# 結合麥克風陣列及模型調整技術之遠距離語音辨識系統

# Far-Distant Speech Recognition System Using Combined Techniques of Microphone Array and Model Adaptation

賴建瑞　簡仁宗

國立成功大學資訊工程學系

Email：jtchien@mail.ncku.edu.tw

## 摘要

本篇論文提出一種可應用於噪音環境下麥克風陣列(Microphone Array)的語音辨識演算法，其主要的目的在於克服傳統電腦語音辨識系統需要使用者頭戴或手持麥克風的不方便。為了消除遠距離麥克風的噪音干擾，我們的方法是先將每個麥克風收集到的語音，利用語音到達每個麥克風角度的不同，使用 Time Domain Cross Correlation (TDCC)演算法找出語者發音的方向及語音到達每個麥可風的時間延遲，再應用 Delay-and-Sum Beamformer 陣列訊號處理技術將語音訊號加強，最後我們再將加強過的語音訊號和語音模型參數間的不匹配用最佳相似度線性回歸(MLLR)的模型調整演算法來克服。在噪音環境下使用麥克風陣列之連續數字辨識實驗中，我們提出來的方法對於提升辨識率有良好的效果。

## 1. 導論

現實生活環境中，充滿了各式各樣的噪音和回音，這些干擾會嚴重的降低語音辨認系統的效能，其中之一的解決方式是使用頭戴式麥克風(Head-Mounted Microphone)，使得聲音源和麥克風盡可能的靠近，來降低環境噪音和回音的影響。然而使用頭戴式麥克風設備會造成使用者的不便，因此如何發展以免持式麥克風(Hands-Free Microphone)為主的語音辨認系統已成為一個相當重要的研究課題。

基本上，使用麥克風陣列可以進行遠距離錄音，因此可以解決頭戴式麥克風造成使用者不便的問題，而我們常用的麥克風陣列訊號處理技術是採用 Delay-and-Sum Beamformer，它可以克服環境噪音和回音對語音訊號的影響，還原出乾淨的語音。而且此一技術並非針對特定噪音環境，它可適用於任何噪音環境下，得到令人滿意的效果。在本論文中，我們將麥克風陣列應用於降低汽車環境噪音的干擾，以達到提高語音辨認率之目的。

一般的語音辨認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汽車裡，語音辨識的效果將大打折扣，因此，如何抑制噪音並加強語音訊號已成為汽車語音辨識的關鍵性技術。因此本論文中我們使用一組遠距離麥克風陣列做語音訊號輸入，然後使用 TDCC 將每個麥克風之間的時間延遲計算出來，再利用 Delay-and-Sum Beamformer 的方式，得到一組具抗噪性且加強過後的語音訊號。為了使加強過的語音訊號在進行隱藏式馬可夫模型(Hidden Markov Model, HMM)為主語音辨識時有更佳的辨識效果，我們使用最佳相似度線性回歸理論(maximum likelihood linear regression, MLLR) (Leggetter and Woodland, 1995)將原始語音訓練出來的隱藏式馬可夫模型參數做調整，以補償測試語音與模型參數之間的不匹配。

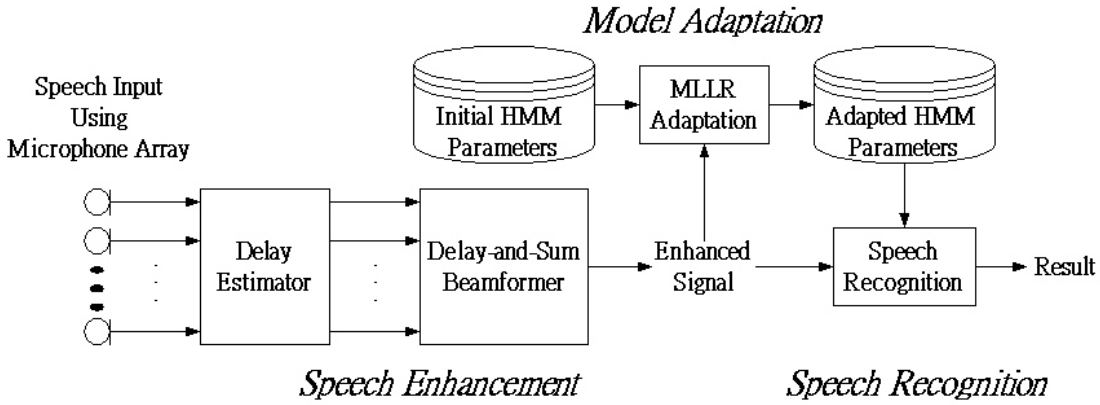
目前台灣的學術研究機構對於架構於麥克風陣列上的語音處理技術尚屬起步階段，在中文語音辨識上的應用發表在相關學術會議及期刊論文尚不多見，然而，國外的研究機構則早已投入此一領域，並且獲得不錯的成果，比較有名的包括三大類的方法，第一類是著重在不同麥克風間時間延遲的計算，它主要是利用語者定位演算法來計算不同麥克風間的時間延遲(Yamada et al., 1996; Inoue et al., 1997)，以及利用不同麥克風之間的頻譜能量來找出不同麥克風的時間延遲(Omologo & Svaizer, 1994, 1996; Giuliani et al., 1996)。第二類的方法是將麥克風間時間延遲併入隱藏式馬可夫模型的參數，它的觀念是擴充傳統語音隱藏式馬可夫模型的參數，加入各種不同的語者角度，並使用一種三維的維特比演算法(Three-Dimensional Viterbi Search)作語音辨識(Yamada et al., 1996, 1998a, 1998b, 1999)。第三類方法是將常用的語音增強技術和麥克風間時間延遲的估測作結合，主要是使用多頻(Multiband)的技術，將語音訊號分成數個不同的頻帶，在各個頻帶上作 Delay-and-Sum Beamformer 然後再將各個頻帶的訊號作合成，產生出強健性的加強語音訊號(Mahmoudi, 1998)。

