖 4-9 麥克風於測試二實驗接收的聲音訊號**

(圖 4-9) 白雜訊能量為-30.12dB，而語音「交通大學工五館 905」與雜訊混合部份的能量為-23.72dB，因此 SNR=6.4dB。

經過空間濾波器對 45 度處理的結果(圖 4-10):

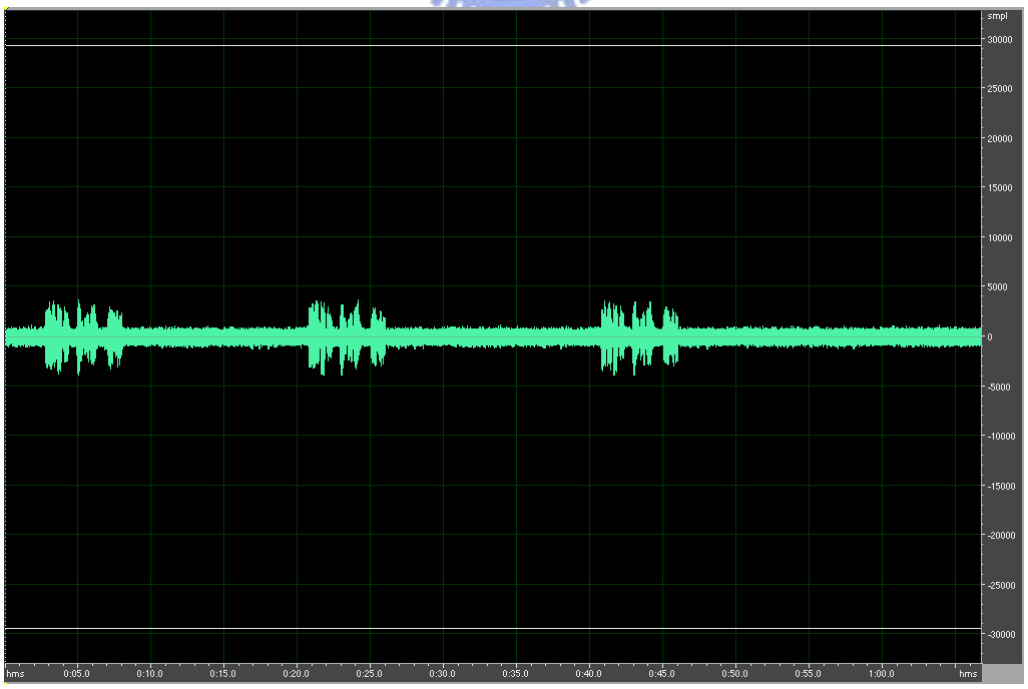


**圖 4-10 通過空間濾波器的處理結果**

(圖 4-10)中，白雜訊能量為-41.47dB，而語音「交通大學工五館 905」與雜訊混合部份的能量為-29.17dB，因此 SNR=12.3dB。

**測試二總結**:

通過空間濾波器的濾波作用，SNR 由原本的 6.4dB 提升到 12.3dB，其 SNR 增加了 5.9dB。經過實驗可發現空間濾波器可以提高 SNR

## 4.2.2 高維度高斯混合模型對於異音監控結果

現在實驗環境中有兩個喇叭，一個喇叭用來播放語言，另一個播放雜訊。並在下列兩種情況下播放測試其辨識效果:

1. 在不同 SNR 的語音環境下
2. 語音在不同角度辨識的結果

**測試一:語音位於麥克風陣列 90 度播放「交通大學工五館 905」與雜訊位於麥克風陣列 140 度方向(SNR=4.14dB)**

麥克風收到的聲音(圖 4-11):

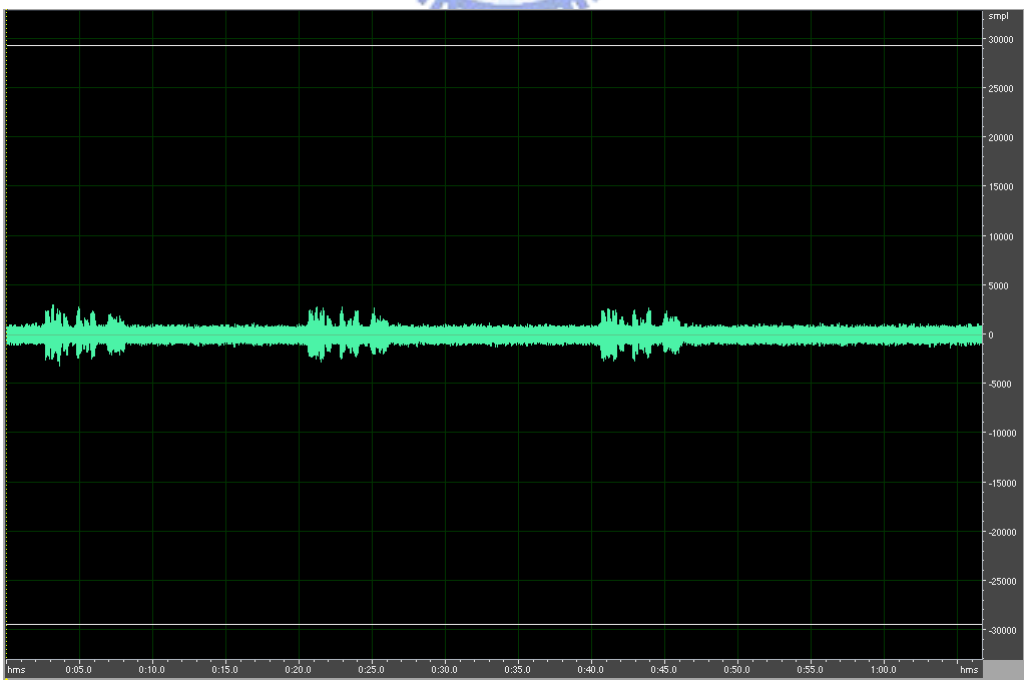


**圖 4-11 麥克風於實驗接收的聲音訊號**

SNR=4.14dB 經過高斯混合模型所辨識的結果(圖 4-12)：

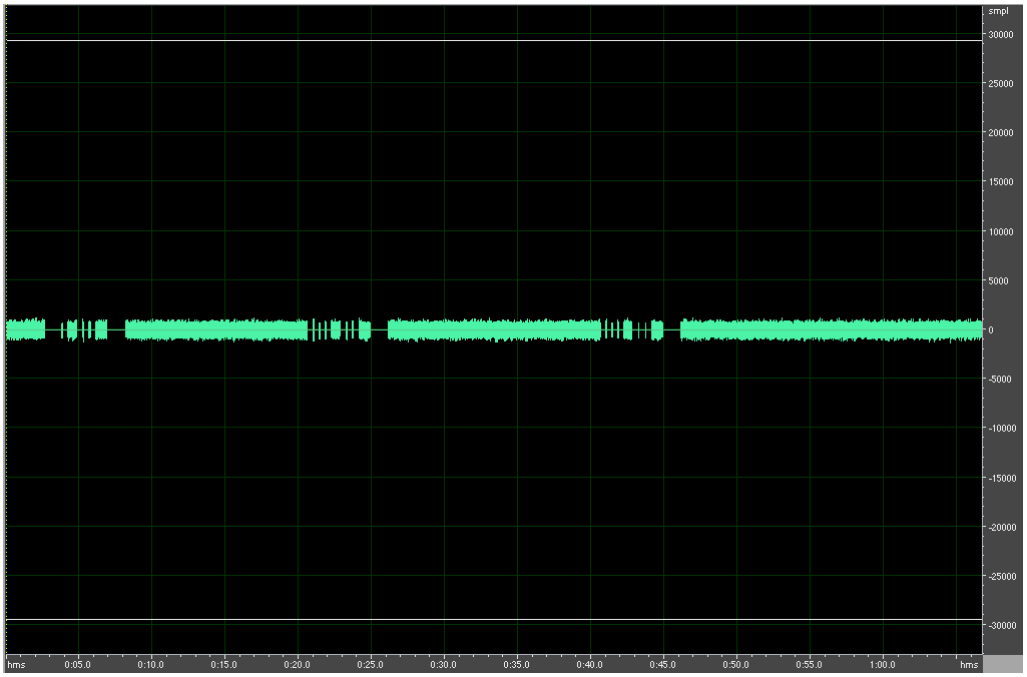


**圖 4-12 SNR=4.14dB 經過高斯混合模型所辨識的結果**

被辨識是屬於異音的部份(圖 4-13):


**圖 4-13 辨識是屬於異音的部份**

**測試一總結**:

當語音的特徵並不明顯時，仍可以使用背景環境和異音頻率不同的特徵區分出來。

**測試二: 語音位於麥克風陣列 90 度播放「交通大學工五館 905」與雜訊位於麥克風陣列 140 度方向(SNR=2.11dB)**

麥克風接收到的聲音但其特徵並不明顯(圖 4-14):

