行政院國家科學委員會專題研究計畫　期中進度報告

子計畫五：結合麥克風陣列與影像之居家監護系統(2/3)

計畫類別： 整合型計畫
計畫編號： NSC94-2218-E-009-009-
執行期間： 94 年 10 月 01 日至 95 年 09 月 30 日
執行單位： 國立交通大學電機與控制工程學系(所)

計畫主持人： 胡竹生

報告類型： 精簡報告

處理方式： 本計畫可公開查詢

中　華　民　國 95 年 8 月 1 日

行政院國家科學委員會專題研究計畫成果報告

子計劃五: 結合麥克風陣列與影像之居家監護系統(2/3)

計畫編號： NSC94－2218－E009－009

執行期限：94 年 8 月 1 日至 95 年 7 月 31 日

主持人： 胡竹生 國立交通大學電機與控制工程學系

計畫參與人員：蘇宗敏、鄭价呈、劉維瀚、楊佳興、林群棋、黃恆嘉、林佩靜

國立交通大學電機與控制工程學系

一、中文摘要

本子計劃將今年度繼續針對居家監護系統進行影像與音訊監控平台之研發，影像監控平台包含兩部份，第一部份是解決 PTZ 攝影機與全向式攝影機之座標轉換關係，以便利用全向式攝影機所具備之大範圍監控優點，以彌補 PTZ 攝影機監控範圍受限的缺點；第二部份是進行三維物件辨識的研究，以期能夠在分離影像中之前景物體與背景影像後，進一步針對前景物體進行後續的處理。而在三維物件辨識之理論研究方面，本子計畫中建構出一套利用傅立葉描述子與邊緣點對點描述子的三維物件資料庫，該資料庫可利用本子計畫所發展之結合演算法來有效的降低每個物件所需要的二維投影影像數量，以及能夠讓每個物件的資料庫隨著所收集的影像增加而越加完善，此外，建立每個物件的資料庫時，用來建構該資料庫的訓練影像並不需要按照物件的拍攝角度排列，大幅的簡化建構三維物件資料庫的流程。而在音訊監控平台的部份，已針對屋內之吵雜實現出一適應性空間濾波器與真人語音活動偵測系統。此系統可與影像監控平台做結合，影像監控平台可將使用者所在的絕對位置資訊傳遞給音訊監控平台，音訊監控平台可對特定的位置達到濾除環境雜音粹取人聲並之需求。傳統的濾除雜訊方法須選取互相匹配的麥克風，而本語音純化系統將麥克風不匹配動態效應納入考量，以達到本系統不需考慮麥克風匹配問題，並將低成本。本計劃亦將著名的 H∞ 理論套用於語音純化中，H∞ 理論對於雜訊並不用做出任何的假設並且對於模型誤差較為穩健。實驗結果展示，本系統能對特定聲源位置抑制干擾源與粹取人聲，並提升語音 SNR。

關鍵詞：居家照護系統，PTZ 攝影機，全向式攝影機，三維物件辨識，真人語音活動偵測系統，適應性空間濾波器，H∞

Abstract

The goal of this project in this year keeps on developing a platform for image and audio surveillance. In the image surveillance platform, the relationship between the coordinate of the omni-directional camera and the PTZ camera is calculated first. Then the advantage of omni-directional camera about the capability of wide area surveillance can be used to compensate the disadvantage of the PTZ camera that has limited view of angle. Secondly, recognizing 3D objects is studied to further process the foreground objects. A 3D object database is established with Fourier descriptor and point-to-point length via the proposed combinational algorithm. The characteristic views of each object can be reduced efficiently. Moreover, the object representation becomes more and more accurate after gathering more new object views. Furthermore, the image database can be built using object views sampled at random intervals. In the audio surveillance system, a real-time interference suppression and voice activity detection (VAD) system has been implemented. The system can be combined with the image surveillance system which can provide the audio surveillance system with the user absolute position for suppressing environmental

interference and extract the human voice. Mutually matched microphones are needed for traditional multidimensional noise reduction methods, the proposed system adapts the mismatch dynamics to maintain the theoretical performance allowing unmatched microphones to be used in an array to make the cost down. An H∞ theory is also applied in the speech purification system. The H∞ filtering approach, which makes no assumptions about noise and disturbance, is robust to the modeling error in a channel recovery process. Consequently, the experimental results show that the proposed system can suppress interference and extract the human voice from the particular position, and enhance the SNR of speech signal.