## 2. 麥克風陣列語音辨識系統

麥克風陣列語音辨認系統架構圖如圖一所示，可分為語音加強、語音辨識及模型參數調整三部分。

在語音加強部分，輸入語音是由麥克風陣列錄得每個麥克風收集到的語句，再利用語音到達每個麥克風角度的不同，找出語者發音的方向及語音到達每個麥克風的時間延遲，應用 Delay-and-Sum Beamformer 的陣列訊號處理技術將原始語音做訊號加強。本論文中我們提出

TDCC 的演算法來計算時間延遲並於實驗中和其它演算法作比較。另外在語音辨識部分，我們是使用傳統的隱藏式馬可夫模型和一階段(One-Pass)演算法，利用最佳相似度(Maximum Likelihood, ML)法則來進行連續語音辨識。




圖一、麥克風陣列語音辨識系統架構圖

第三部分是模型參數的調整，一般較流行的調整方式有兩種，分別為最佳事後機率(Maximum *A Posteriori*, MAP)調整演算法(Gauvain and Lee, 1994)和最佳相似度線性回歸(Maximum Likelihood Linear Regression, MLLR)演算法(Leggetter and Woodland, 1995)。MAP 和 MLLR 都可以依據目前的測試語料來動態的對語音模型進行調整，主要的分別為 MAP 利用最佳事後機率法則來對語音模型參數做調整，調整的部分為測試語音所對應的狀態及混合數的 HMM 參數，MLLR 則是利用線性回歸的方式根據測試語料來對語音模型進行調整，它是藉由估測出的線性回歸函數，調整所有狀態及混合數的 HMM 參數。本論文中我們使用的調整技術為 MLLR。

## 2.1 Delay-and-Sum Beamformer

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)從我們偵測出的最佳方向 $\theta_s$ 傳播過來，麥克風的輸出為 $\mathbf{x}_t^i$，$1 \le i \le M$，則在時間 $t$ 的時候，當第 $i$ 支麥克風收到平面波的訊號，第 $i$+1 支麥克風則需等到聲波再前進距離 $R$ ($R = d\cos\theta_s$)方可收到訊號，如圖二所示。

若聲波的速度為 $C$，則第 $i$+1 個麥克風延遲的時間

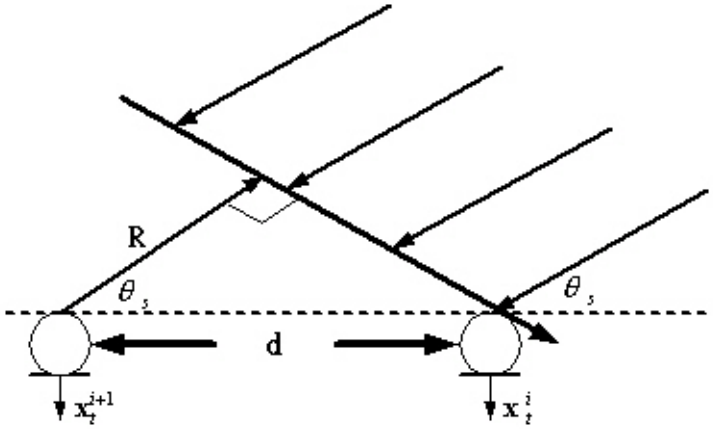
$$\tau = \frac{R}{C} = \frac{d\cos\theta_s}{C} \tag{1}$$

亦即 $\mathbf{x}_t^i = \mathbf{x}_{t+\tau}^{i+1}$，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：

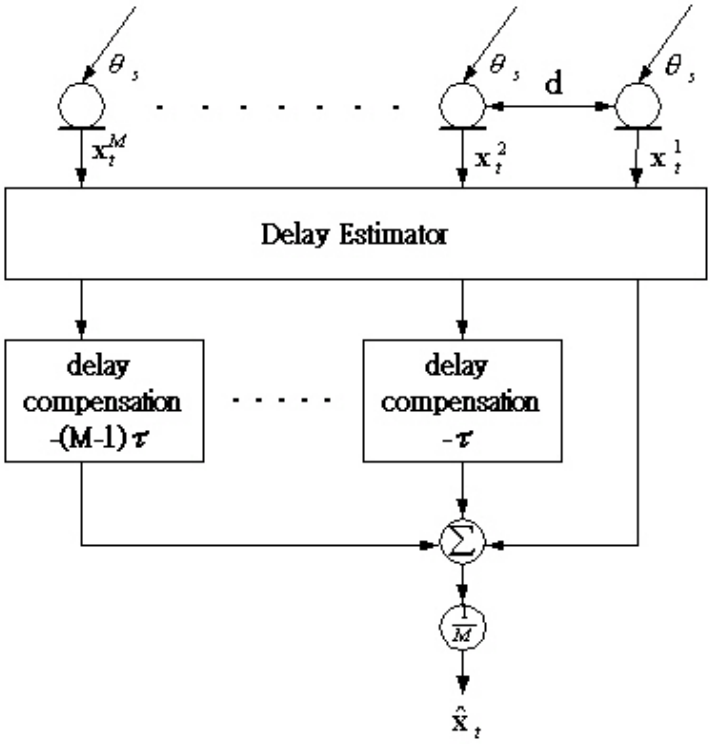
$$\mathbf{x}_t^i = \mathbf{x}_{t+(i-1)\tau}^1 \tag{2}$$