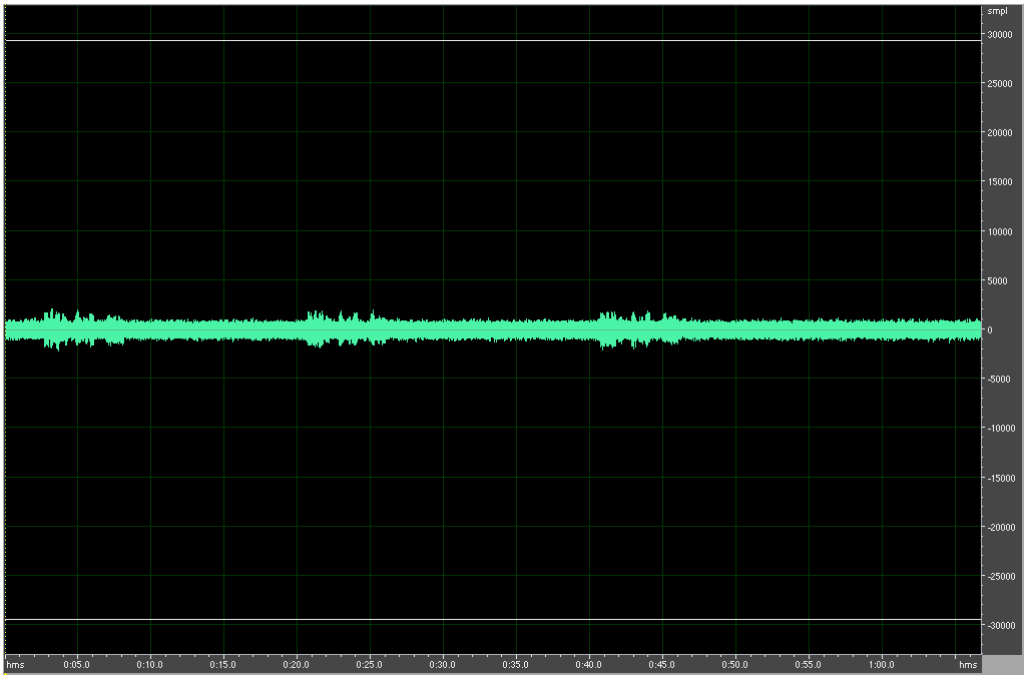


**圖 4-14 麥克風於實驗接收到的聲音**

SNR=2.11dB 經過高斯混合模型所辨識的結果(圖 4-15)：

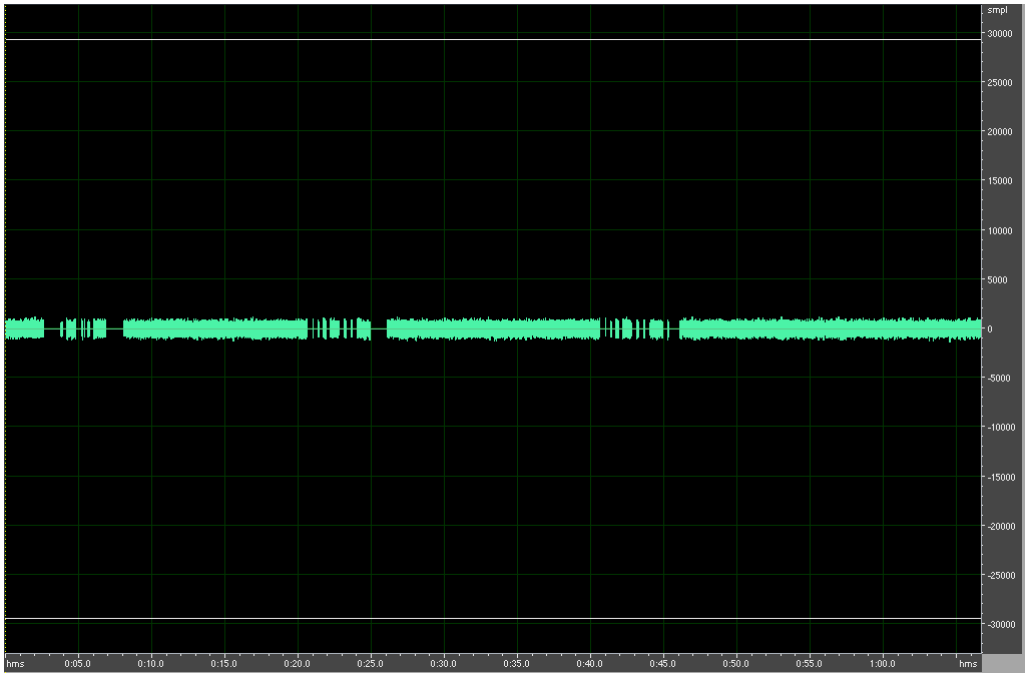


**圖 4-15 SNR=2.11dB 經過高斯混合模型所辨識的結果**

被辨識是屬於異音的部份(圖 4-16):


**圖 4-16 辨識是屬於異音的部份**

**測試二總結**:

當語音的特徵已經不容易只從聲音的波形上辨識出來，只能根據模型的統計資訊去區分發生的機率值，進一步辨識出來。當我們使用高維度的模型去描述背景環境的聲場分佈。高維度的模型可以有更詳細的描述能力，所以對異音的監控也有更好的辨識能力。

**測試三:語音位於麥克風陣列 45 度播放「交通大學工五館 905」與雜訊位於麥克風陣列 140 度方向(SNR=3.64dB)**

麥克風收到的聲音(圖 4-17):

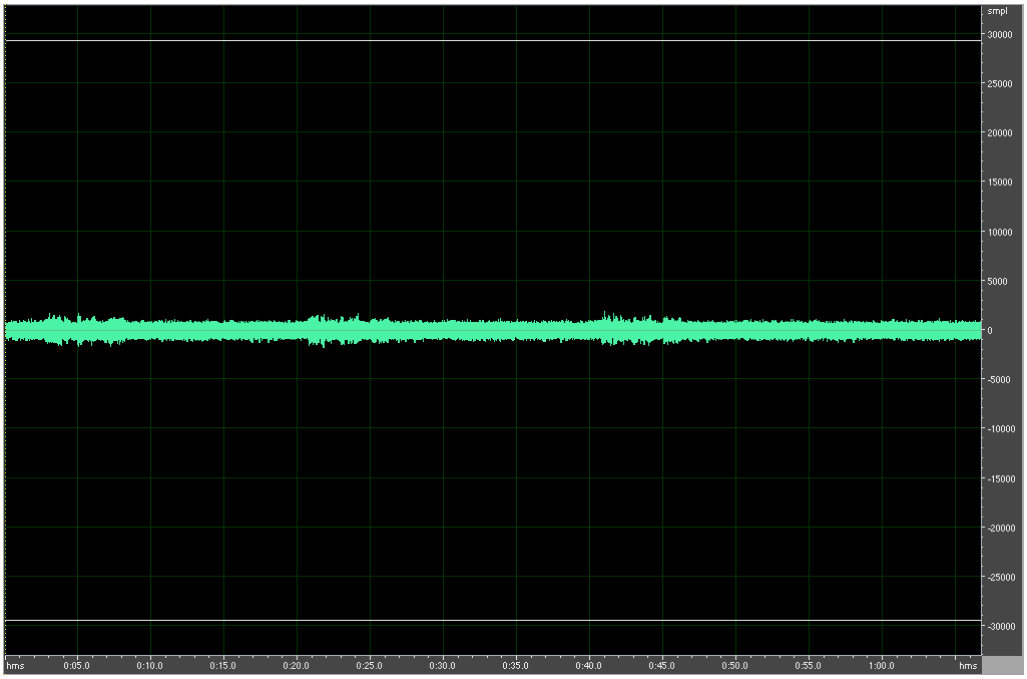


**圖 4-17 麥克風於實驗接收到的聲音**

SNR=3.64dB 經過高斯混合模型所辨識的結果(圖 4-18)：

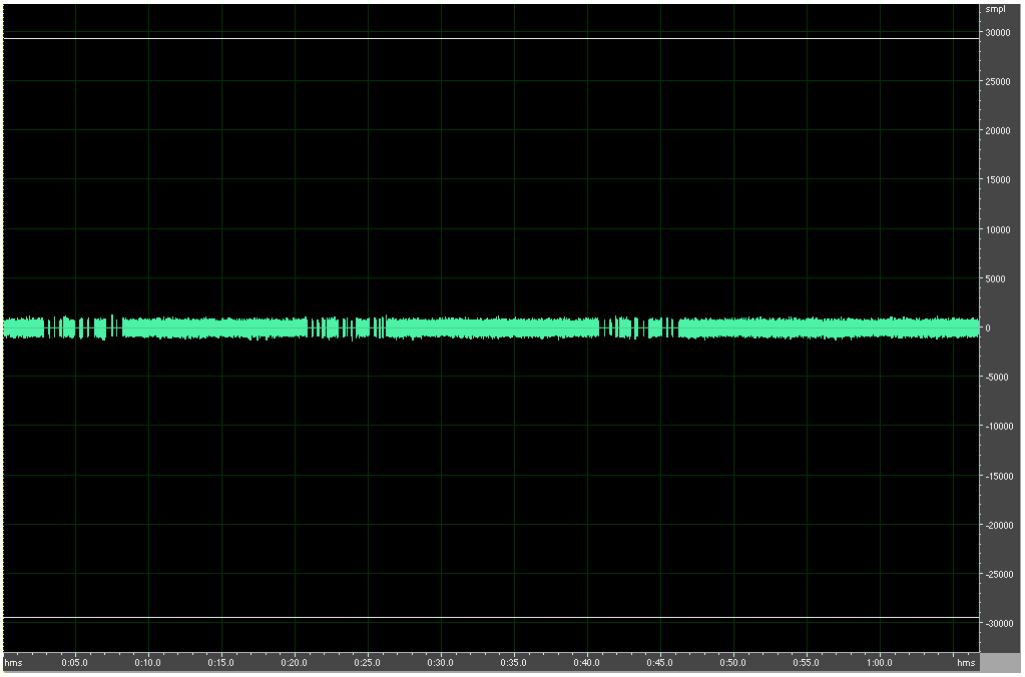


**圖 4-18 SNR=3.64dB 經過高斯混合模型所辨識的結果**

被辨識是屬於異音的部份(圖 4-19):