Keywords: Home-Care System, Omni-directional Camera, Background Subtraction, VAD, microphone array, spatial filter

二、緣由與目的

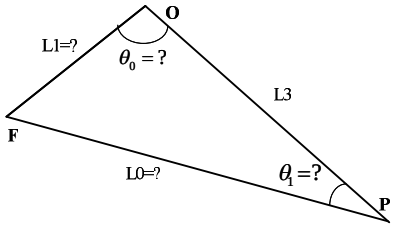
人類利用視覺可以快速的察覺周圍環境的變化狀況，因此如何引入影像的訊號來提高居家照護系統的效能就變成相當重要課題，在本子計劃中，利用全像式攝影機大範圍監控的特性，來針對環境影像做前景背景分離，並估測前景物體所在環境之位置，搭配 PTZ 攝影機以取得高解析度之前景影像；此外，針對三維物體的辨識機制進行資料庫建立、特徵抽取以及比對的主題研究。環境中的語音訊號干擾源總是存在，例如冷氣機、電腦風扇、喇叭、密閉空間反射等等。當語音訊號遭到干擾時，若用於語音辨識中，辨識率會大為降低，若用於通訊中，通話品質也大受影響。因此若能設計出一語音純化系統，降低環境中干擾源的影響，達到語音純化的效果，則在生活中將會有很大的應用面。

三、研究方法與結果

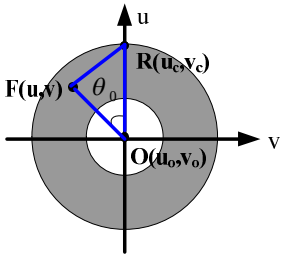
本計畫 94 年度期中報告中完成三項結果。1. 居家環境影像監控平台之建立；2. 居家環境之三維物體辨識機制；3. 結合語音活動偵測與語音純化系統之設計，其內容簡述如下：

1. 居家環境影像監控平台之建立：

架設在室內屋頂的全向式攝影機雖然可以 360 度全方向的監控室內活動，但全向式攝影機的解析度不高，且由於取像角度的關係，無法取得前景物體的詳細資訊以作為進一步的判斷依據，因此，我們在屋內的角落架設了一台 PTZ(Pan-Tilt-Zoom)攝影機，並利用三角定位法推導 PTZ 攝影機與全向式攝影機之座標轉換關係，以建構出一套結合 PTZ 攝影機以及全向式攝影機的居家環境影像監控平台，並且可以在未來因應需求而加入更多 PTZ 攝影機。



圖一. 屋內前景物體之三角定位示意圖



圖二. 全向式影像之成像

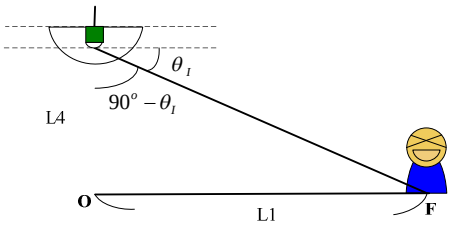
在圖一中，O 代表架設在室內屋頂的全向式攝影機中心點投影於地面的投影點，P 代表架設於室內角落的 PTZ 攝影機之中心點投影於地面的投影點，F 則代表經由前景物體之中心點；其中 L_3 代表是 O 跟 P 兩點之間的距離，可以藉由量測全向式攝影機中心點投影於地面的投影點 O 與 PTZ 攝影機之中心點投影於地面的投影點 P 之間的距離而得知。而 θ_0 則可以由全向式攝影機所擷取的影像(圖二)來估測，其中 R 代表距離中心點 O 最遠之距離，且其實際之距離 $\overrightarrow{RO}$ 亦可藉由事先量測而得知。接著可由式(1)來推導出 θ_0。

$$\theta_0 = \mathrm{a}\cos(\overrightarrow{FO} \cdot \overrightarrow{RO} / \| \overrightarrow{FO} \| \overrightarrow{RO} \|) \qquad (1\text{-}1)$$

其中，

$$\overrightarrow{FO} = (u - u_c, v - v_c)$$

$$\overrightarrow{RO} = (u_c - u_c, v_c - v_c)$$



圖三. 全向式攝影機與前景物體之座標關係示意圖

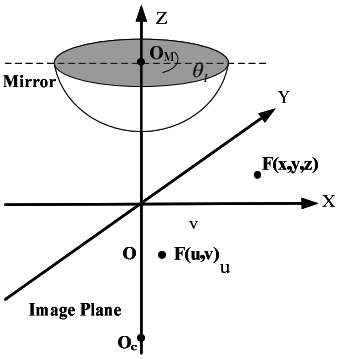
圖一中的L_1代表全向式攝影機中心點投影於地面的投影點O與前景物體中心點F之間的距離，可以藉由圖三的示意圖來估測，其中L_4代表全向式攝影機之成像平面與前景物體所在地面之高度，可以藉由事先量測而得知。而θ_1可以藉由全向式攝影機之成像原理來估測；圖四代表全向式攝影機之成像原理，由於全向式攝影機藉由 CCD 鏡頭擷取凸面鏡中的影像來獲取全方位的影像，因此當前景物體的入射角度θ_1與前景物體成像位置有一定關係時，則可以藉由所擷取之全向式影像來反推θ_1，從圖四中我們可以看出線性關係的趨勢。接著，可以利用式(2)來估測L_1