而整個 Delay-and-Sum Beamformer 的輸出 $\hat{\mathbf{x}}_t$，如圖三所示，就是將每個麥克風間的時間延遲作補償後合成再取平均而得

$$\hat{\mathbf{x}}_t = \frac{1}{M}\sum_{i=1}^{M} \mathbf{x}_{t+(i-1)\tau}^i \tag{3}$$



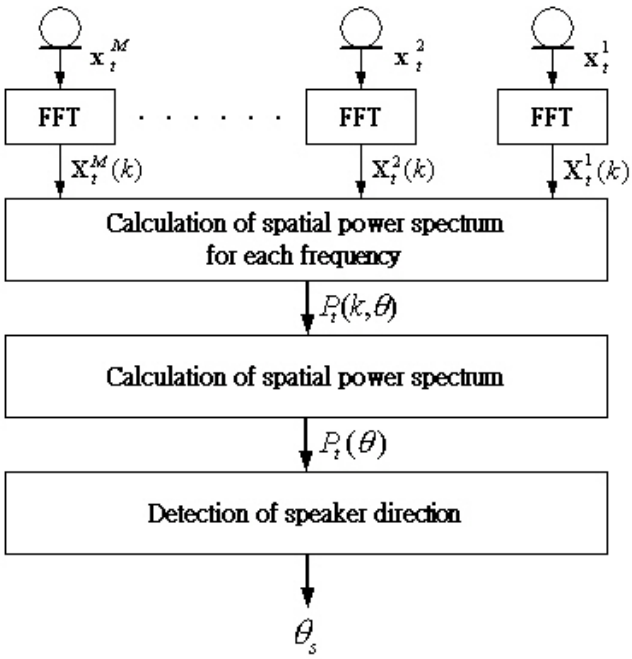

圖二、相鄰麥克風的時間延遲




圖三、Delay-and-Sum Beamformer 流程圖

## 2.2語者定位演算法(Speaker Localization Algorithm, SLA)

語者定位演算法的主要目的在於估測出語者發話的方向，其系統流程圖如圖四所示，我們將分成下列三部份做說明。




圖四、語者定位演算法(SLA)流程圖

首先，我們將 $M$ 個麥克風在時域上的語音訊號 $\{\mathbf{x}_t^i, i=1,...,M\}$ 經過快速傅利葉轉換後得到的麥克風在頻率上的訊號 $\{\mathbf{X}_t^i(k), i=1,...,M, k=0,...,K-1\}$，其中 $i$ 表示麥克風的引數，$k$ 表示頻率的引數，$t$ 表示音框的引數。

第二部分，我們計算不同聲音方向角度 $\theta=1,...,180$ 的空間功率頻譜

$$P_t(\theta)=\sum_{k=0}^{K-1}P_t(k,\theta)\ ,\ \theta=1,\cdots,180 \tag{4}$$

其中

$$P_t(k,\theta)=\left|\sum_{i=1}^{M}\mathbf{X}_t^i(k)\exp\{j2\pi f_k(i-1)\frac{d\cos\theta}{c}\}\right|^2 \tag{5}$$

$f_k$ 表示 $k$ 所對應的頻率，$d$ 表示相鄰麥克風的間距，$c$ 表示聲波速度。

第三部分我們做語者方向的偵測，基本上，語者方向 $\theta_s$ 的偵測是去尋找空間功率頻譜中最大的空間功率頻譜所對應的角度，也就是進行以下的運算

$$\theta_s = \arg\max_{\theta} P_t(\theta) \tag{6}$$

其中，$\theta$ 為聲音方向的隨機變數。

求出$\theta_s$後再利用 2.1 節中的 Delay-and-Sum Beamformer 即可求出相鄰麥克風間的時間延遲 $\tau$，對每一個麥克風做時間延遲補償後即可得到加強過後的語音訊號。

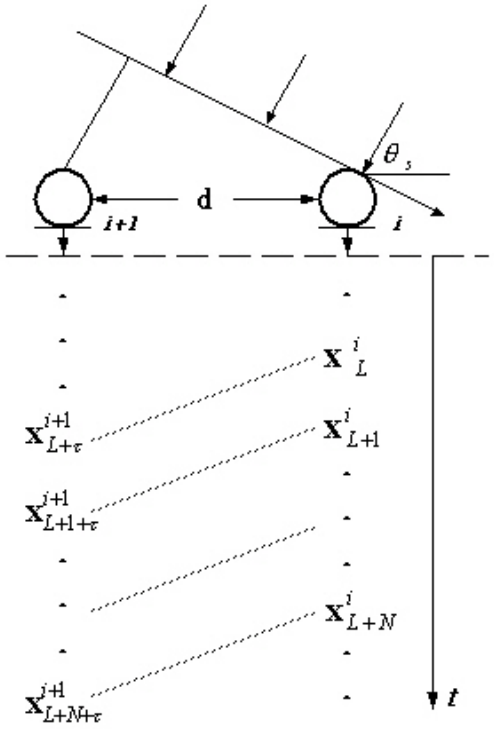
## 2.3 Time Domain Cross Correlation (TDCC)