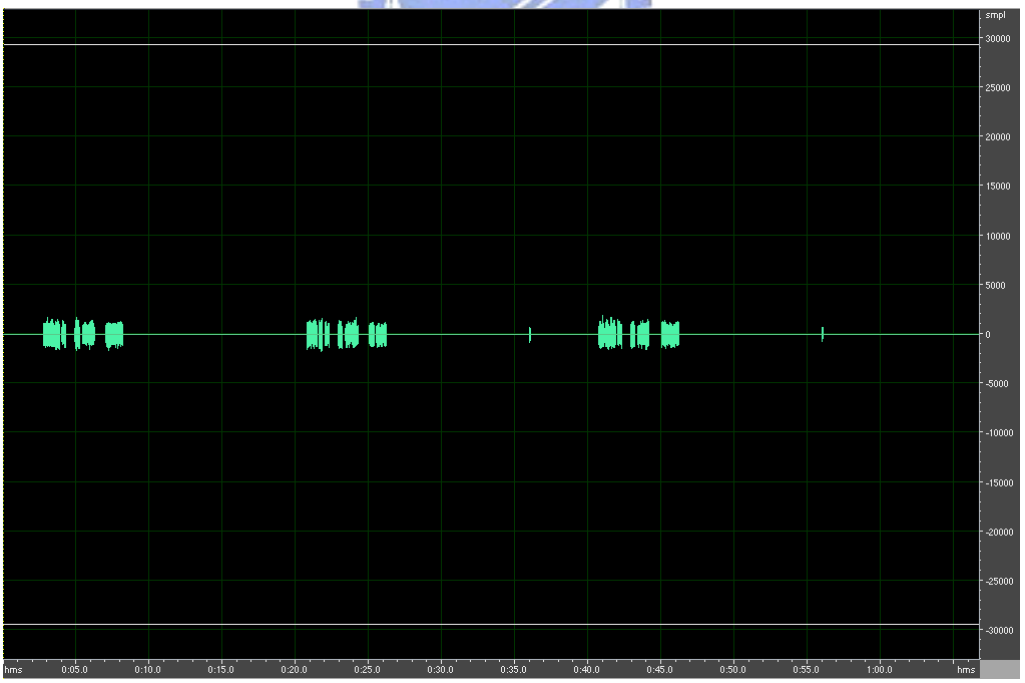


**圖 4-19 辨識是屬於異音的部份**

**測試三總結**:

移動聲源到不同角度。語音的特徵還是不明顯的情形下，還是可以使用高

斯混合模型區分出所要監控的異音出來。

### 4.2.3 單顆麥克風高斯混合模型(維度=1)對於異音監控結果

本節探討的是只用單一顆麥克風單一特徵建立高斯混合模型對於異音監控的結果。

**測試一：語音位於麥克風 90 度播放「交通大學工五館 905」與雜訊位於麥克風陣列 140 度方向**

單顆麥克風收到聲音(圖 4-20):

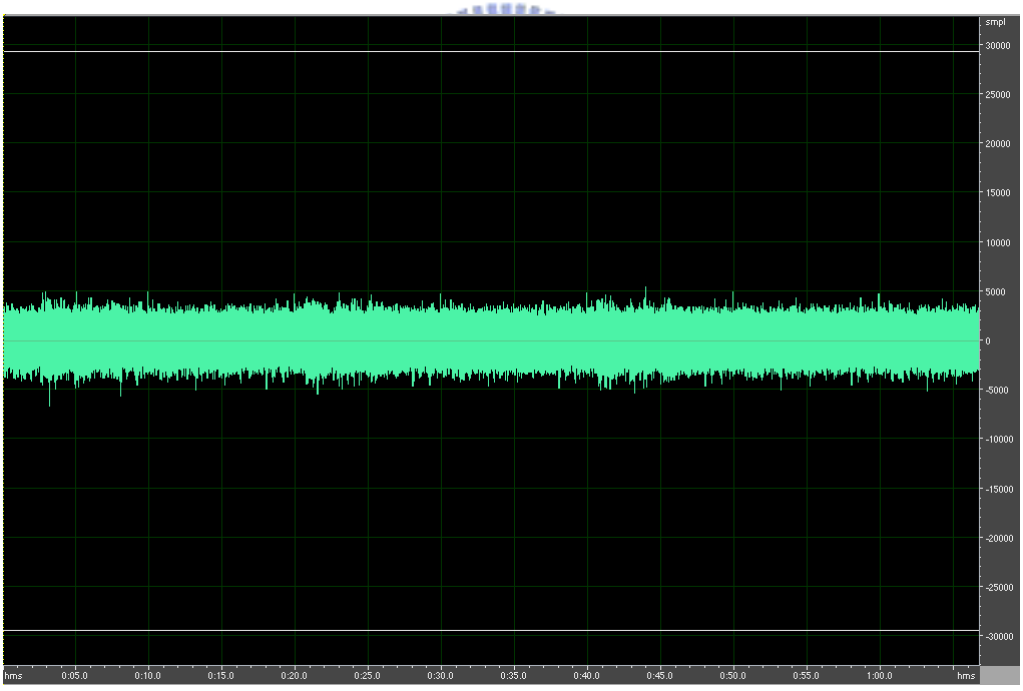


**圖 4-20 單顆麥克風於實驗收到聲音**

SNR=2.1dB 經過高斯混合模型所辨識的結果(圖 4-21)：

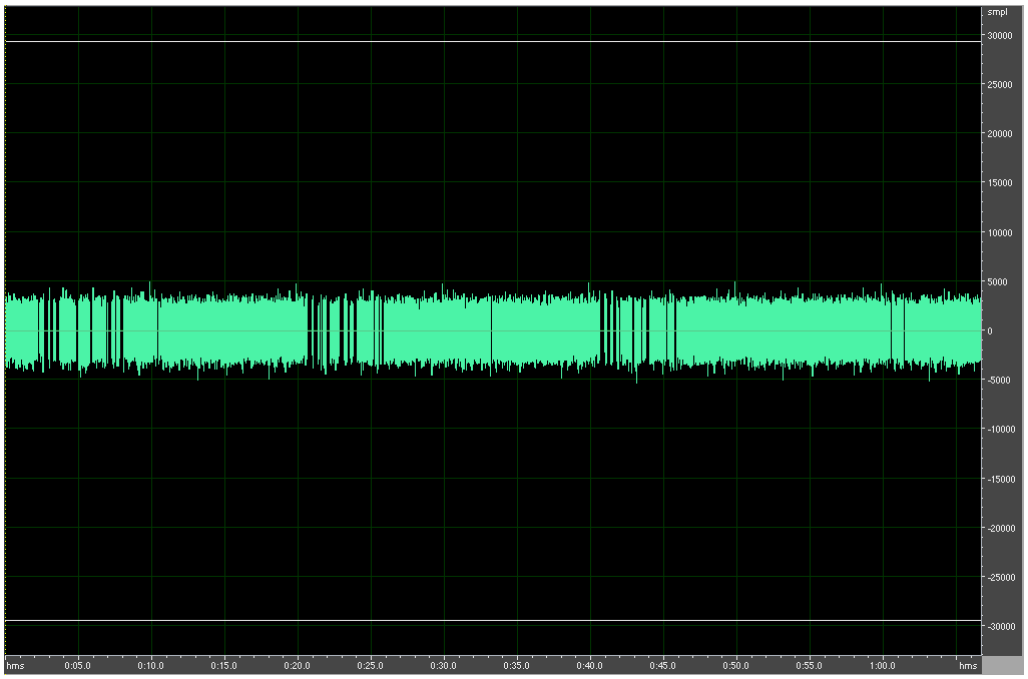


**圖 4-21 SNR=2.1dB 經過高斯混合模型所辨識的結果**

被辨識是屬於異音的部份(圖 4-22):


**圖 4-22 辨識是屬於異音的部份**

**測試一總結**:

利用單顆麥克風建立高斯混合模型，所利用到的聲音資訊只有單顆麥克風所收到的聲音。所以對於環境的聲場分佈的描述能力並沒有 8 顆麥克風所收到的好。所以對於單顆麥克風而言會有部份的異音沒有辨識出來或者是把環境的聲音當作了異音。

### 4.2.4 語音活動偵測對於異音監控的結果

本節探討的是在沒有建模型的情況下，只使用能量偵測和越零率測量來估測語音的發生。實驗環境一樣有兩個喇叭，一個喇叭用來播放語言，另一個播放雜訊。並在下列兩種情況下播放測試其辨識效果:

1. 在不同 SNR 的語音環境下
2. 語音在不同角度辨識的結果

**測試一: 語音位於麥克風 67.5 度播放「交通大學工五館 905」與雜訊位於麥克風陣列 140 度方向**

語音播放「交通大學工五館 905」重覆出現三次與雜訊之混合訊號(圖 4-23)雜訊能量為-42.14dB，而語音「交通大學工五館 905」與雜訊混合部份的能量為-31.1dB，因此 SNR=11.04dB