$$L_1 = \tan(90^o - \theta_I) * L_4 \qquad (1\text{-}2)$$

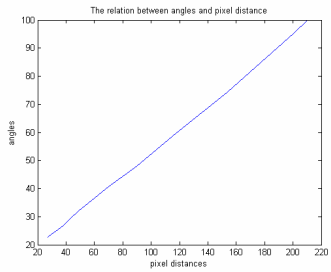
在求得圖一中的L_1、L_3與θ_0之後，便可以利用式(3)來推導得出θ_1，而 PTZ 攝影機即可利用所估測之θ_1來轉向前景物體。

$$\theta_1 = \text{asin}((L_1 / L_0)\sin\theta_0) \qquad (1\text{-}3)$$

$$L_0 = \sqrt{L_1^{\,2} + L_3^{\,2} - 2 * L_3 * L_1 * \cos\theta_0}$$



圖三. 全向式攝影機之前景物體成像示意圖



圖四.前景物體位於全向式影像之位置與入射角度θ_1之關係圖；橫軸代表前景物體位於全向式影像中的位置距離中心點之距離，而縱軸則代表θ_1

2. 居家環境之三維物體辨識機制：

在三維物體辨識機制部分，本子計畫由擷取到辨識出此物體做了一個完整的實現，首先，利用前景偵測結合肯尼邊緣偵測法（Canny edge detection）和加速的**梯度向量流動態輪廓偵測法**（GVF snake），來得到物體輪廓，接著利用此輪廓找尋出物體的特徵，然後配合計算相似度的方法，代入新提出的外觀結合演算法去建資料庫，並利用待測影像與資料庫影像之間的相似度來作為比對的依據，以找出與待測影像最相符之資料庫影像。

2.1 擷取 2D 影像之目標物體輪廓

對於一張實際拍攝的影像，即使在背景單純的情況下，要從中擷取出物體的精確輪廓，還是需要先濾除影像中所包含陰影與強光，在這部份，我們採用去年度本子計劃所發展之背景濾除演算法[1]，接著依序採用肯尼邊緣偵測法(Canny edge detection)[2]與梯度向量流動態輪廓模型(Gradient Vector Flow Snake)[3]來獲得目標物體之輪廓。

整個擷取影像中物體輪廓的步驟如下：

1. 利用背景濾除演算法濾除影像中所包含之背景、陰影與強光，留下前景(目標物體)。
2. 利用肯尼邊緣偵測法搭配前景物體所在區域，估測出目標物體之邊緣影像。
3. 利用梯度向量流動態輪廓模型估測出所需要的物體輪廓。

2.2 利用 2D 影像的輪廓辨識 3D 物體

在辨識裡，特徵擷取與辨識方法是很主要的兩個議題，在此我們採用傅立葉描述子與輪廓取樣點的向量長度來作為主要特徵與輔助特徵。

以下列出實際擷取二種特徵的步驟：

1.將原始物體封閉按順時針順序的輪廓重新

均勻地取樣成 N 點。
2.依據 N 點的重心的位置，重新定義座標。
3.利用此重新均勻地取樣的 N 點算出的輪廓長度 L 與預設的標準輪廓長度 L_c，將輪廓座標做比例放大縮小。
4.利用步驟 3 所得的輪廓中之 N 點座標，做複數傅立葉轉換，並擷取其低頻的部分的強度，即前後各 $N/10$ 點的頻譜強度做為主要特徵。
5.利用步驟 3 所得的輪廓中之 N 點座標，計算點與點之間的向量長度，利用此算出的 N 個長度做為輔助特徵。