不同於語者定位演算法，TDCC 是直接於時域上利用不同麥克風語音間的相關性來求取時間延遲。其基本想法是假設第 $i$ 個麥克風和其所相鄰的第 $i$+1 個麥克風在第 $L$ 個數位點後的語音訊號分別表示如下：

$$\mathbf{x}_L^i, \mathbf{x}_{L+1}^i, ..., \mathbf{x}_{L+N}^i \quad 和 \quad \mathbf{x}_{L+\tau}^{i+1}, \mathbf{x}_{L+1+\tau}^{i+1}, ..., \mathbf{x}_{L+N+\tau}^{i+1}$$

其中我們取出 $N$ 個數位點，如圖五所示。

在不考慮噪音以及訊號衰減的情形下，若 $\tau$ 為麥克風 $i$ 和麥克風 $i+1$ 間的時間延遲，則 $\mathbf{x}_t^i$ 和 $\mathbf{x}_{t+\tau}^{i+1}$ 之間具有最大的相關性且 $\sum_{t=L}^{L+N} \mathbf{x}_t^i \cdot \mathbf{x}_{t+\tau}^{i+1}$ 點積和為最大，此一乘積和可稱之為 Time Domain Cross Correlation。




圖五、TDCC示意圖

經由以上的想法我們發展出 TDCC 的演算法：若現有一麥克風陣列包含有 $M$ 個麥克風，麥克風 $i$ 於時間 $t$ 所收到的訊號稱為 $\mathbf{x}_t^i$，則對語音訊號中任一音框 $m$ 的 TDCC 定義如下

$$C(m)=\sum_{i=2}^{M}\sum_{j=1}^{N}\mathbf{x}_{(m-1)\cdot p+j}^{1}\cdot\mathbf{x}_{(m-1)\cdot p+j}^{i} \tag{7}$$

其中 $N$ 為音框內所包含的點數，$P$ 為音框間的位移點數。我們以第一個麥克風為基準麥克風，所以式子(6)中麥克風的引數 $i$ 從 2 開始累加。若 $\tau$ 為時間延遲的隨機變數，則麥克風陣列中相鄰麥克風間的最佳時間延遲 $\hat{\tau}_H$ 為

$$\hat{\tau}_H=\arg\max_{\tau}C(m,\tau) \tag{8}$$

其中

$$C(m,\tau)=\sum_{i=2}^{M}\sum_{j=1}^{N}\mathbf{x}_{(m-1)\cdot p+j}^{1}\cdot\mathbf{x}_{(m-1)\cdot p+j+(i-1)\tau}^{i} \tag{9}$$

這裡我們是以語句中能量最高的音框為基準來計算時間延遲，另外若將語句內全部音框的 TDCC 累加起來，根據此累加值則相鄰麥克風間的時間延遲 $\hat{\tau}_A$ 為

$$\hat{\tau}_A=\arg\max_{\tau}\sum_{m}C(m,\tau) \tag{10}$$

我們在後面的實驗部分會針對此二種方法分別做實驗並分析其結果。計算出相鄰麥克風間的時間延遲 $\hat{\tau}$ 後，再利用 2.1 節中的 Delay-and-Sum Beamformer 即可求出加強過後的語音訊號。

## 2.4 最佳相似度線性回歸演算法

MLLR 是一種常見使用於語音模型參數調整的技術，它是從測試語料計算出一個轉移矩陣，然後利用此轉移矩陣來調整語音模型中每一個狀態及混合數的平均值向量。

一般我們常使用高斯機率密度函數來表示隱藏式馬可夫模型的觀測機率

$$P(o_t\mid\mu_s,\Sigma_s)=\frac{1}{(2\pi)^{n/2}\left|\Sigma_s\right|^{1/2}}e^{-1/2(o_t-\mu_s)'\Sigma_s^{-1}(o_t-\mu_s)} \tag{11}$$

其中 $\mu_s$ 表示平均值向量，$\Sigma_s$ 表示變異數矩陣，$o_t$ 為觀測到的特徵向量，$n$ 為向量的維度。

定義一個大小為 $n\times(n+1)$ 的轉移矩陣 $W_s$，它可將擴展後的平均值向量 $\xi_s$ 調整而得到新的平均值向量

$$\hat{\mu}_s=W_s\xi_s \tag{12}$$

其中 $\xi_s = [\omega, \mu_1, \mu_2, ..., \mu_n]'$，$\omega$ 是在進行回歸計算時考慮是否使用偏差量(使用則 $\omega$ 為 1，不使用則為 0)。因此調整過後的高斯機率分佈如下所示

$$P(o_t \mid W_s, \mu_s, \Sigma_s) = \frac{1}{(2\pi)^{n/2}\left|\Sigma_s\right|^{1/2}} e^{-1/2(o_t - W_s\xi_s)'\Sigma_s^{-1}(o_t - W_s\xi_s)} \tag{13}$$

根據最佳相似度法則，最佳轉移矩陣為

$$\hat{W}_s = \arg\max_{W_s} P(o_t \mid W_s, u_s, \Sigma_s) \tag{14}$$

這裡我們簡化轉移矩陣內的參數為

$$W_s = \begin{pmatrix} w_{1,1} & w_{1,2} & 0 & .. & 0 \\ w_{2,1} & 0 & w_{2,3} & .. & 0 \\ : & & & & : \\ w_{n,1} & .. & .. & 0 & w_{n,n+1} \end{pmatrix} \tag{15}$$