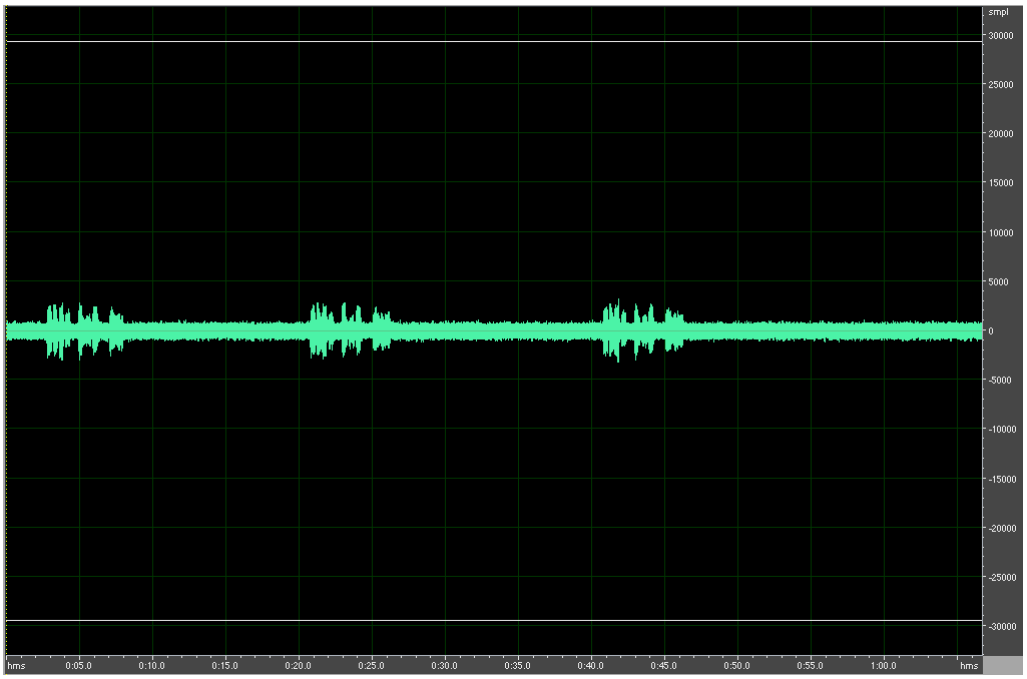


**圖 4-23 麥克風於實驗收到聲音**

經過語音活動偵測所辨識的結果(圖 4-24):

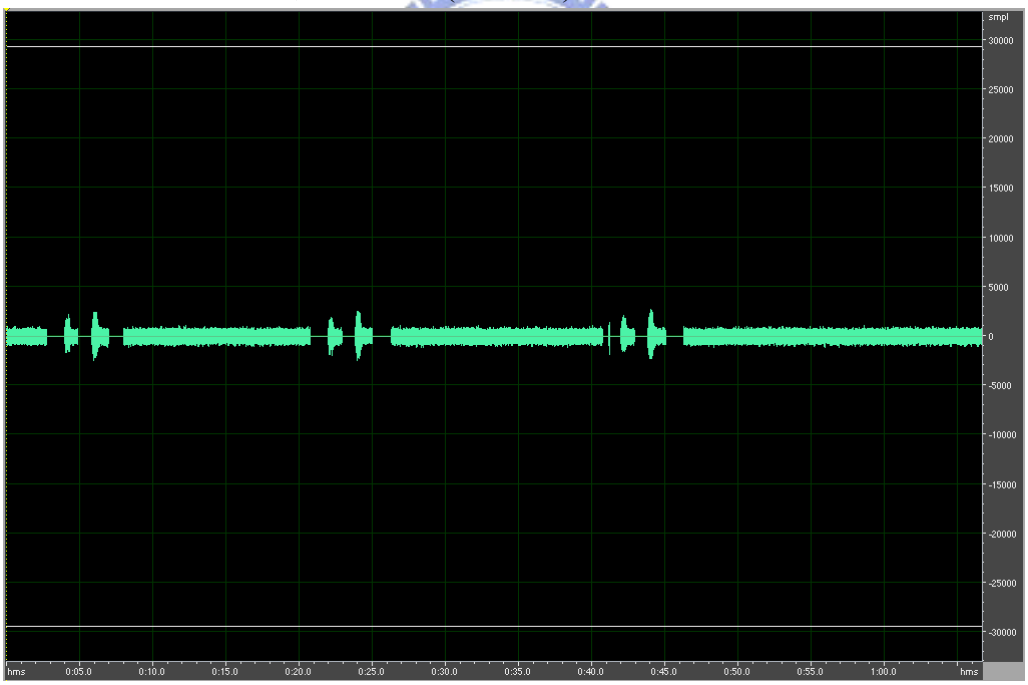


**圖 4-24 SNR=11.04dB 經過語音活動偵測所辨識的結果**

被辨識是屬於異音的部份(圖 4-25):

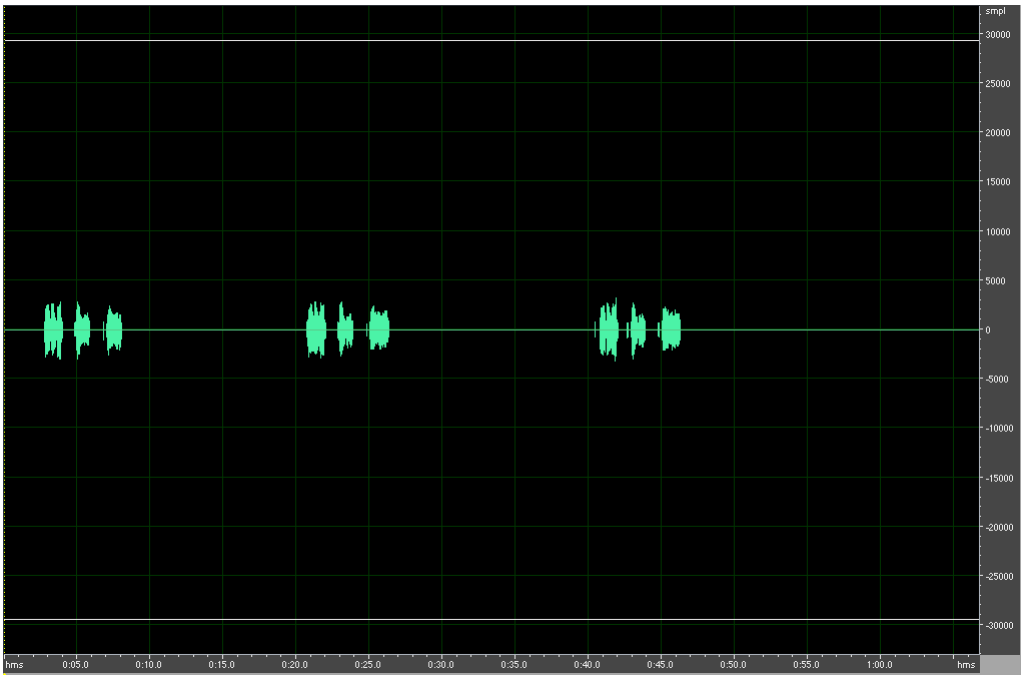


**圖 4-25 辨識是屬於異音的部份**

**測試一總結**:

當特徵明顯時，可以很容易的從視覺上判斷出發生語音的地方。單純設定臨界值就可以監控出異音出來。但是當特徵不明顯時，其效果就沒有那麼好了。

**測試二: 語音位於麥克風 90 度播放「交通大學工五館 905」與雜訊位於麥克風陣列 140 度方向**

語音播放「交通大學工五館 905」重覆出現三次與雜訊之混合訊號(圖 4-26)雜訊能量為-41.4dB，而語音「交通大學工五館 905」與雜訊混合部份的能量為-35.33dB，因此 SNR=6.07dB

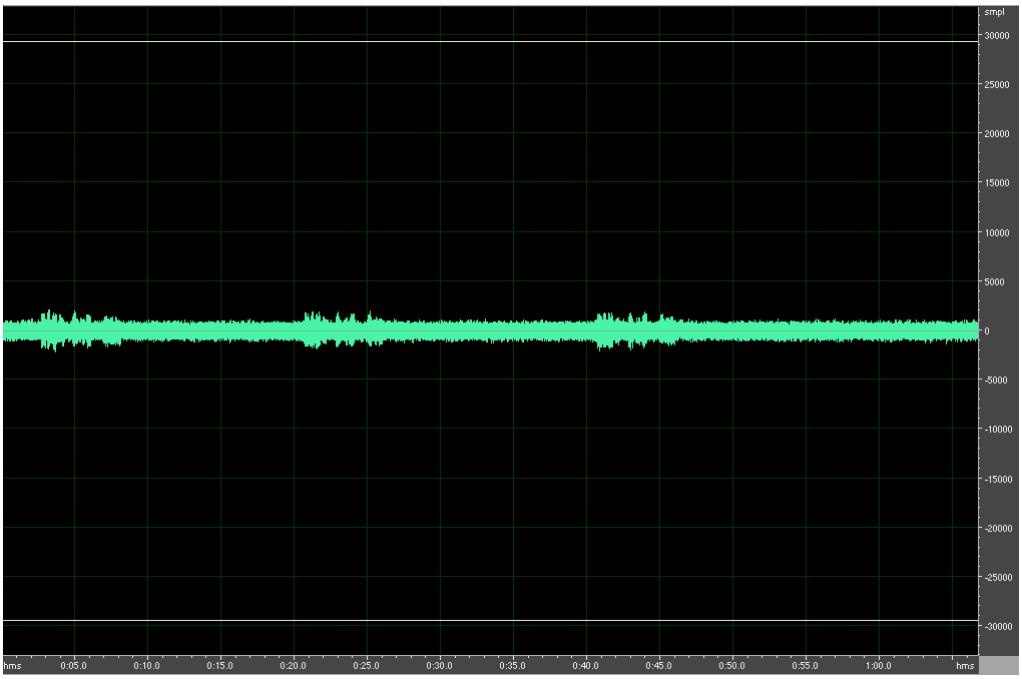


**圖 4-26 麥克風於實驗收到聲音**

經過語音活動偵測所辨識的結果(圖 4-27):