若定義 n 為主要特徵或輔助特徵的長度，且定義兩個物體輪廓擷取出的特徵分別如下：
$U=\{u_0,\cdots,u_i,\cdots,u_{n-1}\}$
$V=\{v_0,\cdots,v_i,\cdots,v_{n-1}\}$
則計算相似度的方法定義如下：

$$D_{1-norm}(u,v)=\sum_{i=0}^{n-1}|u_i-v_i| \qquad (2\text{-}1)$$

2.3 資料庫的建立

對於現實情況而言，物體是可能複雜且多特性的，所以這裡我們利用基於相似度的外觀圖解法，來建立物體的資料庫。
首先，先定義一些用於下列說明的符號，V_{new} 指的是新擷取用來建立某物體資料庫的二維影像，C_m 是此物體資料庫中第 m 個外觀的特徵面，$C_{m^{\min}\pm1}$ 則是此新擷取之影像與所有此物體的特徵面中距離最近的那個特徵面的相鄰左右兩個特徵面，$m^{\min}$ 代表此新擷取之影像與所有此物體的特徵面中距離最近的那個特徵面所代表的外觀。
接著列出此新的外觀結合演算法運作的步驟如下：
1.當新的影像要用來建立某物體資料庫時，先判斷此物體資料庫中的外觀數目。
2. 依據外觀的數目，來做其建立此物體的資料庫的判斷依據。
(a)外觀的數目為 0：
此新進來的面直接形成一個外觀，且此外觀的特徵面就是此新進來的面。
(b)外觀的數目為 1 或 2：
若下式成立，則不增加新的外觀，並且將此新進來的面直接併入此擁有最小距離的外觀，並保有此外觀的原有的特徵面。

$$\min_{all\ C_m} d(V_{new},C_m)<threshold\_1 \qquad (2\text{-}2)$$

若不符合式(2-2)，則此新進來的面就會形成一個新外觀，且此新外觀的特徵面就是此新進來的面。
(c)外觀的數目大於等於 3：

$$\min_{all\ C_m} d(V_{new},C_m)>threshold\_2 \qquad (2\text{-}3)$$

$$\begin{cases}\min\limits_{all\ C_m} d(V_{new},C_m)<threshold\_2\\ d(V_{new},C_{m^{\min}\pm1})>threshold\_2\end{cases} \qquad (2\text{-}4)$$

上式中，$threshold\_2$ 大於 $threshold\_1$，若式(2-3)或式(2-4)其中一式成立且式(2-2)不成立，則此新進來的面就會形成一個新外觀，且此新外觀的特徵面就是此新進來的面。至於此新外觀的位置，則依據式(2-5)來判定，若成立，則此新外觀加在 $m^{\min}$ 和 $m^{\min-1}$ 兩個外觀之間，反之，加在 $m^{\min}$ 和 $m^{\min+1}$ 兩個外觀之間。

$$d(V_{new},C_{m^{\min}+1})>d(V_{new},C_{m^{\min}-1}) \qquad (2\text{-}5)$$

若此時式(2-3)或式(2-4)皆不成立或是式(2-2)成立，則不增加新的外觀，且將此新進來的面直接併入此擁有最小距離的外觀，且此保持該外觀的原來的特徵面。
總而言之，此法是以一個物體的一個面為單位的演算法，可以藉由不斷地新增物體影像來精確完整地表達物體，也可對不同的物體設定適合它們的門檻值，以建立出適合物體的資料庫。

2.4 辨識方法

採用主要特徵與輔助特徵來辨識物體的方法如下：(假設共有 N 個物體)。
1. 對一個不知名的面，取其主要特徵與輔助特徵，並利用此主要特徵結合計算相似度的方法去計算其與資料庫(以主要特徵建出的)中的每一個特徵面的距離，然後依照距離的大小取其前 $(N/2)+1$ 個小的特徵面所屬的物體，用輔助特徵做第二次判斷(步驟 2)。
2. 將由步驟 1 得到的可能物體，用輔助特徵再判斷一次，且在比較距離大小時，以輔助特徵算出的距離加上此物體的最小主要特徵算出的距離，且依照主要與輔助特徵建資料庫時的門檻值比例，將兩者的比重調至一樣，來做為比較距離大小的依據，如式(2-6)

$$d(V_j^i,V_m^n)=d_{\text{assistant}}(V_j^i,V_m^n)+\frac{threshold_{\text{assistant}}}{threshold_{main}}\times d_{main}^{\min}(n) \qquad (2\text{-}6)$$

其中 V_j^i 是未知的面，V_m^n 是第 n 個物體的特徵面 m，$d_{main}^{\min}(n)$ 則是用主要特徵得到的與可能物體最小的距離，$threshold_{assistant}$ 和 $threshold_{main}$ 分別是指利用輔助特徵與主要特徵建立資料庫時所設定的門檻值 $threshold\_2$。同樣地，最小距離的特徵面所屬的物體即是此不知名的面的辨識結果。

