International Journal of Information Acquisition
Vol. 2, No. 1 (2005) 1–22


World Scientific
www.worldscientific.com

# ROBUST SOUND LOCALIZATION USING SMALL NUMBER OF MICROPHONES

SHUZHI SAM GE*, AI POH LOH and FENG GUAN
*Department of Electrical and Computer Engineering*
*National University of Singapore, Singapore*

Received 10 January 2005
Accepted 31 January 2005

In this paper, robust sound localization is proposed for systems with spatially distributed microphones fewer than 4, the minimum required criteria for 3D sound localization in the literature, in an effort to address the robustness issues of microphone failures. It is shown that, for three- and two-microphone systems, two and four samples of the Interaural Time Difference (ITD) measurements are required respectively. In a one-microphone system, five samples of the Interaural Intensity Difference (IID) measurement should be used instead. These investigations are important because if one microphone fails, the remaining system can still function to locate the sound source without significant impact on the localization system. In addition, our investigation also shows that a 3-microphone system can estimate the azimuth and elevation simultaneously, which usually requires 4 microphones in the literature. Simulation and experimental results are presented to illustrate the performance of a three-microphone system. Results show that the system can locate the sound source with satisfactory accuracy.

*Keywords*: 3D sound localization; interaural time difference; multiple sampling.

## 1. Introduction

Sound localization plays important roles in humans daily life. To a passenger, the sound of a fast approaching vehicle warns him to steer clear of the dangerous traffic. Sound localization has attracted much attention in the literature [Rucci *et al.*, 1999; Weng & Guentchev, 2001; Wang *et al.*, 2004; Tabrikian & Messer, 1996] owing to the wide applications including robot perception [Huang *et al.*, 1997], human-machine interfaces [Buchner & Kellermann, 2001], handicappers' aids [Desloge *et al.*, 1997; Welker *et al.*, 1997] and some military applications [Ferguson *et al.*, 2002]. Due to the importance of sound localization, it has been greatly explored. Related works fall into the following categories:

- Smart audio sensors
  Psychoacoustic studies on humans have provided us with information about the cues to locate sound sources, namely, Interaural Time Differences (ITD), Interaural Intensity Differences (IID) and spectrum [Cook, 1999; Shinn-Cunningham *et al.*, 2000;

*To whom all correspondance should be sent. Tel.: (65) 6874-6821; Fax: (65) 6779-1103; E-mail: elegesz@nus.edu.sg

Morimoto, 2001]. To mimic human dimensional hearing, a neuromorphic microphone was proposed by making use of biologically-based monaural spectrum cues [Pu *et al.*, 1997]. This microphone relied on a specially shaped reflecting structure — a parabola simulating the pinna's curved surface and a sound source can be located in the front half plane by using echo-time processing. Based on the analysis of biological sound localization systems (such as the barn owl, bats, etc.), neural networks have been successfully used to locate sound sources with relative azimuth [Rucci *et al.*, 1999]. In [Huang *et al.*, 1995, 1997], a simplified model of the auditory system of human beings was developed, which consists of a three-microphone set and several banks of band-pass filters. Zero-crossing was used to detect the arrival temporal disparities and the histogram mapping method was introduced to localize multiple sound sources in the azimuth direction.

- Microphone arrays [Brandstein & Ward, 2001; Ferguson & Lo, 2002; Gazor & Grenier, 1995; Brandstien *et al.*, 1997]:
  (i) Relative sound localization
  It focuses on sound localization with respect to the reference frame attached to the microphone arrays. It can be grouped into three types, namely, localization based on beamformer, signal correlation matrix, and time difference of arrival (TDOA) [Brandstein, 1995] respectively. The first is similar to that of radar and can be achieved using beamformer-based energy scans [Katkovnik & Gershman, 2000; Gazor & Grenier, 1996], by which output power of a steered-beamformer should be maximized. The second makes use of high resolution spatio-spectral correlation matrix derived from signal received. TDOA-based localization covers the receptive environment of interest, instead of "focalization", based on high resolution TDOA estimation. Related techniques of relative sound localization include triangulation [Ferguson *et al.*, 2002], Least-Square (LS) approach [Huang *et al.*, 2001], Maximum likelihood technique [Chen *et al.*, 2002], sensor fusion [Chen *et al.*, 2000] and so forth.
  (ii) Global sound localization
  Global sound localization is similar to Simultaneous Localization and Mapping (SLAM) in the research domain of map building and robotics [Guivant & Nebot, 2003]. It focuses on the source localization with respect to the global reference frame. It consists of two main issues, namely, the relative source localization with respect to the microphone array and the localization of microphone array itself with respect to the global reference frame. The latter can be referred to as inverse sound localization [Aarabi, 2002] and has significant impact on the overall sound localization.

All these works assume that adequate/redundant audio sensors are provided to locate sound sources or the minimum number of audio sensors is determined for certain applications. For example, four microphones are required for 3D sound localization using both ITD and IID [Weng & Guentchev, 2001]. This raises the question that what if the number of audio sensors available is less than the minimum required. Similar problem exists in GPS system if satellite signals are blocked by tall buildings in urban environments and the available satellites are less than 4. Given the geometry constraints such as sea surface or road, [Cui & Ge, 2003] proposed a solution to GPS-based localization, by which the minimum number of satellites can be reduced to 2. All these considerations motivate us to investigate the robustness issues of microphone failures.

The remainder of this paper is organized as follows. Section 2 briefs the basic issues related to realistic ITD measurements. The capability of a 3-microphone system is discussed in details in Sec. 3. This is followed by the investigation of a 2-microphone system in Sec. 4. The principles of localization of a 1-microphone system are presented in Sec. 5. Simulation results are presented in Sec. 6 to show the performance of

the proposed 3-microphone system. Experimental results are given in Sec. 7. Final conclusions are drawn in Sec. 8.

## 2. Preliminaries on ITD

ITD-based sound localization consists of two independent steps, namely, ITD estimation and source localization based on ITD estimation. Since ITD estimation has been extremely explored in the literatures [Knapp & Carter, 1976; Gustafsson *et al.*, 2003; Tanaka & Kaneda, 1993; Champagne *et al.*, 1996], it is not the primary focus of this paper. For completeness of the paper, ITD estimation used in the paper is briefed in the section.

ITD is the difference in arrival times of the sound signal at any two microphones. ITD is thus proportional to the difference in distance from the sound source $S = [x_a, y_a, z_a]^T$ to any pair of omnidirectional and identical microphones, $m_i = [x_i, y_i, z_i]^T$ and $m_j = [x_j, y_j, z_j]^T$, given by

$$\delta_t(i,j) = \frac{|m_i - S| - |m_j - S|}{c_0} = \frac{r_i - r_j}{c_0}, \quad i,j \in [1,N], \tag{1}$$

where $|\cdot|$ is the Euclidean distance measure, $r_i$ and $r_j$ are the distances from the sound source to the microphones, $m_i$ and $m_j$, $c_0$ is the speed of sound in the air, assumed to be constant, $N$ is the number of microphones. Furthermore,

$$\begin{aligned} |\delta_t(i,j)| &\leq \frac{|m_i - m_j|}{c_0}, \\ \delta_t(i,j) &= -\delta_t(j,i). \end{aligned} \tag{2}$$

In an $N$-microphone system, after each sampling of the sound signal, one can obtain ${}^NC_2$ ITD values as follows:

$$\begin{bmatrix} \delta_t(1,2) \\ \vdots \\ \delta_t(1,N) \\ \vdots \\ \delta_t(i,i+1) \\ \vdots \\ \delta_t(i,N) \\ \vdots \\ \delta_t(N-1,N) \end{bmatrix} = \frac{1}{c_0} \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 0 & 0 & \dots & -1 \\ \vdots & \vdots & \vdots & \vdots & \dots & 0 \\ & \vdots & 1 & -1 & \dots & 0 \\ & \vdots & \vdots & \vdots & \dots & 0 \\ & \dots & 1 & 0 & \dots & -1 \\ & \vdots & \vdots & \vdots & \vdots & \vdots \\ & \dots & \vdots & \vdots & 1 & -1 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ r_N \end{bmatrix} = \frac{1}{c_0}\mathbf{AR}, \tag{3}$$

where the rank of $\mathbf{A}$ is $(N-1)$. It means that only $(N-1)$ ITD measurements are independent.

### 2.1. *ITD estimation*

Cross correlation techniques are typically used in the ITD estimation. The ITD is obtained as the time delay derived from the generalized cross correlation function between two received signals by any two microphones.

For any two microphones, $m_i$ and $m_j$, suppose that $m_j$ is the reference microphone. Then, the received signal $x^i(t)$ at $m_i$ may be modeled as

$$x^i(t) = x^j(t - \delta_t(i,j)) + v_i(t), \tag{4}$$

where $x^j(t)$ is the signal received at $m_j$, $v_i(t)$ are noise components at $m_i$, assumed to be uncorrelated with $x^j(t)$, $\delta_t(i,j)$ is the ITD value with respect to microphones $m_i$ and $m_j$. The generalized cross-correlation function between $x^j(t)$ and $x^i(t)$ is defined as

$$R_{i,j}(\tau) = F^{-1}\{\psi(f)\hat{G}_{i,j}(f)\}, \tag{5}$$

where $F^{-1}\{\cdot\}$ denotes the inverse Fourier transform of $\{\cdot\}$, $\hat{G}_{i,j}(f)$ refers to the estimate of the cross-spectral density function between $x^j(t)$ and $x^i(t)$, $\psi(f)$ denotes a certain weighting function used to minimize the spread of the cross-correlation function in time domain. In [Knapp & Carter, 1976], this $\psi(f)$ is given as

$$\psi(f) = \frac{1}{|G_{i,j}(f)|}. \tag{6}$$

Ideally, when $\hat{G}_{i,j}(f) = G_{i,j}(f)$, then $R_{i,j}(\tau) = \Delta(t - \delta_t(i,j))$ where $\Delta(\cdot)$ denotes the Dirac Delta function. In practice, $G_{i,j}(f)$ is unknown and $\psi(f)$ is replaced by $1/|\hat{G}_{i,j}(f)|$. Thus, $R_{i,j}(\tau)$ is not an ideal Dirac Delta function but has some spread centered around $t = \delta_t(i,j)$.