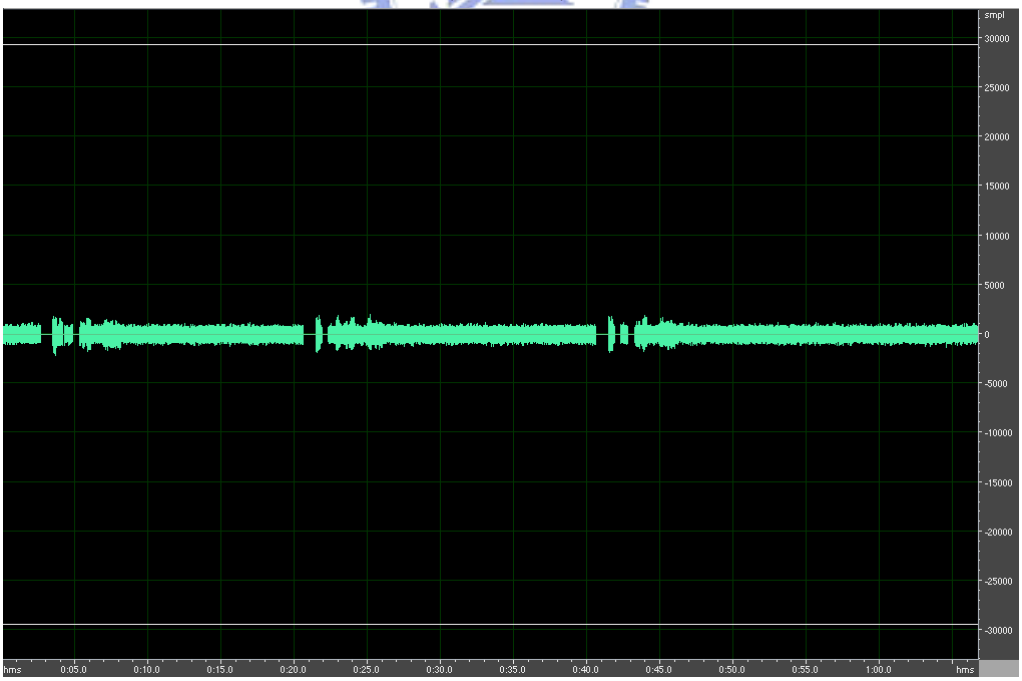


**圖 4-27 SNR=6.07dB 經過語音活動偵測所辨識的結果**

被辨識是屬於異音的部份(圖 4-28):

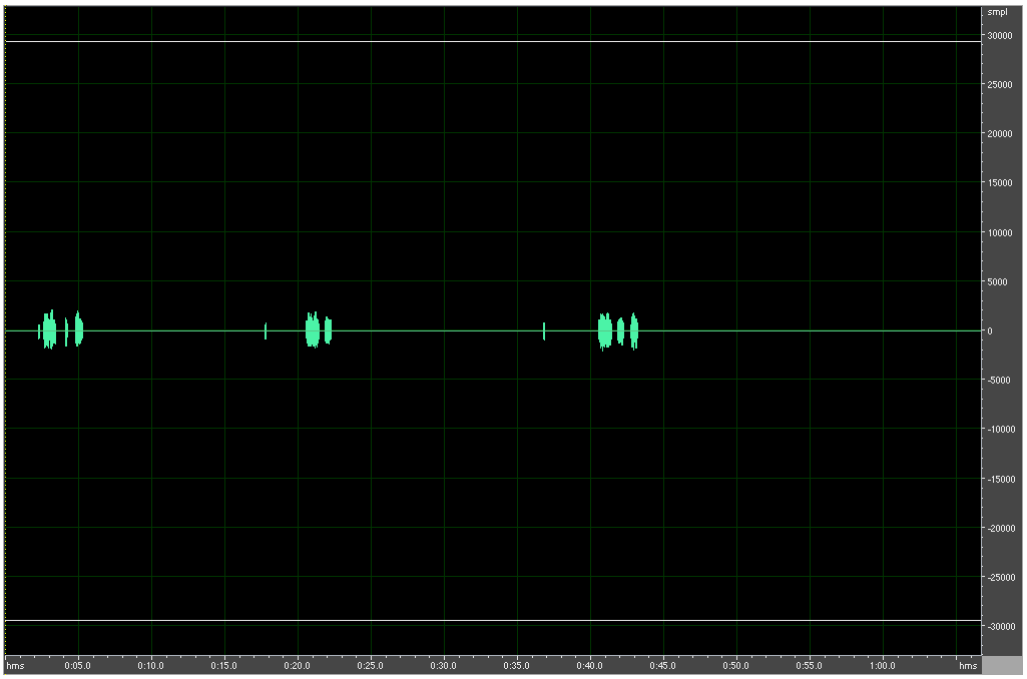


**圖 4-28 辨識是屬於異音的部份**

**測試二總結**:

只使用能量偵測和越零率測量來估測語音的發生。在特徵明顯時，其判斷的結果還可以和高斯混合模型有差不多的效果。但是，當特徵不明顯時其如果有發生異音時，會有一部分的異音無法偵測出來。

## 4.2.5 利用 Support Vector Machine 對於異音監控結果

本章節是探討在聲音監控常用到的二分法:支持向量機分類法(SVC)和高斯混合模型在不同角度和不同 SNR 值上所辨識的效果。實驗環境一樣有兩個喇叭，一個喇叭用來播放語言，另一個播放雜訊。並在下列兩種情況下播放測試其辨識效果:

1. 在不同 SNR 的語音環境下
2. 語音在不同角度辨識的結果

**測試一: 語音位於麥克風陣列 90 度播放「交通大學工五館 905」與雜訊雜訊位於麥克風陣列 140 度方向**

語音播放「交通大學工五館 905」重覆出現三次與雜訊之混合訊號(圖 4-29)雜訊能量為-41.39dB，而語音「交通大學工五館 905」與雜訊混合部份的能量為-30.02dB，因此 SNR=11.37dB

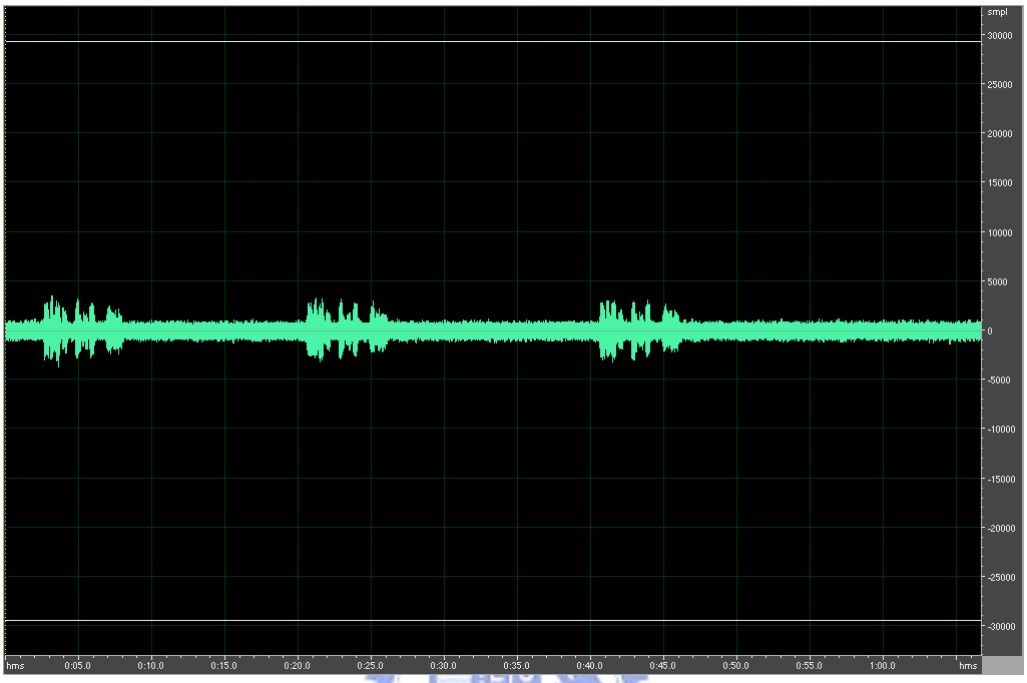


**圖 4-29 麥克風於實驗收到聲音**

SVM 所辨識的結果(圖 4-30):