2.5 實驗結果

圖五中列出 12 種用來建立資料庫的三維物體，每個物體在 viewing sphere 的赤道線上每隔 5 度，按照由小到大排列的視角順序取樣 72 個面做為建立資料庫的面，然後在每隔 5 度裡，等份地取樣 3 張，做為辨識的面，所以一個物體共有 216 個面來當做測試辨識率的測試影像。在之後的說明裡，每個物體 72 個建立資料庫的面，我們統稱其為資料庫影像(database views)，216 個測試影像統稱為未知影像(unknown views)。

表1. 採用主要特徵與輔助特徵所建立的資料庫之外觀數目(統計200次不同的輸入影像次序)

外觀數目(個)		資料庫(圖五)中的物體編號											
		1	2	3	4	5	6	7	8	9	10	11	12
主要特徵	平均	34.66	3.84	27.83	24.75	6.87	9.47	2.04	25.62	17.14	16.16	16.62	28.75
	標準差	1.82	0.56	1.62	1.07	0.82	0.96	0.49	1.28	1.79	1.24	1.00	1.04
輔助特徵	平均	38.72	14.08	14.32	22.84	10.98	20.12	8.41	31.07	25.79	17.68	23.61	19.88
	標準差	2.51	1.67	1.81	1.79	1.17	1.94	1.09	2.07	2.18	1.68	1.81	1.59

表2. 利用主要特徵與輔助特徵作為相似度判斷之依據所統計得到的辨識率結果(各個物體有216張測試影像)

Top1 辨識率	資料庫(圖五)中的物體編號												
(%)	1	2	3	4	5	6	7	8	9	10	11	12	平均
平均	98.25	99.97	97.71	97.39	100	99.81	99.79	99.35	99.90	97.97	98.44	96.83	98.78
標準差	0.957	0.184	1.349	0.979	0.000	0.003	0.241	0.641	0.229	0.915	0.827	0.700	0.197

圖五. 資料庫中的 12 種三維物體，由左至右，由上至下分別為物體 1,物體 2,…,物體 12。

表 1 中列出資料庫中各個物體的外觀數目，分別是採用主要特徵與輔助特徵建立而成，由於不需要按照當初拍攝物體的角度順序地建立資料庫時，所以每次的結果都會有一些差別，但由表 1 中可知，在統計 200 次後，所算出的外觀數目之標準差都不大，所以由此可以推斷這樣方法不會因輸入影像的順序不同而有很大的差異，是一種可以任意地加入物體的影像去建資料庫的方法。表 2 則為利用主要特徵與輔助特徵作為相似度判斷之依據所統計得到的辨識率結果，從表 2 中可以得知各個物體的 Top1 辨識率(相似程度最高的辨識結果)都有相當不錯的結果，且其標準差也都相當小，可以驗證所提出方法在三維物體的辨識上具有相當穩健的辨識率。

3. 結合語音活動偵測與語音純化系統之設計：

本子計畫提出了結合語音活動偵測（Voice Activity Detection ,VAD）與適應性陣列訊號處理架構，其架構圖如圖六所示，在圖六架構中，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，若判定為語音訊號，則語音訊號會直接輸出，若為非語音訊號，系統會將非語音訊號的原始訊號（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。

3.1 語音活動偵測 (Voice Activity

Detection ,VAD)

語音活動偵測是用來判定是否有語音訊號，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。本子計畫使用的 VAD 演算法[4]是使用長時間語音的資訊而非傳統瞬間音框訊，針對長時間語音資訊，定義出下列定義。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（Long-Term Spectrum Envelope）定義為：

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

其 $\mathrm{LTSE}_{\mathrm{N}}(k,l)$ 代表的意義為，從第 l-N 個音框到第 l+N 個音框，這 2N+1 個音框分別對其取頻譜絕對值（Amplitude Spectrum）後，在第 k 個頻率下，這 2N+1 個頻域絕對值內的最大值。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 的定義如下

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

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

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

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

$$\gamma=\begin{cases}\gamma_0 & E\le E_0\\ \gamma_0+\dfrac{\gamma_1-\gamma_0}{E_1-E_0}(E-E_0) & E_0<E<E_1\\ \gamma_1 & E\ge E_1\end{cases} \qquad (3\text{-}4)$$

其中 E_0 和 E_1 代表了在最乾淨和最吵雜的情況下，雜訊的能量，而 E 是指現階段雜訊的能量。γ_0 和 γ_1 代表在最乾淨和最吵雜的情況下與 LTSD 比較的臨界值，因此 E_0, E_1, γ_0 和 γ_1 是先設定好的初始值。從（3-4）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。