也就是將轉移矩陣改寫為以下的轉移參數向量

$$w_s = [w_{1,1}, ..., w_{n,1}, w_{1,2}, w_{2,2}, ..., w_{n,n+1}]' \tag{16}$$

那麼最佳轉移參數向量可以由 EM 演算法(Dempster et al.,1977)推導而得

$$\hat{\mathbf{w}}_s = \left[\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{s_r}(t) D'_{s_r} C^{-1}_{s_r} D_{s_r}\right]^{-1}\left[\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{s_r}(t) D'_{s_r} C^{-1}_{s_r} o_t\right] \tag{17}$$

其中 $r$ 為狀態的索引，$t$ 為時間的索引，$\gamma_{s_r}$ 為一事後機率，若 $o_t$ 經由 Viterbi Decoding 後對應到狀態 $s_r$ 則其值為 1 否則為 0。$D_s$ 定義為

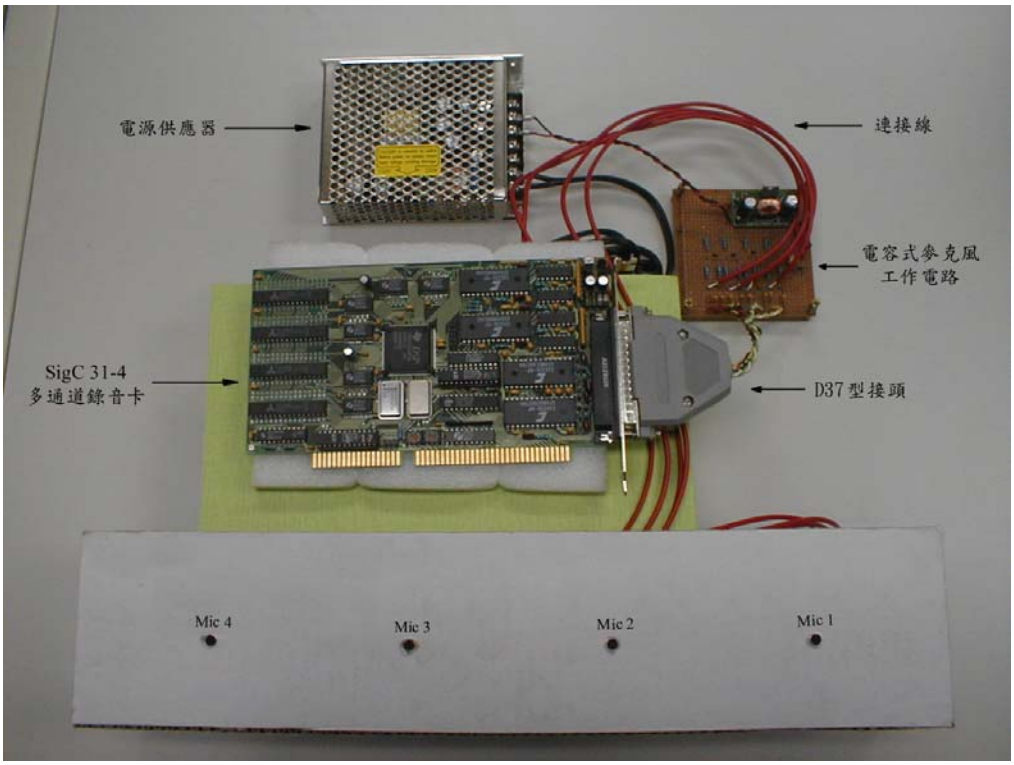
$$D_s = \begin{pmatrix} \omega & 0 & .. & .. & 0 & \mu_1 & 0 & .. & .. & 0 \\ 0 & \omega & 0 & .. & .. & 0 & \mu_2 & 0 & .. & 0 \\ : & & & & & & & & & : \\ 0 & .. & 0 & \omega & 0 & .. & .. & 0 & \mu_{n-1} & 0 \\ 0 & .. & .. & 0 & \omega & 0 & .. & .. & 0 & \mu_n \end{pmatrix} \tag{18}$$

# 3. 實驗結果

## 3.1 麥克風陣列錄音設備

麥克風陣列錄音設備如圖六所示，主要的部分包含多通道錄音卡 SigC31-4、四個全方向電容式麥克風、供給麥克風運作的電源供應器和工作電路以及必要的連接線。

多通道錄音卡 SigC31-4 是由美國 Signalogic 公司所生產的，為一 4 個通道的錄音卡，使用的數位訊號處理晶片(DSP processor)為德州儀器公司(TI)所生產的 TM8320C31，可同時提供 4 個通道進行錄音的動作，此錄音卡的介面為 ISA 介面可裝於個人電腦上，並有提供 D37 型接頭經由工作電路和麥克風相連接，透過所附的軟體即可利用 4 個麥克風同時錄音，電容式麥克風我們使用國內音賜公司所生產的全方向電容式麥克風(Omni-directional Condenser Microphone)，型號為 ECM9D，所對應的頻寬為 20~10000Hz，靈敏度為-38±3dB，訊噪比(signal-to-noise ratio, SNR)大於 60dB，工作電壓則介於 DC 3V 至 DC 10V 之間。工作電路主要的作用是將電源供應器所提供的電力進行穩壓，之後再送至麥克風提供錄音時所需的電力，並保持麥克風錄音時訊號的穩定，並將訊號送至多通道錄音卡 SigC31-4。此工作電路是依據音賜公司針對電容式麥克風的建議電路稍加修改後由我們自行焊接製作的。