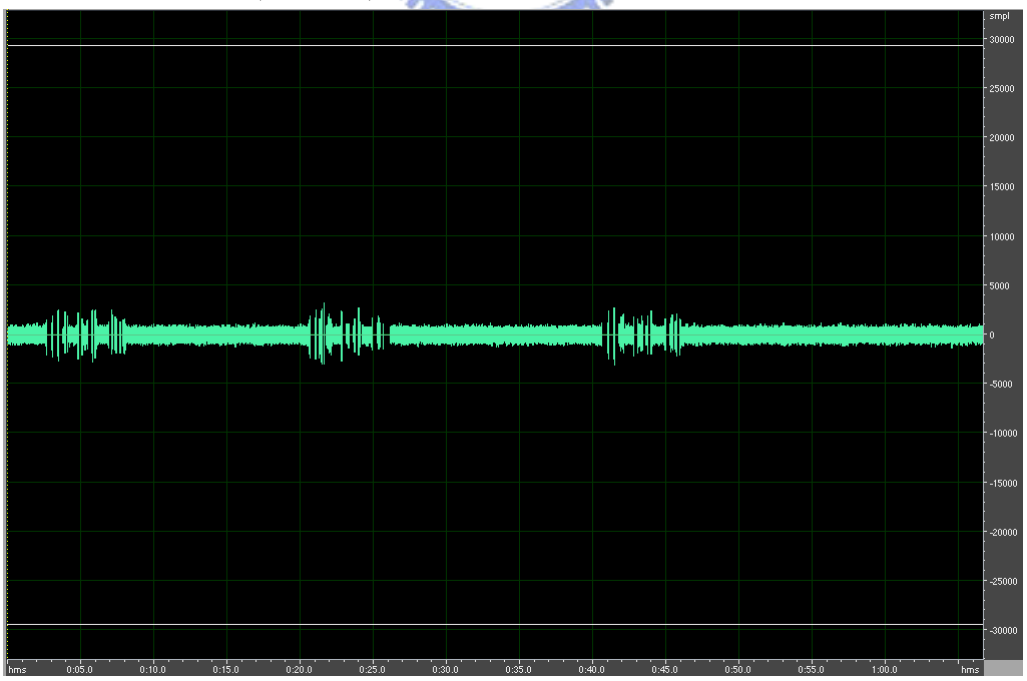


**圖 4-30 SNR=11.37dB 經過 SVM 所辨識的結果**

被辨識是屬於異音的部份(圖 4-31):

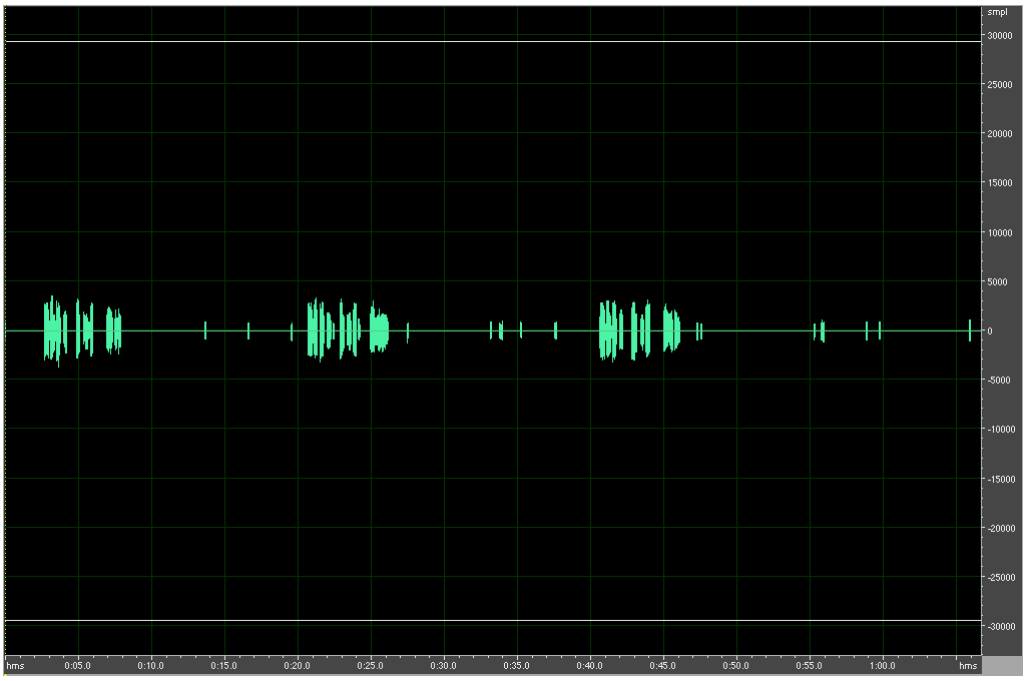


**圖 4-31 辨識是屬於異音的部份**

**測試一總結**:

當特徵明顯時，利用 SVM 可以雖然有偵測到異音的部份。但是也是有相當多的環境聲音被誤認為異音。

**測試二: 語音「交通大學工五館 905」在麥克風陣列 45 度與雜訊**

語音播放「交通大學工五館 905」重覆出現三次與雜訊之混合訊號(圖 4-32)雜訊能量為-41.47dB，而語音「交通大學工五館 905」與雜訊混合部份的能量為-34.67dB，因此 SNR=6.8dB。

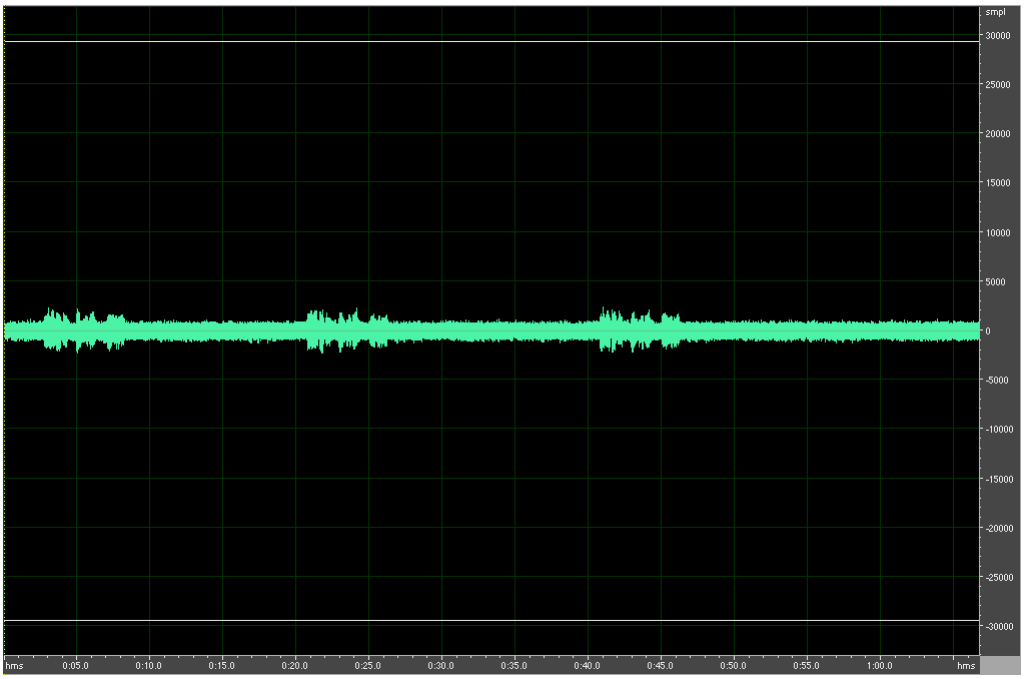


**圖 4-32 麥克風於實驗收到聲音**

SVM 所辨識的結果(圖 4-33):

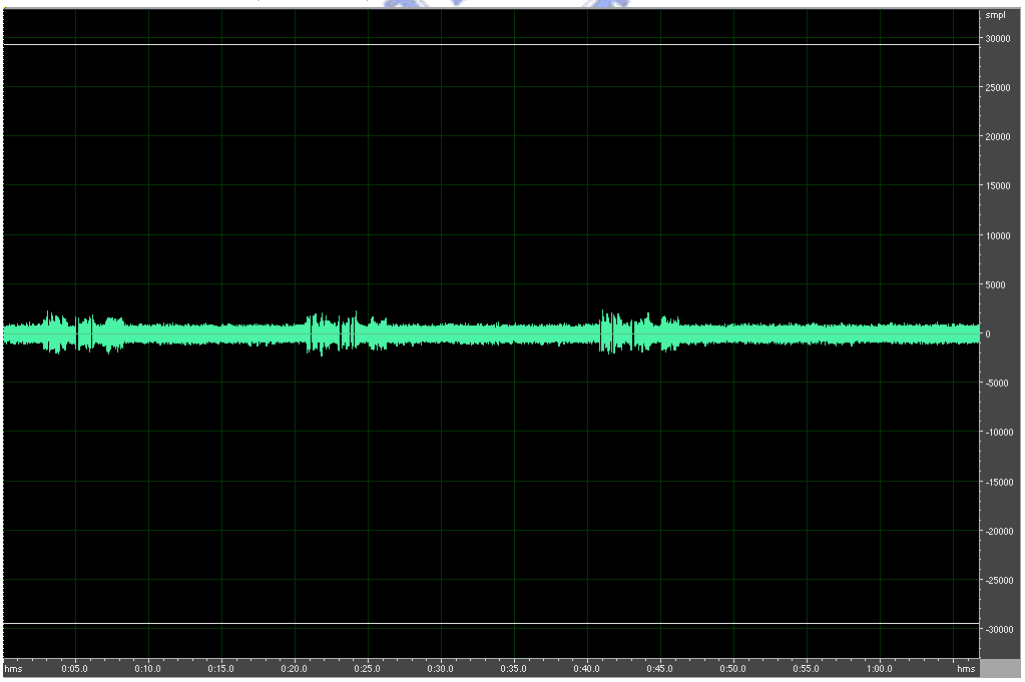


**圖 4-33 SNR=6.8dB 經過 SVM 所辨識的結果**

被辨識是屬於異音的部份(圖 4-34):