3.2 以 Normalize Least Mean Square 為基礎之語音純化系統

於圖六架構中，當須調整 Upper Beamfomer 的係數時，本計劃所使用的演算法為以 NLMS(Normalize Least Mean Square)[5-6]為基礎的適應性空間濾波器，其空間濾波器的輸出可線性模型為（3-5）式

$$r(n)=\boldsymbol{x}(n)^{T}\boldsymbol{w}+e(n) \qquad (3\text{-}5)$$

其中

$$\boldsymbol{x}(n)=\left[\boldsymbol{x}_1(n)\ \cdots\ \boldsymbol{x}_M(n)\right]^{T}$$

$$\boldsymbol{x}_i(n)=\left[x_i(n)\ \cdots\ x_i(n-P+1)\right]^{T} \qquad (3\text{-}6)$$

$$\boldsymbol{w}=\left[w_{11}\ \cdots\ w_{1P}\ \cdots\ w_{M1}\ \cdots\ w_{MP}\right]^{T} \qquad (3\text{-}7)$$

M 代表麥克風個數，P 為每各麥克風的濾波階數，T 為矩陣的轉秩運算，$r(n)$ 為目標訊號，$\boldsymbol{x}(n)$ 為 $MP\times1$ 的訓練向量，由預錄的目標訊號與雜訊訊號所線性組合而成，$e(n)$ 為未知的誤差，$\boldsymbol{w}$ 為空間濾波器係數，大小為 $MP\times1$。而 $\boldsymbol{w}$ 的遞迴關係式如（3-8）所示

$$\hat{\boldsymbol{w}}(n+1)=\hat{\boldsymbol{w}}(n)+\frac{\mu}{\delta+\left\|\boldsymbol{x}(n+1)\right\|^{2}}\boldsymbol{x}(n+1)\times\left[r(n+1)-\boldsymbol{x}(n+1)^{T}\hat{\boldsymbol{w}}(n)\right] \qquad (3\text{-}8)$$

其中 δ 為一微小正數，使（3-8）式中分母不為零，$0<\mu<2$ 使（3-8）式能收斂。

3.3 以 H∞為基礎之語音純化系統

本子計畫亦將著名的 H∞理論[7-9]應用於語音純化系統中。在圖一的架構中，空間濾波器的輸出可線性模型為

$$r(n)=\boldsymbol{x}^{T}(n)\boldsymbol{w}+e(n) \qquad (3\text{-}9)$$

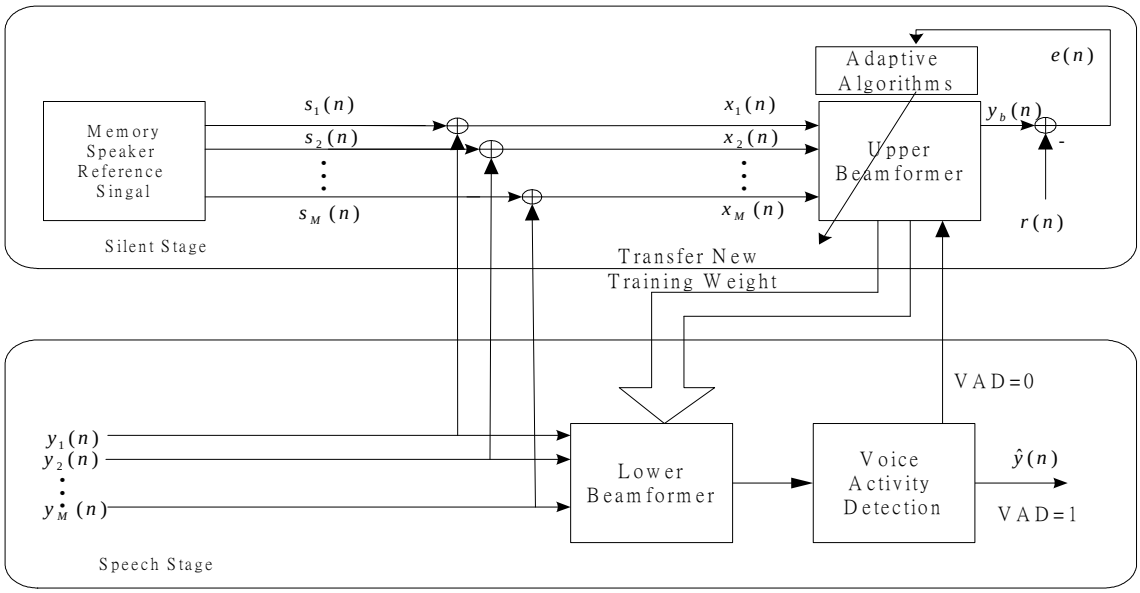
其中

$$\boldsymbol{x}(n)=\left[\boldsymbol{x}_1(n)\ \ \cdots\ \ \boldsymbol{x}_M(n)\right]^T$$

$$\boldsymbol{x}_i(n)=\left[x_i(n)\ \ \cdots\ \ x_i(n-P+1)\right]^T \qquad (3\text{-}10)$$

$$\boldsymbol{w}=\left[w_{11}\ \ \cdots\ \ w_{1P}\ \ \cdots\ \ w_{M1}\ \ \cdots\ \ w_{MP}\right]^T \qquad (3\text{-}11)$$