### 2.2. *Practical issue of ITD*

Clearly, Eq. (6) assumes that $G_{i,j}(f) \neq 0, \forall f$, which can only be satisfied by certain signals such as white noise.

Let the ITD values of an $N$ microphone system, at the time instant $n$, to be estimated be $s_n$, an $(N-1)$-tuple column vector containing independent ITD values. Since the elements in $s_n$ are independent with each other, the motion equation can be modeled as

$$s_{n+1} = A_n s_n + W_s, \tag{7}$$

where $A_n = I_{N-1}$ is an $(N-1)\times(N-1)$ identity matrix, and $W_s = [\omega_{s(1)}, \ldots, \omega_{s(N-1)}]^T$. As the motion of the sound source can be regarded as a random movement, $W_s$ is zero-mean normal distributed white noise and is characterized by variance matrix $Q_s$. It drives the actual ITD values from $s_n$ to $s_{n+1}$. It worths pointing out that the value of $W_s$ has close relationship to the motion of the sound source, e.g. large value of $W_s$ should be applied if the sound source moves quickly.

The measurement model can be written as

$$z_n = H_n s_n + W_z, \tag{8}$$

where $H_n = I_{N-1}$, $z_n$ is the ITD measurements, $W_z = [\omega_{z(1)}, \ldots, \omega_{z(N-1)}]^T$ is measurement disturbance which may result from misadjustment of microphones, signal corruption and cross-correlation algorithm error. Since audio drift is a long-term effect, it can be regarded as output bias which has no impact on ITD measurements. Other measurement disturbance can be modeled as white noise. Therefore, elements in $W_z$ are referred to as independent normal distributions with zero mean and are characterized by variance matrix $Q_z$. Clearly, $W_z$ is closely related to the level of measurement disturbance, e.g. $W_z$ is large if measurement noise is large. Due to the geometry constraints in Eq. (2), the ITD measurements can be modified as

$$\hat{\delta}_t^n(i,j) = \begin{cases} \Delta T_n(i,j), & \text{if } |\Delta T_n(i,j)| \le \left|\frac{m_i - m_j}{c_0}\right| \\ \hat{\delta}_t^{n-1}(i,j) & \text{otherwise} \end{cases}, \tag{9}$$

where $\Delta T_n(i,j)$ is the actual measurements. Since the ITD sampling interval is only 20–30 ms, it is reasonable to assume that the ITD values remain constant as shown in Eq. (9).

Given the motion and measurement models, the update equations are

$$\begin{aligned} \hat{s}_n^- &= A_n \hat{s}_{n-1}, \\ P_n^- &= A_n P_{n-1} A_n^T + Q_s, \end{aligned}$$

where $\hat{s}_n^-$ and $\hat{s}_{n-1}$ are *a priori* and *a posterior* estimation of $s_n$ and $s_{n-1}$ respectively, $P_n^-$ and $P_{n-1}$ are *a priori* and *a posterior* estimation of error covariance at time instants $n$ and $n-1$ respectively. The measurement update equations are

$$\begin{aligned} K_n &= P_n^- H_n^T (H_n P_n^- H_n^T + Q_z)^{-1}, \\ \hat{s}_n &= \hat{s}_n^- + K_n(z_n - H_n \hat{s}^-), \\ P_n &= (I - K_n H_n) P_n^-, \end{aligned} \tag{10}$$

where $K_n$ is the gain factor that minimizes the posterior error covariance. It can be observed from Eq. (10) that measurement $z_n$ will have weak impact on the update of $\hat{s}_n$ if $W_z$ is large, which will be shown in the simulation Sec. 6.

## 3. Three-Microphone System

Consider three non-collinear microphones, $m_1$, $m_2$ and $m_3$, in the system shown in Fig. 1. These three microphones form a Y-shaped frame whereby all the nodes are connected to a common point $O_a$. The distance of each microphone from $O_a$ is denoted by $r$ and they are equally spaced at 120° apart. As the distance between the microphones affects the estimation process, having this structure ensures that the time difference range between any two microphones are equal. This Y-shaped structure is fixed on a pan-tilt unit whose center $O$ is coincident with the common point $O_a$ and the node $m_3$

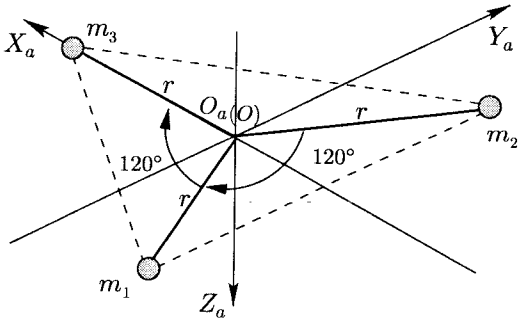



Fig. 1. Configuration of the 3-microphone system.

is on the $X_a$ axis. The spherical coordinates of the sound source, $[\alpha, \beta, d_0]^T$ is taken with respect to $O_a X_a Y_a Z_a$ where $d_0$ is the distance of the sound source from the origin, $O_a$, $\alpha$ and $\beta$ its azimuth and elevation angles. Where convenient, the cartesian coordinates of the sound source, denoted by $[x_a, y_a, z_a]^T$ may also be used.

Due to the Y-shaped structure, the 3D space is divided into 6 subspaces, I–VI, by planes $S_i$ $(i = 1, 2, \ldots, 6)$ as shown in Fig. 2 where $O_a m_3$ is the reference axis. Based on geometry, the $r$-vector in (3) can be computed from

$$\begin{aligned} r_i^2 &= d_0^2 + r^2 - 2rd_0 \cos\beta \cos(\phi_i - \alpha), \\ \phi_i &= \frac{2}{3}\pi(3 - i), \quad i = 1, 2, 3. \end{aligned} \tag{11}$$

Considering (11), the subspace or plane in which the sound source lies can be determined by the signs of $\delta_t(i, j)$ $(i, j = 1, 2, 3,\ i \neq j)$ according to (3) for $N = 3$. Hence if the sound source is between $m_i$ and the plane of symmetry between $m_i$ and $m_j$, then $r_k > r_j > r_i$, $k \neq j \neq i$, and hence $\delta_t(k, j) > 0$, $\delta_t(i, k) < 0$ and $\delta_t(i, j) < 0$. Table 1 shows the signs of $\delta_t(i, j)$ when the source is in each subspace I–VI or on each plane, $S_1 - S_6$. Using this table, the approximate position of the sound source can quickly be determined by examining the signs of $\delta_t(i, j)$.

According to (3), when $N = 3$, the rank of $\mathbf{A}$ is 2 and this implies that there are only 2 independent ITD values available to determine $[\alpha, \beta, d_0]^T$. Thus $[\alpha, \beta, d_0]^T$ cannot be solved uniquely based on one sampling of the sound signal. In this case, the locus of a given $\delta_t(i, j)$ value is a hyperboloid, while the locus of $\delta_t(i, k)$, $k \neq j$ is another hyperboloid, the intersection of the two hyperboloids forms a 3D curve $C^i_{j,k}$ which thus defines an infinite number of possible sound positions as illustrated in Fig. 3.

In the following subsections, we investigate the capability of the 3-microphone system to locate the source under some special conditions. A full 3D localization strategy will be proposed in Sec. 3.2.

### 3.1. *Capability of the 3-microphone system*

To investigate the capability of sound localization for the 3-microphone system, re-write (1) as

$$\begin{aligned} c_0 \delta_t(i, j) &= r_i - r_j \\ &= \frac{2rd_0 \cos\beta(\cos(\phi_j - \alpha) - \cos(\phi_i - \alpha))}{r_i + r_j}. \end{aligned} \tag{12}$$

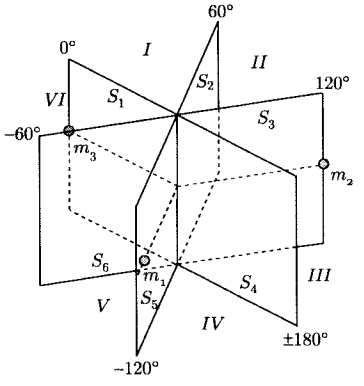



Fig. 2. Vectors determined by ITD values.

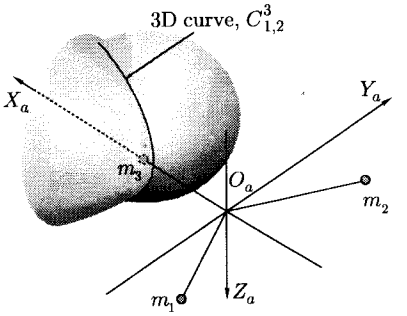



Fig. 3. 3D curve on which the sound source lies.