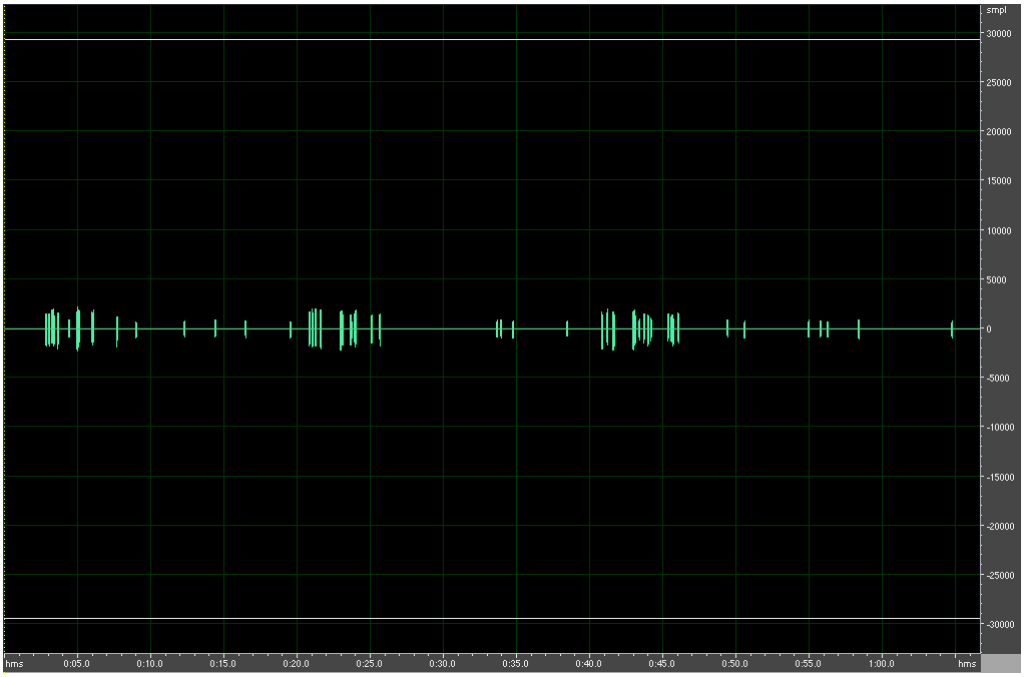


**圖 4-34 被辨識是屬於異音的部份**

**測試二總結**:

在語音監控上也常使用 SVM 二分法來區分屬於環境既有的聲音和非環境的聲音。但是當特徵並不是很明顯時，SVM 分類的效果很差。

## 4.2.6 方法比較

本章節探討使用高維度高斯混合模型在聲音監控上對於整個空間環境的描述能力比較。並且也探討高維度高斯混合模型和語音活動偵測、Support Vector Machine 分類器等方法之間的比較結果。

實驗方式是播放同組資料進行上述分析，觀察各演算法之間差異。

**測試一: 語音位於麥克風陣列 90 度播放「交通大學工五館 905」與雜訊雜訊位於麥克風陣列 140 度方向**

語音播放「交通大學工五館 905」重覆出現三次與雜訊之混合訊號(圖 4-35)

SNR=2.09dB:

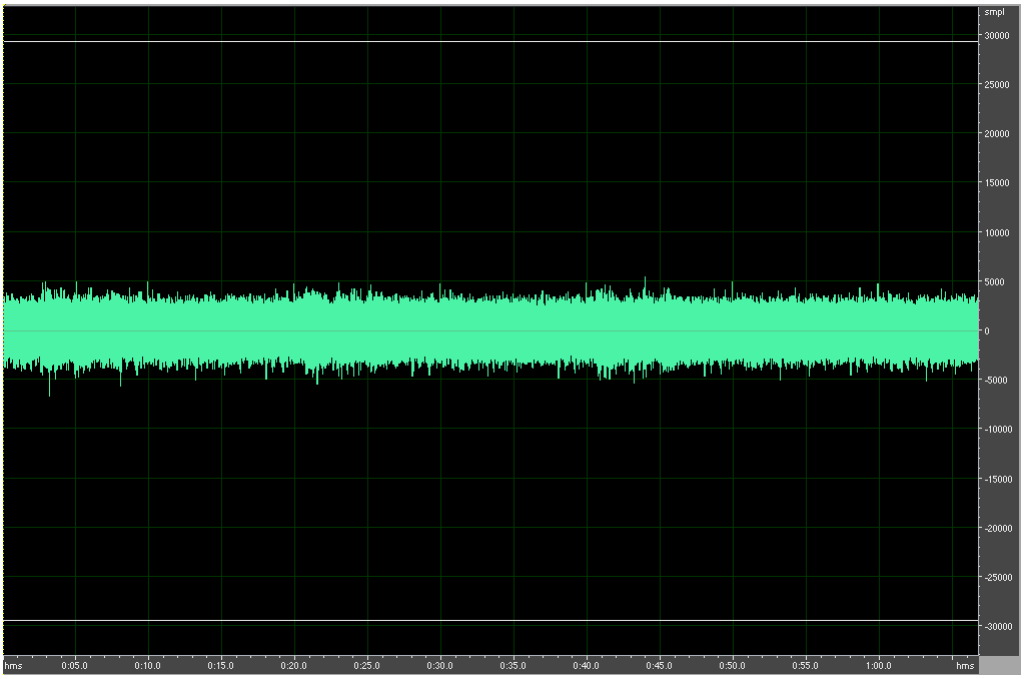


**圖 4-35 麥克風於實驗收到聲音**

辨識異音結果(圖 4-36):

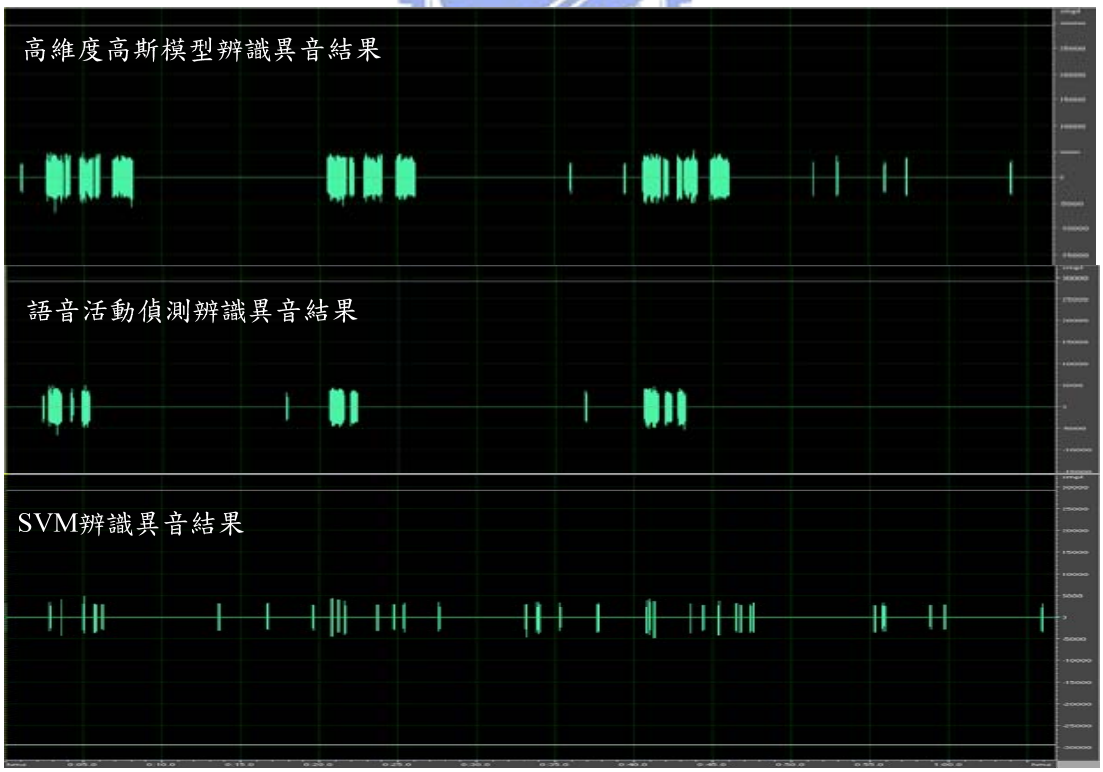


**圖 4-36 辨識異音結果比較**

判斷為背景聲音(圖 4-37):

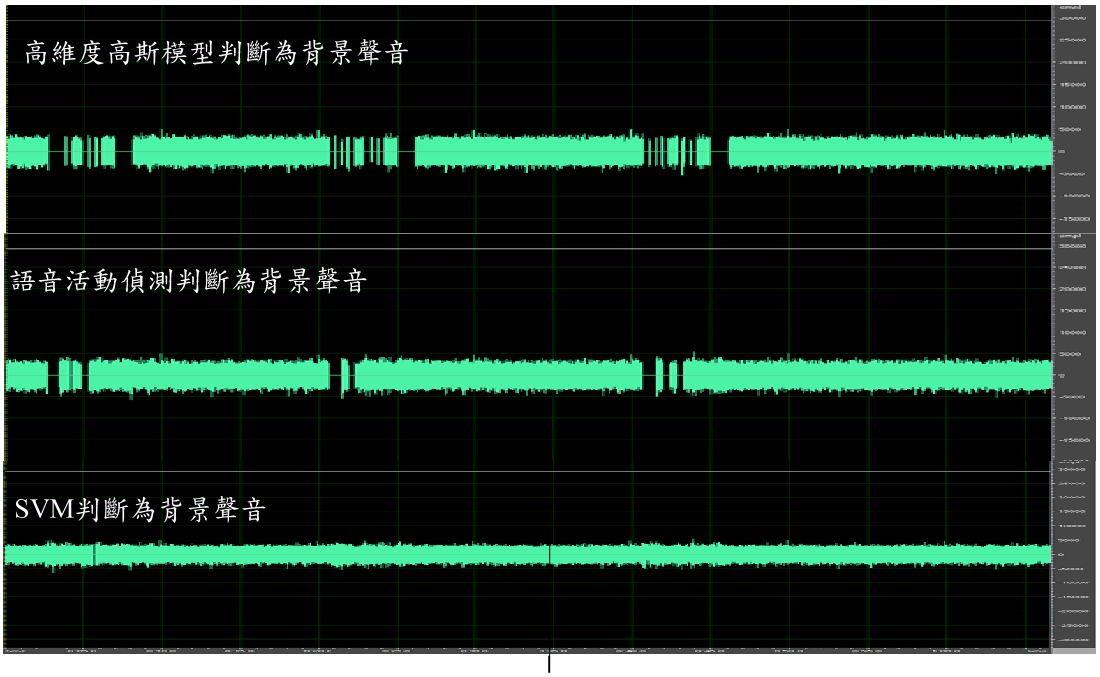


**圖 4-37 判斷為背景聲音**

**測試一總結**:由(圖 4-36) (圖 4-37)可以看出使用高斯混合模型在 SNR 值很低的情況下，語音活動偵測會有部分異音沒有偵測出來。而 SVM 幾乎已經沒有判斷的能力了。相較於語音活動偵測和 SVM 分類器依然能夠有高的辨識能力。