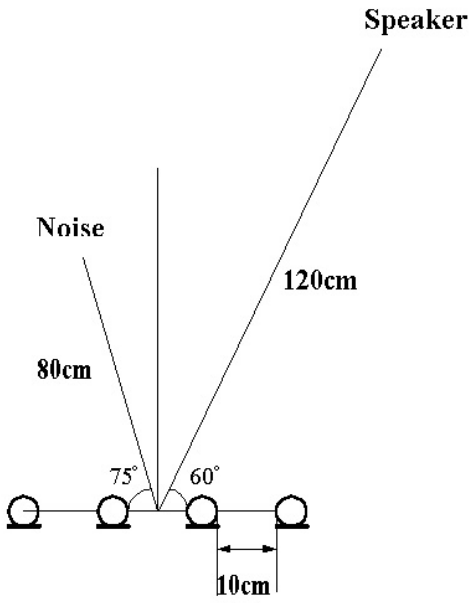

圖六、麥克風陣列錄音設備連接圖

## 3.2 語料庫

我們使用的語音特徵參數為 12 階 MFCC 和 12 階 delta MFCC 和 1 階 delta log energy 和 1 階 delta delta log Energy 共 26 階。訓練語料是在一般辦公室環境下使用近距離麥克風所錄製的，共有 1400 句中文連續數字，其中包含 70 位男生和 70 位女生。語音模型是使用隱藏式馬可夫模型來表示，每一個中文數字使用 7 個狀態，背景噪音則使用 3 個狀態，分別表示音

檔前後的噪音和數字間的噪音，所以總共的狀態數目為 73 個。每一個狀態包含 4 個混合數，因此共有 292 個混合數。

測試語料是在實驗室中使用遠距離麥克風陣列所錄製，我們模擬了三種不同車速的路況，分別為 0 km/h、50 km/h 和 90 km/h。在 0 km/h 路況下不加任何噪音，而 50 km/h 和 90 km/h 路況則利用喇叭放出汽車於時速 50 km/h 和 90 km/h 時所錄下的噪音來模擬。錄音時語者距離麥克風中心約 120 公分和麥克風陣列的夾角約為 60 度，噪音源則擺放於距離麥克風中心約 80 公分處和麥克風的的夾角約 75 度，麥克風陣列是線性配置的，相鄰麥克風的間距為 10 公分。語者和麥克風陣列以及噪音的相對位置如圖七所示。總共有 15 人參與錄音，包含 12 位男生和 3 位女生，每一種路況有 30 句不同的中文連續數字。每種路況總共錄得 450 句音檔。連續語音辨認所使用的演算法為一階段演算法。實驗結果我們以數字錯誤率(Digit Error Rate)來表示。




圖七、錄音時麥克風陣列和語者以及噪音間的相對位置圖

## 3.3 Delay-and-Sum Beamformer實驗結果

對於每一個麥克風(Mic1, Mic2, Mic3, Mic4)所收集的語音訊號其個別的字元錯誤率以及錯誤率的平均值如表一所示，此結果可視為基本系統(Baseline)的錯誤率。在此我們使用所有麥克風辨認錯誤率的平均值作為麥克風陣列的整體錯誤率。

我們所進行的第一組實驗是事先預設一些聲音源角度值來進行實驗以找出最佳效果的角度，這裡我們將聲音源的方向固定從 30° 到 150° 每間隔 30° 做一次實驗，再對經由

Delay-and-Sum Beamformer 處理後的語音訊號分別計算其辨認結果，辨認結果如表二所示。觀察實驗結果我們可以發現在不同路況下最佳辨認率都出現在60°，此一結果和我們實際上錄製語音時的方向是十分吻合的。

| 麥克風／Digit Error Rate (%)／路況 | 0 km/h | 50 km/h | 90 km/h |
|---|---|---|---|
| Mic 1 | 47.0 | 52.1 | 55.1 |
| Mic 2 | 42.0 | 48.3 | 53.4 |
| Mic 3 | 51.0 | 54.8 | 58.7 |
| Mic 4 | 46.5 | 53.0 | 57.2 |
| Mic 平均 | **46.6** | **52.1** | **56.1** |

表一、基本系統的辨認結果

| 路況 / Digit Error Rate (%) / 角度 | 30° | 60° | 90° | 120° | 150° |
|---|---|---|---|---|---|
| 0 km/h | 35.2 | **31.9** | 60.6 | 58.6 | 55.4 |
| 50 km/h | 40.9 | **38.1** | 66.9 | 68.3 | 62.6 |
| 90 km/h | 42.8 | **40.4** | 70.0 | 69.6 | 65.7 |

表二、不同聲音源角度下Delay-and-Sum Beamformer的辨認結果

| 演算法／Digit Error Rate (%)／路況 | 0 km/h | 50 km/h | 90 km/h |
|---|---|---|---|
| Mic 平均 | 46.7 | 52.1 | 56.1 |
| 固定角度 60° | **31.9** | **38.1** | **40.4** |
| SLA | 43.8 | 48.6 | 52.1 |
| TDCC H | 37.5 | 43.6 | 47.0 |
| TDCC A | **31.7** | **38.4** | **40.9** |