Table 1. Differential Time Distribution.

| | I | II | III | IV | V | VI | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\text{sgn}(\delta_t(3,1))$ | −1 | −1 | 1 | 1 | 1 | −1 | −1 | 1 | 0 | 1 | −1 | 0 |
| $\text{sgn}(\delta_t(3,2))$ | −1 | 1 | 1 | 1 | −1 | −1 | −1 | 0 | 1 | 1 | 0 | −1 |
| $\text{sgn}(\delta_t(1,2))$ | 1 | 1 | 1 | −1 | −1 | −1 | 0 | 1 | 1 | 0 | −1 | −1 |

We now consider 4 special cases where the sound source is (i) very far away ($d_0 \to \infty$), (ii) on the $O_aX_aZ_a$ plane when $\alpha = 0$, (iii) on the $O_aX_aY_a$ plane when $\beta = 0$, and (iv) has both $\alpha = \beta = 0$.

### *Case (i)*

If $d_0 \to \infty$, (12) becomes

$$c_0\delta_t(i,j) = r\cos\beta(\cos(\phi_j - \alpha) - \cos(\phi_i - \alpha)). \tag{13}$$

Given $\delta_t(i,j)$ and $\delta_t(i,k)$, we have

$$\begin{aligned} c_0\delta_t(i,j) &= r\cos\beta(\cos(\phi_j - \alpha) - \cos(\phi_i - \alpha)), \\ c_0\delta_t(i,k) &= r\cos\beta(\cos(\phi_k - \alpha) - \cos(\phi_i - \alpha)). \end{aligned} \tag{14}$$

Choosing $i = 3$, $j = 1$ and $k = 2$, and dividing the second equation by the first in (14), we have

$$\tan\alpha = \frac{\sqrt{3}(R_{\delta_t} - 1)}{R_{\delta_t} + 1}, \tag{15}$$

where

$$R_{\delta_t} = \frac{\delta_t(3,1)}{\delta_t(3,2)}.$$

Substituting (15) into (14) results in

$$\alpha = \arctan\left(\frac{\sqrt{3}(R_{\delta_t} - 1)}{R_{\delta_t} + 1}\right), \tag{16a}$$

$$\beta = \begin{cases} \arccos\left(\frac{c_0\delta_t(3,1)}{r\left(\cos\left(\frac{4}{3}\pi - \alpha\right) - \cos\alpha\right)}\right), & z_a \geq 0 \\ -\arccos\left(\frac{c_0\delta_t(3,1)}{r\left(\cos\left(\frac{4}{3}\pi - \alpha\right) - \cos\alpha\right)}\right), & z_a < 0 \end{cases}, \tag{16b}$$

$$\alpha \in [-\pi/2, \pi/2], \quad \beta \in [0, \pi/2] \quad \text{or} \quad [-\pi/2, 0].$$

Equations (16a) and (16b) show that a 3-microphone system can estimate the azimuth and elevation angles of a distant sound source if it is at the front side of the Y frame and positioned either above or below the horizontal plane.

### *Case (ii)*

Let us consider the case when $\alpha = 0$, $y_a = 0$, i.e. the source is at $[x_a, 0, z_a]^T$. Therefore, using (11), we have $r_1 = r_2$, i.e. there is only one independent ITD value. In other words, there are infinite solutions in this case.

### *Case (iii)*

When $\beta = 0$, $z_a = 0$ and the source is at $[x_a, y_a, 0]^T$ and lies on the same plane as the 3 microphones. Therefore

$$r_1^2 = \left(x_a + \frac{r}{2}\right)^2 + \left(y_a + \frac{\sqrt{3}r}{2}\right)^2, \tag{17a}$$

$$r_2^2 = \left(x_a + \frac{r}{2}\right)^2 + \left(y_a - \frac{\sqrt{3}r}{2}\right)^2, \tag{17b}$$

$$r_3^2 = (x_a - r)^2 + y_a^2. \tag{17c}$$

Equations (17a)–(17c) defines two hyperbolas on the $O_aX_aY_a$ plane, one each from $\delta_t(3,1)$ and $\delta_t(3,2)$. Thus, depending on the position of the source, either one or two solutions may be obtained as illustrated in Figs. 4 and 5.

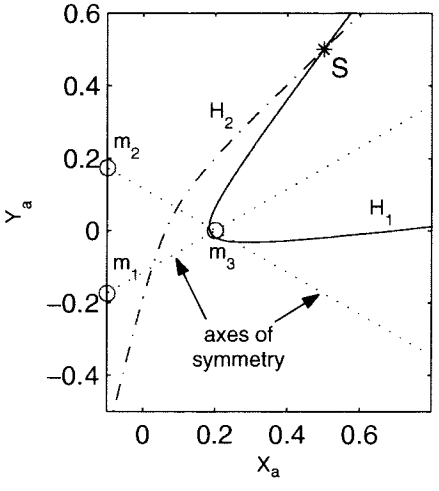



Fig. 4. Single solution for a special case in (iii).

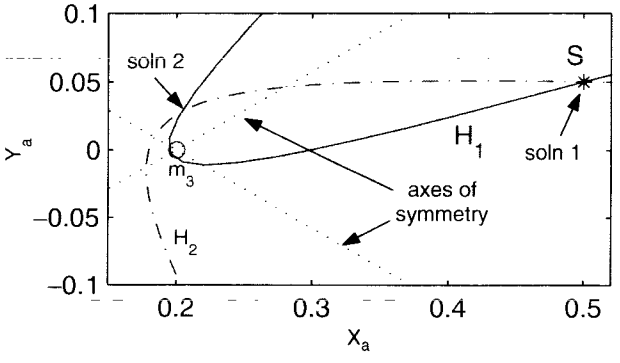



Fig. 5. Two solutions for a special case in (iii).

Figure 4 was simulated for a source position at $[0.5, 0.5, 0]^T$ with $r = 0.2$. The two hyperbolas are $H_1$ and $H_2$ with their axes of symmetry indicated. For this source location, there is only one solution derived from (17a)–(17c). Figure 5 was simulated for a source at $[0.5, 0.05, 0]^T$. In this case, there are clearly two solutions. Even for this special case where $\beta = 0$, a 3-microphone system cannot resolve the source position uniquely.

### *Case (iv)*

When $\alpha = \beta = 0$, the sound source not only shares the same plane as the 3 microphones but is also colinear with $O_a$ and $X_a$. Thus the position of the source is at $[x_a, 0, 0]^T$ with $x_a = d_0$. The solution of $x_a$ can be obtained by

$$x_a = \begin{cases} \dfrac{\delta_d(3,1)(\delta_d(3,1) - 2r)}{3r - 2\delta_d(3,1)} & x_a < r \\ \dfrac{\delta_d(3,1)(\delta_d(3,1) + 2r)}{3r + 2\delta_d(3,1)} & x_a \geq r \end{cases},$$

where $\delta_d(i,j) = c_0\delta_t(i,j)$. For the two solutions, one of them is inside the $Y$-frame containing the 3 microphones and hence is not a valid source location. Thus, in principle, a source that is on the $O_aX_aY_aZ_a$ plane and aligned with the $O_aX_a$ axis can be determined by a single sample of the sound signal. An illustration is given in Fig. 6 for a source at $[1, 0, 0]^T$.

The above detailed analysis shows that a three-microphone system, with a single sample of the signal (giving 2 independent ITD values), is only able to uniquely estimate the position of the sound source under very restrictive

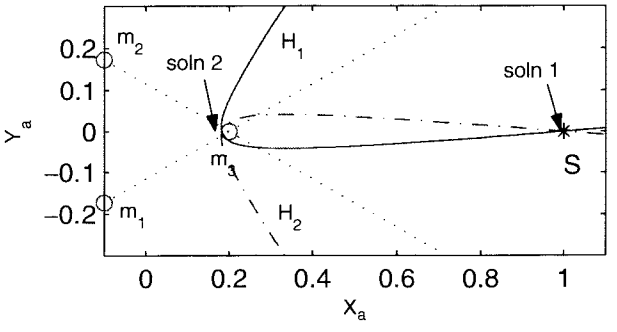



Fig. 6. Two solutions for case (iv).

conditions. In general, the full coordinates of the source in 3D space cannot be obtained uniquely. The main problem is that there are insufficient equations to determine 3 unknowns, $[\alpha, \beta, d_0]^T$ uniquely. In order to overcome this, an additional sample has to be taken at a different position of the 3 microphones. This can be achieved by rotating the pan-tilt unit through one or two axes. As an additional sample has to be taken for the same source position, we assume that the sound source is either fixed or moving slowly in space.

**Remark:** It is shown in [Brandstein & Silverman, 1997] that only a 20–30 ms sound signal is required to estimate ITD. Due to the efficiency of the ITD estimation, multiple-sampling should not yield significant errors when the sound source moves relatively much slower compared to the movement of the pan-tilt unit.

## 3.2. *3D localization by multiple-sampling*

To estimate the full 3D position of the source, another sample should be taken by rotating the pan-tilt unit. Basically, we aims to locate the sound source at the intersection of a set of 3D curves $C^3_{1,2}$, ${}^1C^3_{1,2}$, i.e. $S = {}^1C^3_{1,2} \cap C^3_{1,2}$. Since the coordinate system, $O_aX_aY_aZ_a$, is assumed to be "attached" to the Y-frame, when the unit moves, the position of the sound source has to be re-computed with respect to the new position of the Y-frame with a new coordinate system denoted by $O^i_aX^i_aY^i_aZ_a$ where the leading superscript indicates the $i$th rotation of the pan tilt unit. In this new coordinate system, the position of the sound source is denoted by ${}^ip = [{}^ix_a, {}^iy_a, {}^iz_a]^T$ or in spherical coordinates, $[{}^i\alpha, {}^i\beta, {}^id_0]^T$. Note that the origin $O_a$ remains the same after each rotation.