最後為了判斷模型對於語音監控辨識能力的高低，首先定義正確率和錯誤率以區分辨識的結果好壞[8]:

$$\text{正確率} = \frac{\text{判斷正確的個數}}{\text{正確資料的總個數}}$$

$$\text{錯誤率} = \frac{\text{將環境聲音判斷為非環境聲音}}{\text{環境聲音的總個數}}$$

現在實驗是將聲源放置麥克風陣列 6 種不同角度距離 120 公分播放語音。比較高維度高斯混合模型、語音活動偵測、SVM 等方法，在不同 SNR 可達到的正確率和錯誤率(正確率, 錯誤率)。

45 度(表 4-1)

| ( dB) | 12.53 | 8.44 | 5.6 | 3.64 |
|---|---|---|---|---|
| GMM | (98.44%,1.56%) | | | (72.66%,1.04%) |
| 語音活動偵測 | (90.63%,6.25%) | (79.7%,4.7%) | (10.7%,2.1%) | |
| SVM | (61.72%,0%) | (32.81%,0.52%) | | |

**表 4-1 聲音監控演算法對同一角度異音判斷比較一**

67.5 度(表 4-2)

| ( dB) | 15.28 | 11.04 | 8.06 | 5.92 |
|---|---|---|---|---|
| GMM | (100%,4.5%) | | | (79.17%,1.5%) |
| 語音活動偵測 | (90.8%,9.5%) | (79.1%,7%) | (40%,2.5%) | |
| SVM | (83.33%,1%) | (45%,1%) | | |

**表 4-2 聲音監控演算法對同一角度異音判斷比較二**

90 度(表 4-3)

| ( dB) | 15.66 | 11.37 | 6.13 |
|---|---|---|---|
| GMM | (100%,3.55%) | | (85.37%,2.54%) |
| 語音活動偵測 | (99.2%,11.68%) | (90.24%,9.14%) | (34.15%,2.03%) |
| SVM | (78.05%,2.54%) | (50.41%,1.02%) | |

**表 4-3 聲音監控演算法對同一角度異音判斷比較三**

112.5 度(表 4-4)

| ( dB) | 16.23 | 11.71 | 6.82 |
|---|---|---|---|
| GMM | (100%,5.03%) | | (86.78%,1.01%) |
| 語音活動偵測 | (91.7%,11.56%) | (78.5%,7.54%) | (29.75%,5.51%) |
| SVM | (80.99,1.01%) | (48.76%,0.5%) | |

**表 4-4 聲音監控演算法對同一角度異音判斷比較四**

135 度(表 4-5)

| ( dB) | 13.55 | 8.94 | 4.57 |
|---|---|---|---|
| GMM | (100%,2.56%) | | (80%,1.54%) |
| 語音活動偵測 | (91.2%,7.69%) | (79.2%,5.13%) | |
| SVM | (80.8%,0.51%) | (49.6%,0.51%) | |

**表 4-5 聲音監控演算法對同一角度異音判斷比較五**

157.5 度(表 4-6)

| ( dB) | 13.81 | 9.26 | 4.64 |
|---|---|---|---|
| GMM | (100%,3.09%) | | (80.16%,1.55%) |
| 語音活動偵測 | (88.89%,7.73%) | (68.25%,4.12%) | |
| SVM | (65.87%,2.58%) | (38.89%,1.55%) | |

**表 4-6 聲音監控演算法對同一角度異音判斷比較六**

# 第五章 結論與未來展望

## 5.1 結論

本論文結合 Delay and Sum Beamformer 空間濾波器與高斯混合模型偵測異音的出現，並測試在不同訊噪比和不同方位的異音所偵測的結果。在一般異音監控方面，使用語音活動的方式判斷異音的機制是對音框判斷能量和頻率的變化，所以在低訊噪比的情況下變化不大，其判斷結果並不理想。所以結合空間濾波器的好處在於，空間濾波是針對空間資訊對聲源作純化。但該方向仍包含雜訊，所以利用高斯混合模型判斷是否符合聲場資訊的統計資料。之後利用高斯混合模型的判斷機制，還可以根據機率值判斷異音的方位。

所以利用麥克風陣列除了可以比一單麥克風辨識出更吵雜的環境中仍然有異音的出現之後，還可以知道異音出現的方位。這也是單顆麥克風無法達成的。這也是聲音監控為什麼要結合空間濾波的原因和優勢。

## 5.2 未來展望

使用高斯混合模型對於異音判斷能夠有好的結果，但是高斯混合模型需要大量的運算，所以需花相當的時間才可建立完成，為了有效減少運算量，我們將可探討如何自動選取 GMM 個數、訓練取樣音框數…等。

目前異音監控一開始所監控的角度也是隨機設定，未來可以先設定好異音較可能出現的角度監控。可避免異音一開始出現在沒有監控的角度上。

# Reference:


[1] Pradeep K. Atrey, Namunu C. Maddage and Mohan S. Kankanhalli, ” Audio Based Event Detection For Multimedia Surveillance”, IEEE Digital Object Identifier, vol. 5, May 2006.

[2] Aki Harma, Martin F. McKinney, and Janto Skowronek, “Automatic surveillance of the acoustic activity in our living environment,” in IEEE International Conference on Multimedia and Expo, Amsterdam, July 2005.

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

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

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

[7] 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.

[8] C. Clavel, T. Ehrette, and G. Richard, “Event detection for an audio-based surveillance system”, in IEEE International Conference on Multimedia and Expo, Amsterdam, July 2005.

[9] Radhakrishnan, R.; Divakaran, A., “Systematic Acquisition of Audio Classes for Elevator Surveillance”, SPIE Image and Video Communications and Processing, Vol. 5685, pp. 64-71, March 2005.

[10] Härma, M. F. McKinney, J. Skowronek: “Automatic Surveillance of the Acoustic Activity in our Living Environment”, Proc. IEEE International Conference on Multimedia and Expo., Amsterdam, July 2005.

[11] P. K. Atrey et al.: “Audio Based Event Detection for Multimedia Surveillance”, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse (France), May 2006.

[12] C. Clavel et al.: “Events detection for an audio-based surveillance system”, Proc.IEEE International Conference on Multimedia and Expo., Amsterdam, July 2005.

[13] G. Valenzise et al.: “Scream and Gunshot Detection and Localization for Audio-Surveillance Systems”, Proc. IEEE International Conference on Advanced Video and Signal based Surveillanace, London, September 2007.

[14] W. Zajdel et al.: “CASSANDRA: Audio-video Sensor Fusion for Aggression Detection”, Proc. IEEE International Conference on Advanced Video and Signal based Surveillanace, London, September 2007.

[15] Simon Moncrieff, Svetha Venkatesh, Geoff West, “Unifying Background Models over Complex Audio using Entropy”, IEEE Digital Object

Identifier, vol. 5,2006.

[16] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., volume 2, pages 246–252, Fort Collins, CO USA, 1999.

[17] Marco Cristani, Manuele Bicego, and Vittorio Murino, "On-line adaptive background modelling for audio surveillance", IEEE Digital Object Identifier, vol. 2,2004.

[18] M. F. McKinney and J. Breebaart, "Features for audio and music classification," in Proc. Int. Symp. Music Inf. Retrieval (ISMIR'2003), (Baltimore, USA), October 2003.

[19] T. K. Moon, "The Expectation-Maximization algorithm, "IEEE Signal Processing Magazine, 1996.

[20] C. H. Knapp, and G. C. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoustic, Speech, Signal Processing, vol. 24, pp. 320-327, Aug 1976.

[21] 鄭士奇,"以高斯混合模型為基礎並使用陰影濾除之動態背景影像模型建立",國立交通大學,碩士論文,民國 94 年。

[22] 張永融,"利用聲場特徵及光流影像定位之全方向運動平台",國立交通大學,碩士論文,民國 95 年。

[23] 楊佳興,"使用麥克風陣列實現即時語音純化與真人語音活動偵測",國立交通大學,碩士論文,民國 94 年。

[24] 黃楷祥,"利用聲場特徵與 SVM 實現可結合輪式機器人之避障與導航",國立交通大學,碩士論文,民國 96 年。

[25] 王小川 編著,語音訊號處理,全華科技,2002

[26] 丁家群，“語音辨識與 Visual Basic”，義守大學，碩士論文，民國 92 年。

[27] 賴建瑞，簡仁宗，“結合麥克風陣列及模型調整技術之遠距離語音辨識系統”，國立成功大學，民國 89 年。