M 代表麥克風個數，P 為每各麥克風的濾波階數，T 為矩陣的轉秩運算，$r(n)$ 為目標訊號，$\boldsymbol{x}(n)$ 為 $MP\times1$ 的訓練向量，由預錄的目標訊號與雜訊訊號所線性組合而成，$e(n)$ 為未知的誤差，$\boldsymbol{w}$ 為空間濾波器



圖六. 結合語音活動偵測與適應性陣列訊號處理架構圖

係數，大小為 $MP\times1$。而 H∞的限制函式為（3-12）

$$\min_{\hat{\boldsymbol{w}}(n)}\max_{(e(n),\hat{\boldsymbol{w}}(0))} J=-\frac{1}{2}\xi^2\mu_0^{-1}\left|\boldsymbol{w}-\hat{\boldsymbol{w}}(0)\right|^2 +\frac{1}{2}\sum_{n=0}^{N}\left[\left|\boldsymbol{w}-\hat{\boldsymbol{w}}(n)\right|^2-\xi^2\left|e(n)\right|^2\right] \qquad (3\text{-}12)$$

μ_0 為權重參數 $|\cdot|^2$ 代表 2-norm 的平方，根據[10]，$\boldsymbol{w}$ 的遞迴關係式如（3-13）所示

$$\mathbf{M}^{-1}(n+1)=\mathbf{M}^{-1}(n)+\boldsymbol{x}(n)\boldsymbol{x}^T(n)-\xi^{-2}\mathbf{I} \qquad (3\text{-}13)$$

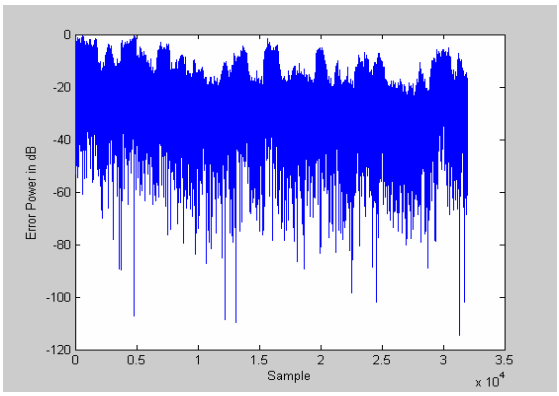
$$\hat{\boldsymbol{w}}(n+1)=\hat{\boldsymbol{w}}(n)+\mathbf{M}(n)\boldsymbol{x}(n)\frac{(d(n)-\boldsymbol{x}^T(n)\hat{\boldsymbol{w}}(n))}{(1+\boldsymbol{x}^T(n)\mathbf{M}(n)\boldsymbol{x}(n))} \qquad (3\text{-}14)$$

$$\hat{\boldsymbol{w}}(0)=\boldsymbol{0},\ \mathbf{M}^{-1}(0)=(\mu_0^{-1}-\xi^{-2})\mathbf{I} \qquad (3\text{-}15)$$

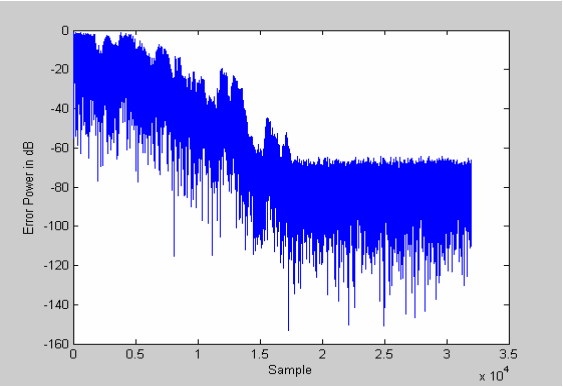
其中 $\mathbf{M}(n)$ 為 $MP\times MP$ 矩陣 $(\cdot)^{-1}$ 為反秩運算. 為了使 $\mathbf{M}(n)$，為正定 ξ 應選為 $\mathbf{M}^{-1}(n)+\boldsymbol{x}(n)\boldsymbol{x}^T(n)-\xi^{-2}\mathbf{I}>0$ 。