Figure 7 shows the spatial relationship between $O_aX_aY_aZ_a$, $O^i_aX^i_aY^i_aZ_a$ and the sound source, S. Suppose that the sound source is at ${}^0p = [{}^0x_a, {}^0y_a, {}^0z_a]^T$ or $[{}^0\alpha, {}^0\beta, {}^0d_0]^T$ with respect to the initial Y-frame ($O_aX_aY_aZ_a$) position. After the pan tilt unit rotates through ${}^1\delta_\alpha$ and ${}^1\delta_\beta$ in the azimuth and elevation directions respectively, the new relative angles, with respect to

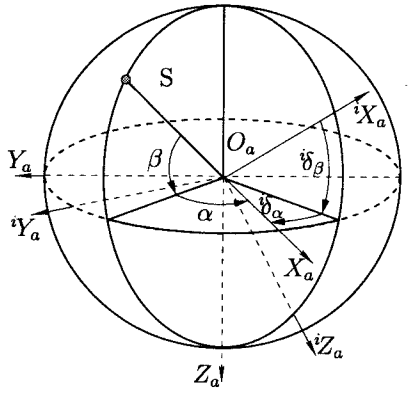



Fig. 7. Coordinate system.

the new coordinate system $O_a^1X_a^1Y_a^1Z_a$ can be computed as

$$
\begin{aligned}
{}^1\beta &= \arcsin\big(\cos{}^1\delta_\beta \sin{}^0\beta - \cos{}^0\beta \cos({}^0\alpha - {}^1\delta_\alpha)\sin{}^1\delta_\beta\big),\\
{}^1\alpha &= \begin{cases} \arcsin\left\{\frac{\cos{}^0\beta \sin({}^0\alpha - {}^1\delta_\alpha)}{\cos{}^1\beta}\right\} & \text{if } {}^1x_a \geq 0\\ \pi - \arcsin\left\{\frac{\cos{}^0\beta \sin({}^0\alpha - {}^1\delta_\alpha)}{\cos{}^1\beta}\right\} & \text{if } {}^1y_a \geq 0\\ -\pi - \arcsin\left\{\frac{\cos{}^0\beta \sin({}^0\alpha - {}^1\delta_\alpha)}{\cos{}^1\beta}\right\} & \text{if } {}^1y_a < 0 \end{cases}
\end{aligned}
\tag{18}
$$

and ${}^1r_i$, $i = 1, 2, 3$ can be obtained using (11). The new coordinates of the sound source with respect to the rotated frame are determined via transformations with rotation matrices as shown in Appendix 1. Note that since there is no translation of the origin, the distance of the sound source to $O_a$ remains unchanged and ${}^1d_0 = {}^0d_0$.

**Remark:** If the rotation is only in the azimuth direction, then

$$
{}^1\alpha = {}^0\alpha - {}^1\delta_\alpha; \quad {}^1\beta = {}^0\beta. \tag{19}
$$

By examining (11), we observe that if $\beta_0$ is one solution to these equations, $-\beta_0$ must be another solution. Therefore, the position of the sound source cannot be resolved uniquely if the pan-tilt is only rotated in the azimuth direction for the second sample. In terms of geometry, this problem can be explained by observing that the plane of symmetry corresponding to $O_aX_aY_a$ remains unchanged when the second sample is obtained from just a rotation about the $Z_a$ axis.

With the second sample where the rotation is in both the $\alpha$ and $\beta$ directions, there are now four nonlinear equations derived from ${}^0\delta_t(i,j)$ and ${}^1\delta_t(i,j)$, as follows

$$
\frac{(x_c\cos\theta + y_a\sin\theta + c_1)^2}{a_1^2} - \frac{(y_a\cos\theta - x_c\sin\theta)^2 + z_a^2}{b_1^2} = 1, \tag{20a}
$$

$$
\frac{(x_c\cos\theta - y_a\sin\theta + c_1)^2}{a_2^2} - \frac{(y_a\cos\theta + x_c\sin\theta)^2 + z_a^2}{b_2^2} = 1, \tag{20b}
$$

$$
\frac{(x_c'\cos\theta + {}^1y_a\sin\theta + c_1)^2}{a_1^{2'}} - \frac{({}^1y_a\cos\theta - x_c'\sin\theta)^2 + {}^1z_a^2}{b_1^{2'}} = 1, \tag{20c}
$$

$$
\frac{(x_c'\cos\theta - {}^1y_a\sin\theta + c_1)^2}{a_2^{2'}} - \frac{({}^1y_a\cos\theta + x_c'\sin\theta)^2 + {}^1z_a^2}{b_2^{2'}} = 1, \tag{20d}
$$

where $i \neq j \neq k$ and

$$
\begin{aligned}
a_1 &= \frac{c_0\,{}^0\delta_t(3,1)}{2}, & b_1 &= \sqrt{\frac{3r^2}{4} - a_1^2},\\
a_2 &= \frac{c_0\,{}^0\delta_t(3,2)}{2}, & b_2 &= \sqrt{\frac{3r^2}{4} - a_2^2}\\
a_1' &= \frac{c_0\,{}^1\delta_t(3,1)}{2}, & b_1' &= \sqrt{\frac{3r^2}{4} - a_1^{2'}},\\
a_2' &= \frac{c_0\,{}^1\delta_t(3,2)}{2}, & b_2' &= \sqrt{\frac{3r^2}{4} - a_2^{2'}}\\
c_1 &= \sqrt{3}r/2, & x_c &= x_a - r,\\
\theta &= \frac{\pi}{6}, & x_c' &= {}^1x_a - r.
\end{aligned}
$$

Solving the four equations yields the position of the sound source at $[{}^0\alpha, {}^0\beta, d_0]^T$. The nature of the solutions will be investigated next.

Figure 8 shows the 3D curves generated by the equations when two samples of the same sound source are taken. The sound source is at $[0.2, 0.2, 0.2]^T$ with $r = 0.1$. The second sample is obtained after the pan-tilt unit is rotated

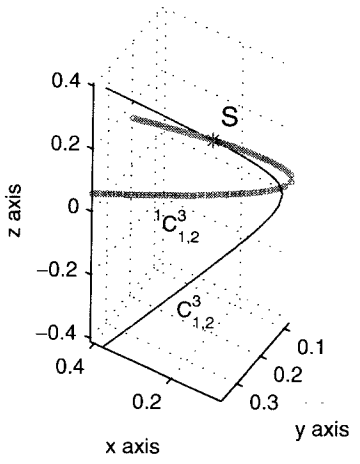



Fig. 8. Different 3D curves after rotation by ${}^1\delta_\alpha$ and ${}^1\delta_\beta$.

in both the elevation and azimuth directions by ${}^1\delta_\alpha = 30°$ and ${}^1\delta_\beta = 20°$ respectively. $C^3_{1,2}$ is the 3D curve when the Y-frame is in its original horizontal position with the $m_3$ microphone on the $X_a$-axis. ${}^1C^3_{1,2}$ is the 3D curve generated when the Y-frame is rotated. Both 3D curves are not on the same plane and they intersect only at the source position, S, thereby uniquely determining the source location.

Figure 9 shows the simulation for the same source position as above. But in this case, the pan tilt unit was rotated only in the azimuth direction by ${}^1\delta_\alpha = 30°$ for the second sample. As discussed previously, a single rotation in the azimuth direction cannot uniquely resolve the

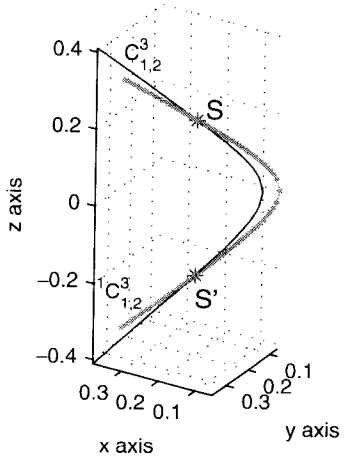



Fig. 9. Different 3D curves after rotation by only ${}^1\delta_\alpha$.

source position. As shown in Fig. 9, although both $C^3_{1,2}$ and ${}^1C^3_{1,2}$ do not exist on the same plane, they share the same plane of symmetry and intersects at two different locations which are mirror images of one another. Thus, there is another possible solution for the sound source at S′. This result shows that the second sample should be taken with rotation in both the azimuth and elevation directions in order to resolve the source location uniquely.

The multiple sampling method can be applied to 2-microphone systems, which will be investigated in the next section.

## 4. Two-Microphone System

It is well known that a 2-microphone system with a single sample cannot resolve a sound source uniquely in 3D space. This is because the locus of a constant ITD value are once again hyperbolas which are symmetric about the line joining microphones $m_1$ and $m_2$ as shown in Fig. 10.