表三、SLA和TDCC辨認結果比較圖

第二組實驗是將所錄得的測試語料分別經由語者定位演算法和 TDCC 求取不同麥克風間的時間延遲，經過補償後產生加強過後的語音訊號，再分別進行辨認，實驗結果如表三所示。其中 TDCC 有兩種不同的計算方法，TDCC H 表示每一個語句僅使用最高能量的音框來計算時間延遲，而 TDCC A 則表示每一個語句的所有音框皆被考慮來計算時間延遲。此外表三亦列出麥克風陣列的平均辨識率和固定角度60°時的辨認錯誤率以方便比較。

觀察表三的辨認結果比較，我們可以發現使用 Delay-and-Sum Beamformer 的 SLA 和 TDCC 確實能夠有效降低錯誤率，而 TDCC 和傳統的語者定位演算法 SLA 相比，TDCC 更能有效降低辨認錯誤率，且 TDCC 使用所有的音框的效果不但比使用一個音框效果還好，而且

幾乎和固定角度60°的錯誤率相同，顯示 TDCC 對於計算時間延遲是十分有效的。

## 3.4 取樣頻率對SLA和TDCC的影響

經由以上的實驗結果，我們發現語者定位演算法的效能較差。仔細研究其原因後發現，語者定位演算法是先求出語者的方向$\theta_s$，再利用公式

$$b = (\text{Sampling Rate}) \cdot \tau = \frac{(\text{Sampling Rate}) \cdot d \cdot \cos\theta_s}{C} \tag{19}$$

來求出取樣點上的位移$b$。因為聲音的速度$C$和麥克風的間距$d$都是固定的，因此我們設計了一些實驗來瞭解取樣頻率對 SLA 和 TDCC 兩種演算法的影響。

實驗時我們先對測試語料利用內差法提高取樣頻率，經由 Delay-and-Sum Beamformer 求出增強過的語音訊號後，再將取樣頻率降為 8KHz。然後再進行辨識，辨認結果如表四(8KHz)、表五(16KHz)和表六(24KHz)所示。

我們可以發現 SLA 的辨認錯誤率有明顯的改變，取樣頻率由 8KHz 提高至 16KHz 和 24KHz 時錯誤率在 3 種不同路況都有顯著的下降。而提高至 16KHz 和提高至 24KHz 相比時則在 3 種不同路況上錯誤率僅有些許的改變。至於 TDCC 則因為其運算的對象就是時域上的取樣點，因此辨認錯誤率並無明顯改變。此一結果顯示由於 TDCC 不需要計算語者方向，因此可以適用於各種取樣頻率，能維持一定的辨認效果。

| 演算法／Digit Error Rate (%)／路況 | 0 km/h | 50 km/h | 90 km/h |
|---|---|---|---|
| SLA H | 43.79 | 48.64 | 52.12 |
| TDCC H | 37.50 | 43.58 | 47.03 |

表四、取樣頻率 8KHz 時 SLA 和 TDCC 的辨認結果

| 演算法／Digit Error Rate (%)／路況 | 0 km/h | 50 km/h | 90 km/h |
|---|---|---|---|
| SLA H | 34.94 | 39.97 | 42.00 |
| TDCC H | 36.90 | 42.69 | 47.09 |

表五、取樣頻率 16KHz 時 SLA 和 TDCC 的辨認結果

| 演算法／Digit Error Rate (%)／路況 | 0 km/h | 50 km/h | 90 km/h |
|---|---|---|---|
| SLA H | 34.17 | 39.52 | 42.79 |
| TDCC H | 38.27 | 44.08 | 49.07 |

表六、取樣頻率 24KHz 時 SLA 和 TDCC 的辨認結果

## 3.5 加入語音模型調整的實驗結果

接下來的實驗著重於瞭解語音模型調整對麥克風陣列語音辨認系統的影響。基本系統經由 MLLR 調整後的實驗結果如表七所示。此外，SLA 和 TDCC 分別加上 MLLR 的實驗結果如表八所示。

| 麥克風／Digit Error Rate (%)／路況 | 0 km/h | 50 km/h | 90 km/h |
|---|---|---|---|
| Mic 1 + MLLR | 29.0 | 31.7 | 33.4 |
| Mic 2 + MLLR | 25.5 | 29.9 | 33.1 |
| Mic 3 + MLLR | 31.6 | 33.4 | 36.2 |
| Mic 4 + MLLR | 28.3 | 31.4 | 34.5 |
| (Mic + MLLR)平均 | **28.6** | **31.6** | **34.4** |
| Mic 平均 | 46.6 | 52.1 | 56.1 |

表七、基本系統經由MLLR調整後的實驗結果

| 演算法／Digit Error Rate (%)／路況 | 0 km/h | 50 km/h | 90 km/h |
|---|---|---|---|
| (Mic + MLLR)平均 | 28.6 | 31.6 | 34.4 |
| SLA + MLLR | 29.9 | 31.5 | 35.1 |
| TDCC H + MLLR | 25.2 | 28.8 | 31.4 |
| TDCC A + MLLR | **21.1** | **24.9** | **28.2** |

表八、SLA和TDCC經由MLLR調整後的辨認結果比較圖

經由表七的實驗結果我們可以發現，基本系統使用 MLLR 的語音模型調整技術後，不管在哪一種路況都可有效的降低辨認錯誤率(0 km/h 由 46.64%降至 28.61%，50 km/h 由 52.07%降至 31.60%，90 km/h 由 56.12%降至 34.42%)，其原因為使用傳統麥克風於乾淨環境下所錄製的訓練語料和利用麥克風陣列於噪音環境下所錄製的測試語料間的不匹配現象相當嚴重。