3.4 Normalize Least Mean Square 與 H∞模擬比較

做 NLMS 演算法時，必須先針對雜訊作出零平均的假設，若雜訊本身並非零平均，則 NLMS 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零平均，而 H∞並不用針對雜訊做出零平均的假設，但其運算量會較 NLMS 大。在基本原理上，NLMS 的原理是將估測誤差能量最小化，而 H∞是在基於量測誤差、模組誤差、初使誤差這三種誤差影響估測誤差最嚴重的情況下，去調整係數使誤差最小化。在此模擬中，本子計劃模擬聲源訊號經過二十階的通道效應，並由六顆麥克風接收。而每顆麥克風的有限脈衝響應（FIR）階數皆為十階來模型目標訊號。圖七與圖八展示了利用 NLMS 與 H∞所求得的誤差比較圖。誤差為目標訊號與模型訊號的差，即為模型誤差（modeling error），從圖中可看出 H∞ 理對於模型誤差較 NLMS 為穩健。



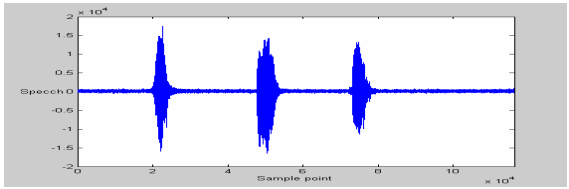
圖七　以 NLMS 為基礎的模型誤差能量



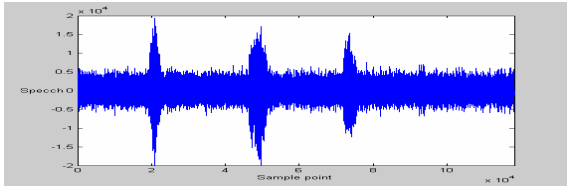
圖八　以 H∞為基礎的模型誤差能量

3.5 語音純化系統實驗結果

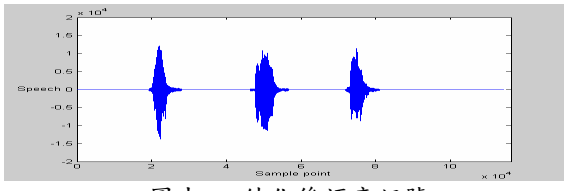
本實驗環境為將本子計畫所開發出的麥克風陣列平台放置於室內環境，而語音訊號從麥克風陣列正前方0°位置發出，白雜訊（white noise）由左方45°位置發出。圖九為無雜訊之純語音訊號，圖十為摻有雜訊之語音訊號，圖十一為純化後的語音訊號。語音純化系統會將 VAD 判定為非語音部分輸出為零。



圖九 無雜訊之純語音訊號



圖十　摻有雜訊之語音訊號



圖十一 純化後語音訊號

3.6 結論

本子計畫已將語音活動偵測（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。本子計畫亦將著名的 H∞ 理論套用於語音純化中，並從模擬中得知 H∞ 理對於模型誤差較 NLMS 為穩健。並由實驗結果展示，本系統能對特定聲源位置抑制干擾源與粹取人聲，並提升語音 SNR。

四、計畫成果自評

在計畫書中所列舉之項目均已執行，並將結果開始與其他子計畫作初步之整合。

五、參考文獻

[1] Hu, J., Cheng, Chieh-Cheng, Yang, Chia-Hsing, Su, Tzung-Min and Liu, W.H, “Removal of Background Information for Dynamic Speech and Image surveillance system,” *CACS Automatic Control Conference, Nov. 18 - 19, 2005.*

[2] J. Canny, “A Computational Approach to Edge Detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-8, No.6, 1986.

[3] C. Xu and J. L. Prince, “Gradient Vector Flow: A New External Force for Snakes,” *IEEE Conference on Computer Vision and Pattern Recognition*, 66-71, 1997.

[4]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

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

[6] 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

[7] Sayyarrodsari, B., How, J.P., Hassibi, B., and Carrier, A., “Estimation-based synthesis of H∞-optimal adaptive FIR filters for filtered-LMS problems,” IEEE Transactions on Signal Processing, vol. 49, pp.164-178, Jan 2001.

[8] De Soua, C.E., Palhares, R.M., and Peres, P.L.D., “Robust H∞ filter design for uncertain linear systems with multiple time-varying state delays,” IEEE Transactions on Signal Processing, vol.49, pp. 569-576, March 2001.

[9] U. Shaked and Y. Theodor, “H -optimal estimation: A tutorial,” in Proc. 31st IEEE Conf. Decision Contr., Tucson, AZ, Dec. 1992, pp.2278–2286.

[10] X. Shen and L. Deng, “A dynamic system approach to speech enhancement using the H∞ filtering algorithm,” IEEE Trans. Speech and Audio Process., vol. 7, pp. 391-399, July 1999.