Additional samples are required to obtain more equations, the intersection of which leads to the unique solution. In the 2-microphone system, since each sample gives one ITD value, a total of 4 samples ${}^i\delta_t(1,2)$, $i = 1,\ldots,4$ are required to set up the four equations similar to (20a)–(20d). New positions of the sound source with respect to the rotated frame $O^i_a X^i_a Y^i_a Z_a$ can

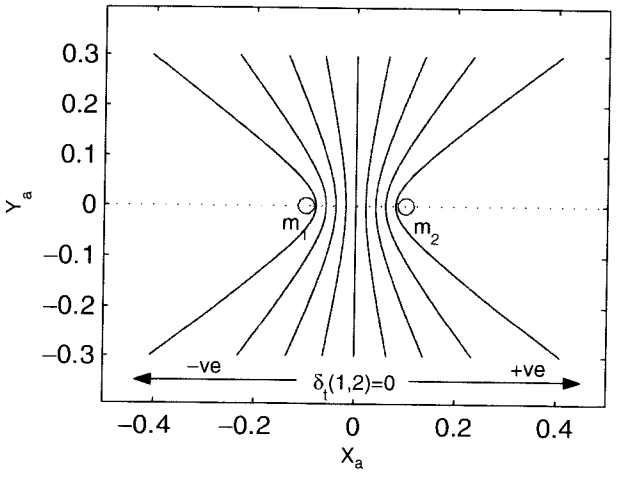



Fig. 10. Symmetric hyperbolas for 2-microphone system.

be re-computed as follows:

$$\begin{cases} {}^1p = {}^1R({}^1\delta_\alpha, {}^1\delta_\beta){}^0p & \text{(21a)} \\ {}^2p = {}^2R({}^2\delta\alpha, {}^2\delta\beta){}^1p\ , & \text{(21b)} \\ {}^3p = {}^3R({}^3\delta\alpha, {}^3\delta\beta){}^2p & \text{(21c)} \end{cases}$$

where ${}^iR({}^i\delta_\alpha, {}^i\delta_\beta)$ and ${}^ip$ are the rotation matrix and source position after the $i$th sequential rotation respectively. Recall that ${}^ip$ is taken with respect to the new coordinate frame $O_a^iX_a^iY_a^iZ_a$. As in the 3-microphone case, multiple sampling must be carried out with rotations in both the elevation and azimuth directions.

## 5. One-Microphone System

For a single microphone, only the spectral intensity or the interaural intensity difference (IID) at a point in space can be used as the localization cue. To calculate IID, the amplitude of the sound is assumed constant for the time it takes to move the pan-tilt unit during sampling.

Consider the 3-microphone system where only one microphone, $m_1$ is working. The sound source, S, is assumed to be located at $[{}^0x_a, {}^0y_a, {}^0z_a]^T$ with respect to the Y-frame. The sound intensity at $m_1$ can be computed from the sound sample via spectral density estimations. Since spectral intensities vary inversely as the square of the distance of the source, S from $m_1$, we have

$${}^iI_1 = \frac{k_s}{{}^id_1^2}$$

for the $i$th rotation, where ${}^id_1$ is the distance of S from $m_1$, ${}^iI_1$ denotes the spectral intensity at $m_1$, $k_s$ is a unknown constant. Since absolute intensities are not measurable, only relative intensities can be computed as follows:

$$\frac{{}^iI_1}{{}^{i+1}I_1} = \left(\frac{{}^{i+1}d_1}{{}^id_1}\right)^2 = f_i''({}^0x_a, {}^0y_a, {}^0z_a),$$

where $f_i''(.)$ is the equation of a sphere in 3D. In 2D, each $f_i''(.)$ is the equation of a circle. To locate the sound source on a 2D plane, 4 sound samples and 3 IID calculations are required to resolve the source position uniquely. An example is shown in Fig. 11 where the source is located at $[0.8, 0.4, 0]^T$ and 4 sound samples were obtained at positions $m_1$–$m_4$ yielding 3 IID calculations.

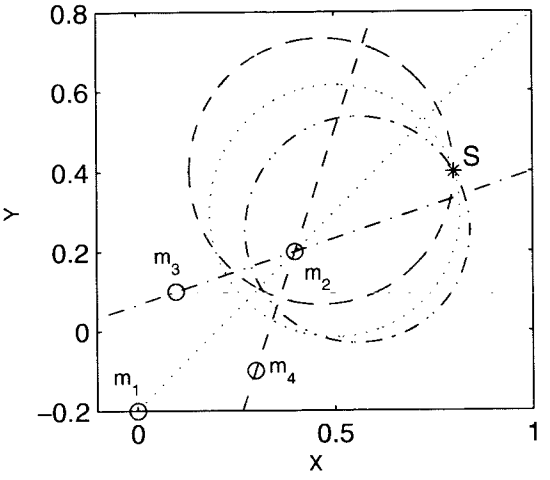



Fig. 11. Source location using a 1-microphone system.

Hence 3 circles can be observed which intersects at the source location, S. In general, 5 sound samples (4 IID values) must be captured by the single microphone to resolve the source location in 3D space.

## 6. Simulation Study

In this section, using 3-microphone system, two sets of simulations are presented to verify the angular estimation based on Kalman Filter and 3D sound localization based on feedforward neural network, respectively.

To estimate the angular value of a sound source, three different sound sources are used as shown in Fig. 12. In each simulation, we define by $x^1(t)$ the primary sound source while the other two, $x^2(t)$ and $x^3(t)$, are regarded as the secondary sources. All these signals arrive at $m_i$ and form

$$x_{m_i}(t) = \sum_{j=1}^{3} \rho_j x^j(t - \tau_{j,i}), \quad i = 1, 2, 3,$$

where

$$\tau_{j,i} = f_s \frac{|S_j - m_i|}{c_0}, \quad \rho_1 = 1, \quad \rho_2 = \rho_3,$$

$$\rho_2 = \sqrt{\frac{E_1}{(E_2 + E_3) * 10^{SNR/10}}},$$

$$E_i = \int_0^t (x^i(u))^2\, du,$$

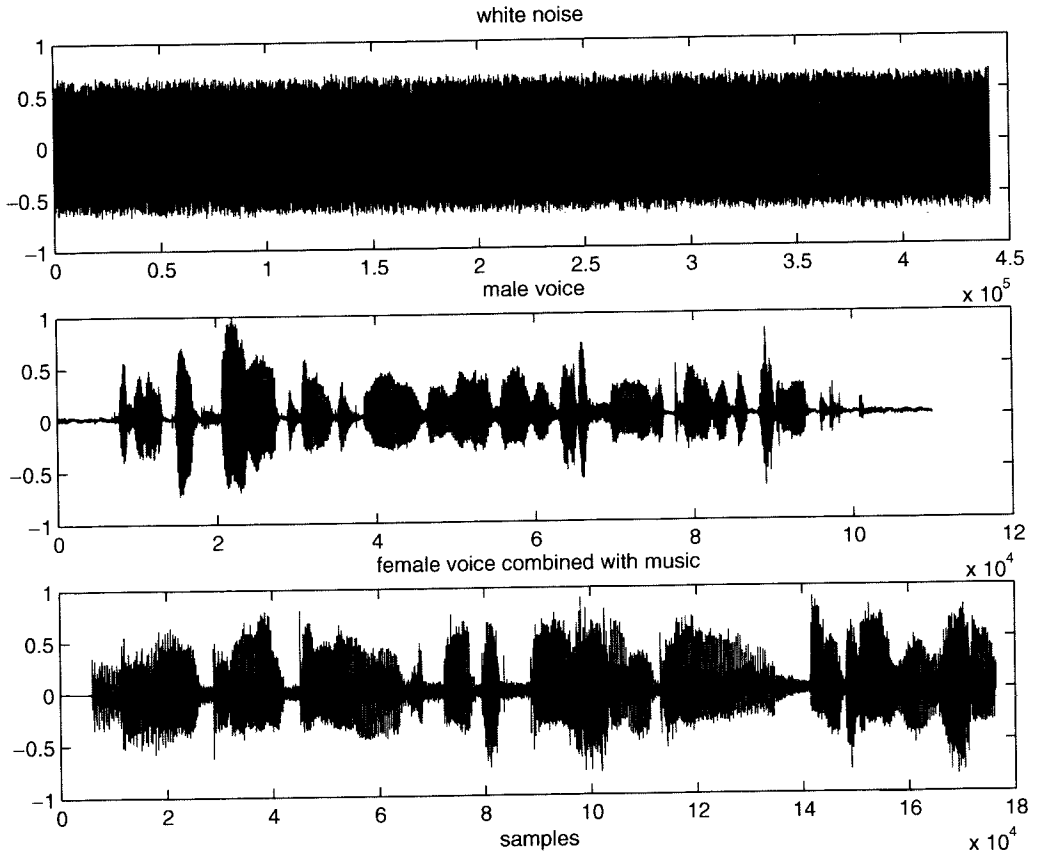



Fig. 12. Sound sources.

Table 2. Simulation Cases.

| | Primary Source $x^1(t)$ | Secondary Source $x^2(t)$ | Secondary Source $x^3(t)$ |
|---|---|---|---|
| Case I | White noise | Male human voice | Female human voice |
| Case II | Male human voice | Female human voice | White noise |
| Case III | Female human voice | White noise | Male human voice |

$f_s$ is the sampling rate; $S_j$ is the $j$th sound source; and $SNR$ is signal-noise-ratio. The primary sound source is at $[2\,\mathrm{m}, -2\,\mathrm{m}, 2\,\mathrm{m}]^T$ or $[-45°, 37.7281°, 3.4641\,\mathrm{m}]^T$ in spherical coordinates system, while the secondary sources are at $[1\,\mathrm{m}, 0\,\mathrm{m}, 0\,\mathrm{m}]^T$ and $[1\,\mathrm{m}, 1\,\mathrm{m}, 0\,\mathrm{m}]^T$ respectively. Three cases are simulated as shown in Table 2. Since the white noise only has strong correlation at certain sample delay, best ITD estimations can be achieved, so are the angular estimations. However, due to the distance constraint, the angular estimations are $[-44.3135°, 34.1298°]^T$ as shown in the top graph of Fig. 13. The graph at the bottom shows that angular estimation results if Kalman Filter is applied. It is shown that angular values converge to $[-44.3135°, 34.1298°]^T$. If human voices are the primary source as cases II and III, the ITD estimations will deviate from the actual ITD values as the analysis in Sec. 2.2. Angular estimation results are shown at the top of Figs. 14 and 15 respectively. It is observed that angular values jump randomly. However, if Kalman Filter is applied, the angular estimation converge to acceptable level at $[-50.5904°, 37.3268°]^T$ and $[-44.7559°, 34.9876°]^T$ respectively. In these