分析表八的結果，我們發現加入語音模型調整的 SLA 和 TDCC 在降低辨認錯誤率上亦有顯著的效果，在三種不同路況上 TDCC 的效能仍然優於 SLA。最低的辨認錯誤率(21.10%)為路況 0 km/h 下使用全部音框來計算的 TDCC 演算法。

## 3.6 辨認時間的比較

辨認時間的計算是統計所有測試語料(共 1350 句，平均一句包含 6 個中文連續數字)經由時間延遲的計算、Delay-and-Sum Beamformer 的處理、語音特徵參數的求取和語音辨認的所有時間再做平均而得，實驗結果如表九所示。基本系統則僅計算特徵參數和語音辨認的時間

再平均。執行測試的電腦配備為 Pentium II 350 處理器和 128MB 記憶體的個人電腦，作業系統則為 Windows 98。觀察實驗結果，我們發現不管是僅使用最大能量的音框或是全部音框的 TDCC 演算法在執行速度上皆優於傳統的 SLA 演算法。

| | Baseline | SLA | TDCC H | TDCC A |
|---|---|---|---|---|
| Without MLLR | 0.28 | 0.58 | 0.41 | 0.56 |
| With MLLR | 0.50 | 0.79 | 0.63 | 0.77 |

表九、SLA 與 TDCC 執行速度之比較 (速度計算單位為秒／句)

## 4. 結論

本論文中我們建立一個應用麥克風陣列的語音辨認系統，此一系統利用 Delay-and-Sum Beamformer 來降低環境噪音對於語音訊號的影響。同時我們也提出了一個應用於麥克風陣列上計算時間延遲的演算法 TDCC。實驗的部分我們進行了基本系統的實驗、給定各種不同角度的實驗、取樣頻率改變的實驗、使用 SLA 演算法、使用最大能量音框的 TDCC 演算法和使用全部音框的 TDCC 演算法以及執行速度比較。經由實驗結果我們可以證明 TDCC 的有效性(在不同路況下平均約可降低 15%的辨認錯誤率)。和傳統的語者定位演算法 SLA 相比較，TDCC 不論是在辨認錯誤率降低的幅度上或執行速度上皆優於 SLA。

本論文中亦結合了語音模型調整的技術。經由實驗我們可以發現，單純只使用 MLLR 來調整語音模型即可獲得不錯的效果。然而若將麥克風陣列和語音模型調整的技術相結合，對於降低辨認錯誤率(在不同路況下平均約可降低 25%的辨認錯誤率)會產生更顯著的效果。從我們研究的結果，可以發現仍然還有許多值得研究的課題，如更精確語者方向的定位、麥克風位置的考量…等。未來我們將主要致力於研究麥克風陣列中麥克風的擺放位置和辨認率間的關係，以及實際將麥克風陣列的演算法應用在汽車環境或有回音、噪音的語音辨識系統上。

## 5. 參考文獻

[1] A. P. Dempster and N. M. Laird, D. B. Rubin. “Maximum Likelihood from Incomplete Data via the EM Algorithm”, *J. Roy. Stat. Soc.*, 39(1) : 1-38, 1977.

[2] J.-L. Gauvain and C.-H. Lee, “Maximum a Posterior Estimation for Multivariate Gaussian Mixture Observations of Markov Chains”, *IEEE Trans. Speech, Audio Processing*, Volume 2,

pages 291-298, April 1994.

[3] D. Giuliani, M. Omologo and P. Svaizer, “Experiments of Speech Recognition In a Noisy and Reverberant Environment Using a Microphone Array and HMM Adaptation”, In Proc. of ICSLP ’96, pages 1329-1332, October 1996.

[4] M. Inoue, S. NAKAMURA, T. YAMADA and K. SHIKANO, “Microphone Array Design Measures for Hands-Free Speech Recognition”, In Proc. of Eurospeech ’97, Volume 1, pages 331-334, September 1997.

[5] C. J. Leggetter and P. C. Woodland, “Maximun Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models”, Computer Speech and Language, Volume 9, pages 171-185, September 1995.

[6] D. Mahmoudi, “Combined Wiener and Coherence Filtering in Wavelet Domain For Microphone Array Speech Enhancement”, In Proc. of ICASSP ’98, pages 385-388, May 1998.

[7] M. Omologo and P. Svaizer, “Acoustic Event Localization Using a Crosspower-Spectrum Phase Based Technique”, In Proc. of ICASSP ’94, Volume 2, pages 273-276, 1994.

[8] M. Omologo and P. Svaizer, “Acoustic Source Location in Noisy and Reverberant Environment Using CSP Analysis”, In Proc. of ICASSP ’96, pages 921-924, 1996.

[9] T. YAMADA, S. Nakamura and K. Shikano, “Robust Speech Recognition with Speaker Localization by a Microphone Array”, In Proc. of ICSLP ’96, pages 1317-1320, October 1996.

[10] T. YAMADA, S. Nakamura and K. Shikano, “Hands-Free Speech Recognition Based on a 3-D Viterbi Search Using a Microphone Array”, In Proc. of ICASSP ’98, pages 245-248, May 1998a.

[11] T. YAMADA, S. Nakamura and K. Shikano, “An Effect of Adaptive Beamforming on 3-D Viterbi Search”, In Proc. of ICSLP ’98, pages 381-384, December 1998b.

[12] T. YAMADA, S. Nakamura and K. Shikano, “Simultaneous Recognition of Multiple Sound Sources Based on 3-D N-Best Search Using Microphone Array”. In Proc. of Eurospeech ’99, Volume 1, Page 69-72, September 1999.