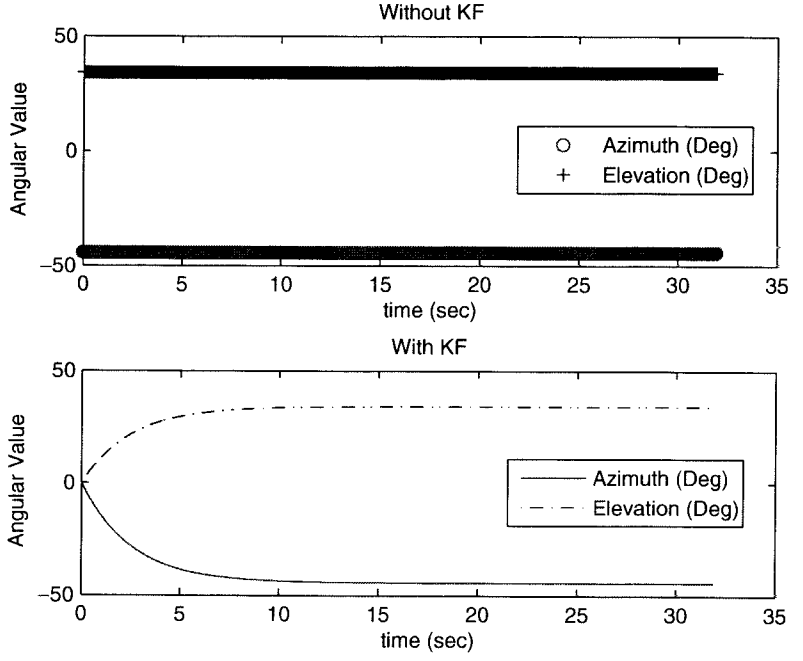



Fig. 13. Case I. White noise is the primary source.

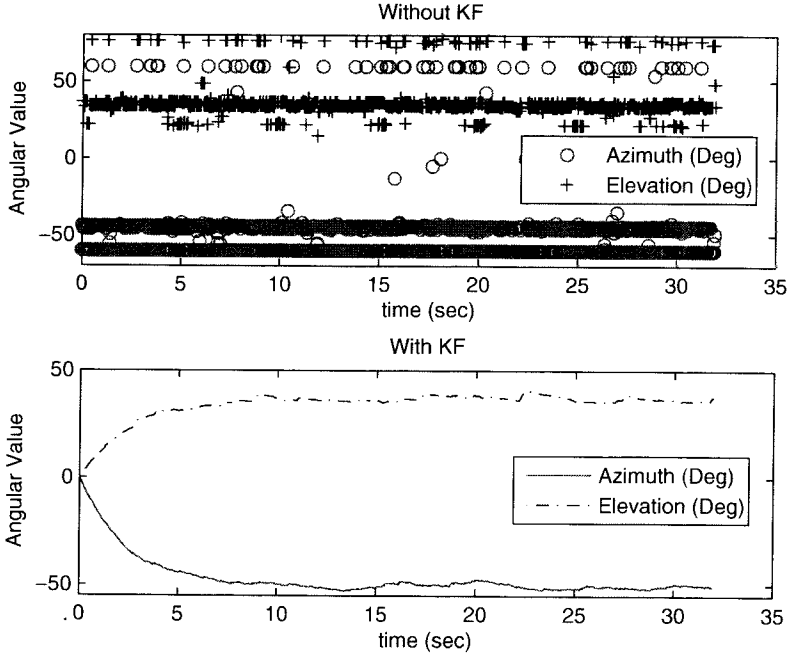



Fig. 14. Case II. Male human voice is the primary source.

simulations, the initial values of angular estimations are 0 while the error covariance matrices are initialized as $I_{2\times 2}$.

The second set of simulation is used to verify the 3D sound localization using multiple sampling and feedforward neural network. The key problem of such locators addressed in earlier sections is the computational effort in finding the solution to simultaneous equations that are set by the geometry of the system. Taylor-series

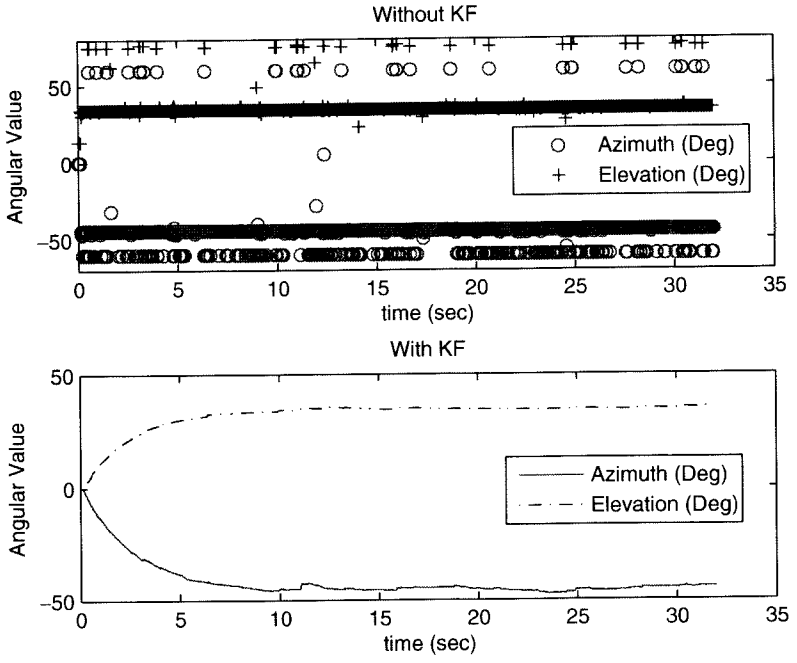



Fig. 15. Case III. Female human voice is the primary source.

expansion is ever used to linearize the equations, the solution after which is derived from optimization. However, this is iterative and depends very much on the initial guesses which should be closed to the solution. In [Chan & Ho, 1994], the nonlinear equations are transformed to a set of linear equations by simple substitution. Though a solution can be obtained by applying the least squares method, the computations involve the inverse of a matrix which may be ill-conditioned.

Since the solutions to the localization problem involve simultaneous nonlinear functions, a neural networks approach is proposed in this section. The neural networks (NN) approach is preferred because of its potential to give accurate instantaneous results after some initial off-line training [Ge *et al.*, 1998, 2001]. Moreover, it is robust when measurements are noisy under the condition that there are not redundant ITD measurements. Furthermore, in this problem where the solutions involve finding $\alpha$ and $\beta$ which are bounded, an NN approach may be appropriate. The final test to our proposal is in the experimental results that is presented in Sec. 7. As a simple verification of the NN-based multiple sampling method, 3D localization simulation for a 3-microphone system is presented in this section.

Since there are different possibilities to obtain the second sample, three scenarios are tested. Each scenario corresponds to a second sampling which involves (i) a single rotation in ${}^1\delta_\alpha$, (ii) a single rotation in ${}^1\delta_\beta$ and, (iii) rotation in both ${}^1\delta_\alpha$ and ${}^1\delta_\beta$. In each case (where $i = 3$, $j = 1$ and $k = 2$), 1000 random sets of $[\alpha, \beta, d_0]^T$ and their corresponding $({}^0\delta_t(3,1), {}^0\delta_t(3,2))$ (first sample) and $({}^1\delta_t(3,1), {}^1\delta_t(3,2))$ (second sample), were used to train the network. Finally an additional 100 sets of $[\alpha, \beta, d_0]^T$ were used to test the network after training.

The NN has 3 layers, each having 30 neurons. It has 4 inputs corresponding to ${}^0\delta_t(3,1)$, ${}^0\delta_t(3,2)$, ${}^1\delta_t(3,1)$ and ${}^1\delta_t(3,2)$. The outputs of the NN are the coordinates $[\alpha, \beta, d_0]^T$ of the sound source. Each node in the NN has an activation function of the form

$$g(x) = \frac{2}{1 + e^{-2x}} - 1$$

except the output node which is linear. The weights are adjusted using backpropagation.

### *Scenario 1*

In this case, the Y-frame is rotated by ${}^1\delta_\alpha$ in the azimuth direction to obtain the second sample. Figure 16 shows the estimation result after training. The solid, dashed and dash-dot lines denote the given $\alpha$ (rad), $\beta$ (rad) and $d_0$ (m), respectively; while the other three lines labeled star (*), circle (o) and plus (+) denote the estimated $\alpha$, $\beta$ and $d_0$ obtained from the trained neural network respectively. Figure 16 shows that the neural network cannot generate a good mapping under the condition that only a rotation in the azimuth direction is employed in the second sample. This problem was alluded to earlier in Sec. 3.2 where it was shown how the solution to the nonlinear equations is not unique due to the symmetry about the horizontal plane.

### *Scenario 2*

In this case, the Y-frame is rotated by ${}^1\delta_\beta$ in the elevation direction. Figure 17 shows the test results after the training.

This figure shows that the neural network is able to satisfactorily map values of ${}^0\delta_t(3,1)$, ${}^0\delta_t(3,2)$, ${}^1\delta_t(3,1)$ and ${}^1\delta_t(3,2)$ to $[\alpha, \beta, d_0]^T$. This indicates that the second sample in the elevation direction is reliable for sound localization.

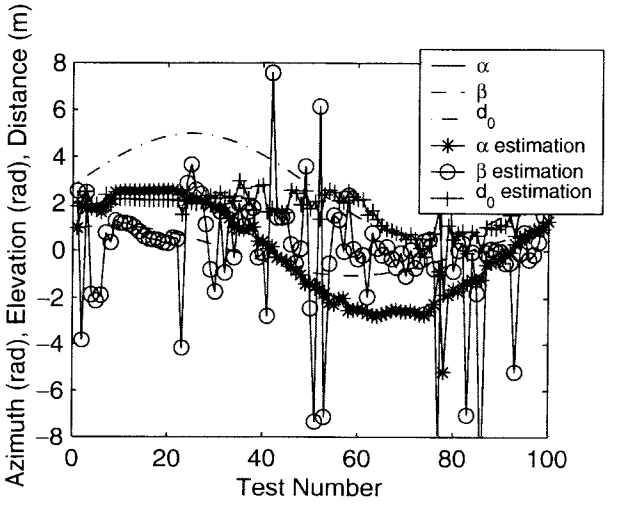



Fig. 16. Turn in azimuth.

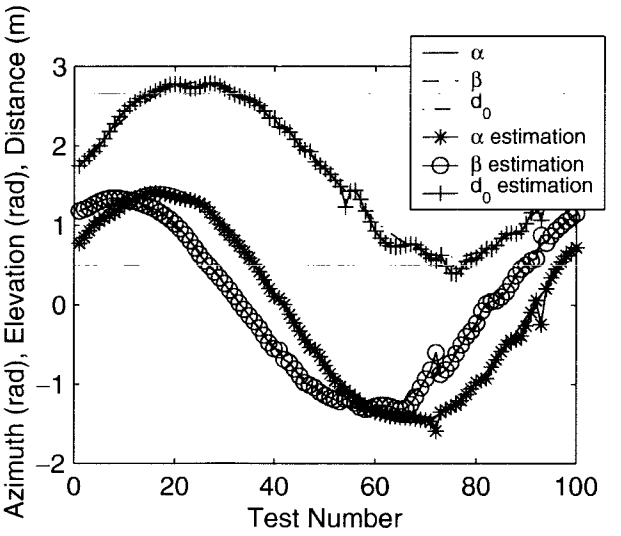



Fig. 17. Turn in elevation.

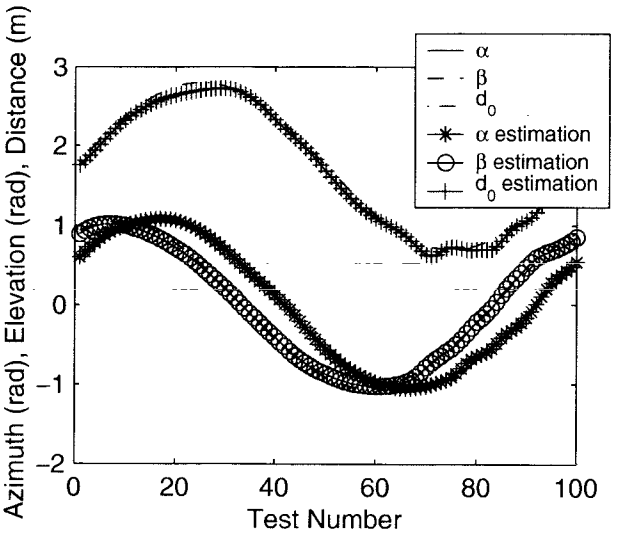



Fig. 18. Turn in both azimuth and elevation.

### *Scenario 3*

In this case, the Y-frame is rotated by both ${}^1\delta_\alpha$ and ${}^1\delta_\beta$ to obtain the second sample. Figure 18 shows the test results.

This figure shows that the neural network can estimate $\alpha$, $\beta$ and $d_0$ better than in scenario 2. Hence, the second sample taken by moving the pan-tilt unit in both the azimuth and elevation directions is more effective in the sound localization problem.

## 7. Experiments

Two main sets of experiments are developed to illustrate the performance of the sound localization algorithms using a Y-shaped structure with dimension $r = 13\,\text{cm}$. The first set illustrates how real time localization is possible with certain assumptions about the source position as illustrated by the different cases in Sec. 3.1. These assumptions are: (A1) the source is relatively far away compared to $r$, the spacing between the microphones, and (A2) the source locates in a certain half space. With these assumptions, only one sampling is required and (16a) and (16b) can be used to compute $\alpha$ and $\beta$.

The second set of experiment solves the full 3D problem without these assumptions. This involves multiple sampling followed by the construction of a neural networks to provide the mapping between the ITD space and the source coordinates. In this experiment, real ITD data was gathered and a NN was trained to provide the source coordinates corresponding to the ITD values.

Firstly, the experimental environment is described in the following subsections.

### 7.1. *Experimental environment*

All the experiments were carried out in the Mechatronics and Automation lab at the National University of Singapore, as shown in Fig. 19. The test space is a small area in the lab, which is surrounded by a number of working personal computers and other devices. For the sake of clarification, the term "primary source" stands for the speaker while "noise" or "secondary source" stands for sound signals generated by other devices such as weak background sound from nearby PCs, printers, conversations and other sounds from fellow lab members, and other sounds generated by the construction outside the lab.

The speaker and signal generator are placed on a holder so that they can be easily moved in the lab. A string suspends the speaker via the extended beam of the holder. The $X_aY_a$ plane is parallel to the floor which is assumed to be

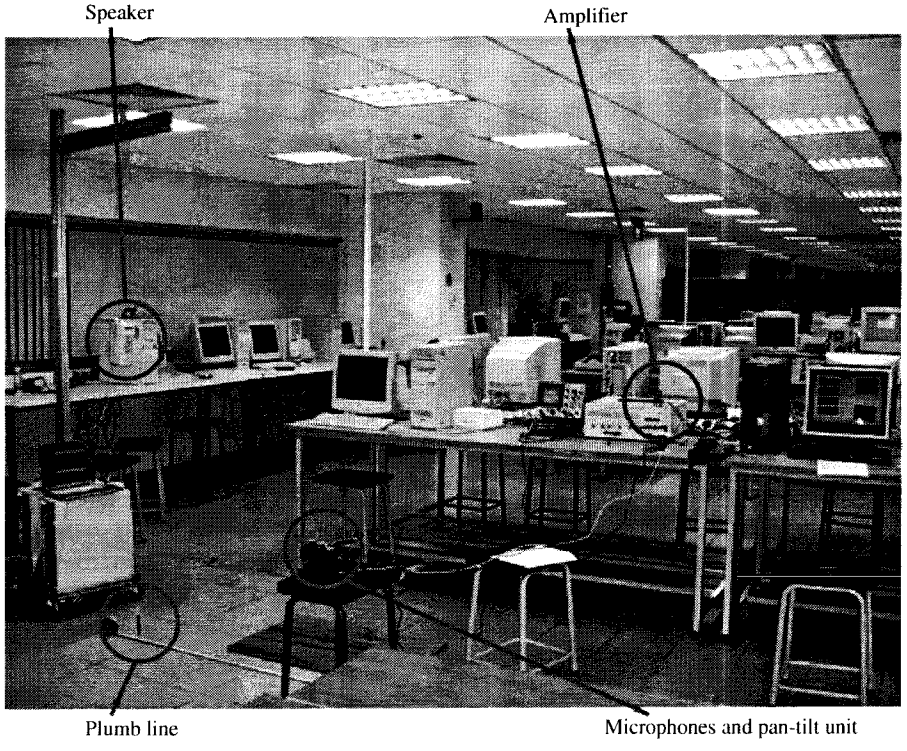



Fig. 19. Experimental environment.

horizontal. A plumb line is placed under the speaker to pinpoint a marker on the floor. The $z$ coordinate can be fixed by adjusting the string. The background noise level is 45.1 dB before the experiment. It was found that the source is perceivable as long as the sound level arriving at the center of the Y frame is larger than 52.5 dB, which is referred to as the perceivable level. During the experiment, the sound level generated by the speaker is constant and higher than the perceivable level. Since white noise has a wide spectrum with a constant sound level without pause, it is selected as the source signal.

### 7.2. *Experimental results*

The first set of results shows how real time azimuth and elevation estimation can be achieved under assumptions (A1) and (A2) using 3 microphones without multiple sampling. With these assumptions, the azimuth and elevation angles of the source can be computed from (16a) and (16b). The results are shown in Fig. 20. The top half shows the azimuth estimation while the bottom half indicates the elevation estimation. The numbers at the top left hand corner of each rectangular area indicates the estimated angular position in degrees. The diamond markers indicate the positions with respect to the middle line which has been calibrated at 0° azimuth. When the pan tilt unit is moved, these markers move in real time to indicate the source position with respect to

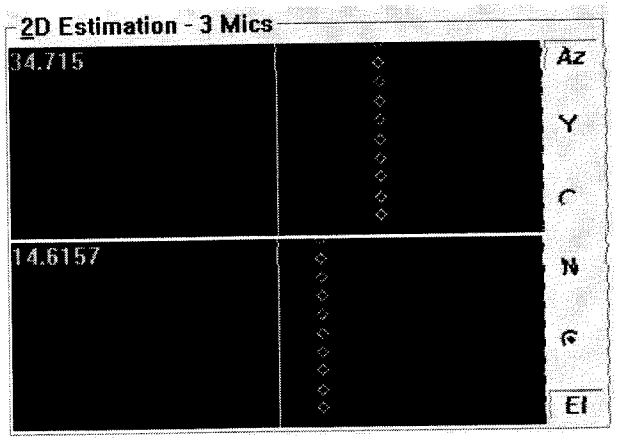


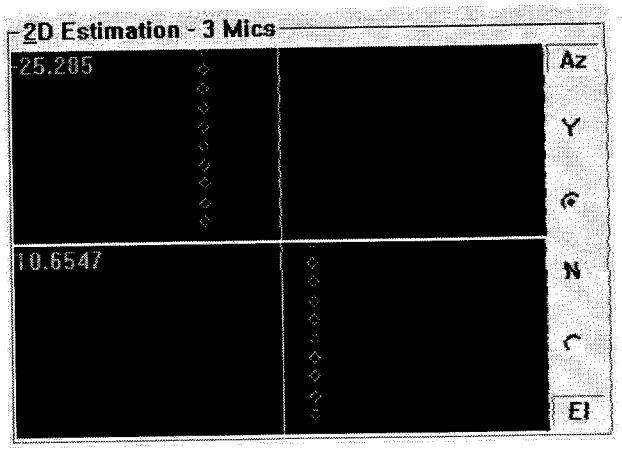

Fig. 20. Azimuth and elevation estimations derived from 3 microphones.

the coordinate system that is attached to the Y-frame. For realistic implementation, different sound sources are used, namely, white noise, male and female voice, as shown in Fig. 12. The speaker is positioned at $[1700\,\text{mm}, -840\,\text{mm}, 465\,\text{mm}]^T$ or $[-26.2948°, 14.1952°, 1950\,\text{mm}]^T$ in spherical coordinates system. Angular estimation of these sound sources is shown in Fig. 21 in which (a)–(c) show the results without the application of Kalman Filter. Due to the inherent property of the ITD method adopted, the angular estimation will deviate from the actual angular position of the primary sound source. Figures 21(d)–(f) show the results with the application of Kalman Filter which indicates that the angular estimation become continuous and reliable. However, due to the ITD measurement error, there exists the deviation of final estimation which depends strongly on the measurement noise in the experimental environments and the spectrum property of the primary sound source.

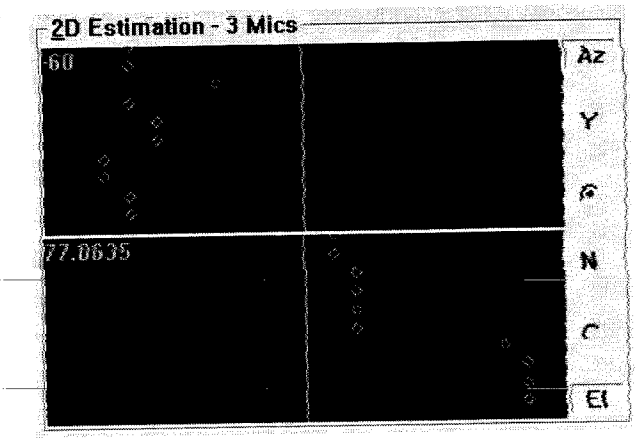
The second set of experiment involves the full 3D localization with multiple sampling. Since the nonlinear equations cannot be solved in real time, a neural network is trained to provide the mapping between ITD values and source coordinates. The experimental setup is used to gather real training data. A total of 92 training samples are taken in the training phase. These samples are distributed randomly and uniformly in the region $\alpha \in [-60°, 60°]$, $\beta \in [10°, 50°]$ and $d \in [0.5, 3]$ m corresponding to one of the subspaces in Fig. 2.

Five samples are taken for each pan tilt position. The average ITD values from these samples are used as inputs into the NN. Since the number of training samples is small, the NN has only two hidden layers, each having 13 neurons. Figure 22 shows the outputs of the trained NN using only the training samples. As can be observed, the NN can estimate trained positions accurately.

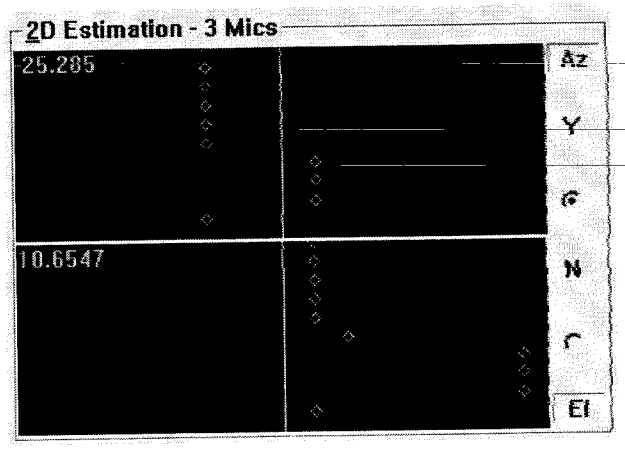
In the test phase, 11 different source positions in space were selected for testing. At each location, 2 samples of sound source were taken by the 3-microphone system and 4 ITD values calculated. The ITD values from these samples were then fed into the trained NN. Figures 23 show the results of the NN mapping from the



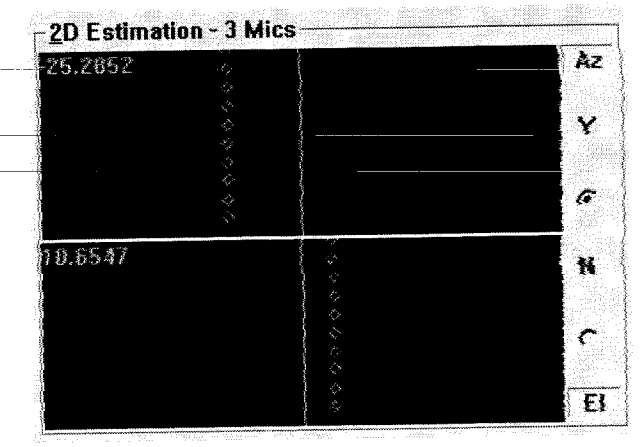

(a) White Noise without KF



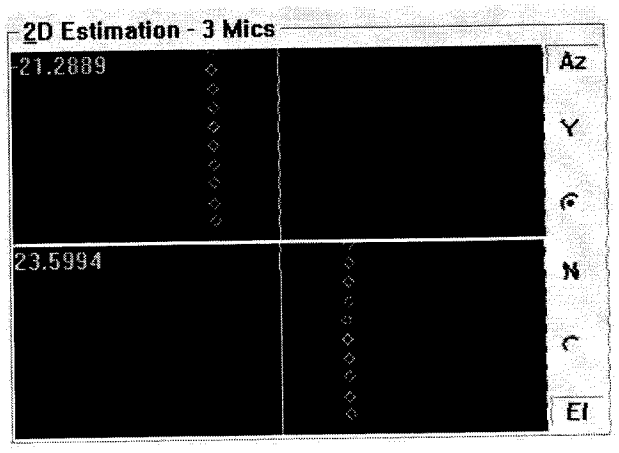

(b) Male Voice without KF



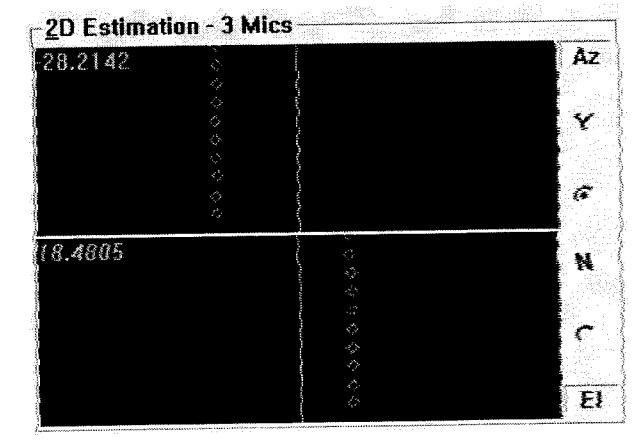

(c) Female Voice without KF




(d) White Noise with KF




(e) Male Voice with KF




(f) Female Voice with KF

Fig. 21. Azimuth and elevation tracking without and with Kalman Filter.

ITD values to the source locations. It can be observed that the results are reasonably accurate despite the fact that the signal samples are noisy under the experimental conditions. The effect of noise can be seen in Fig. 24 where the same source positions gave slightly different NN outputs due to differences in ITD values because of noise.

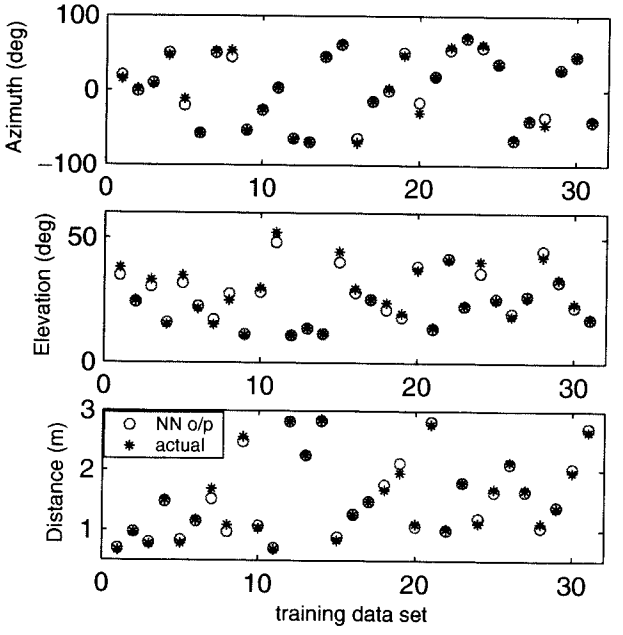



Fig. 22. NN outputs tested with training samples.

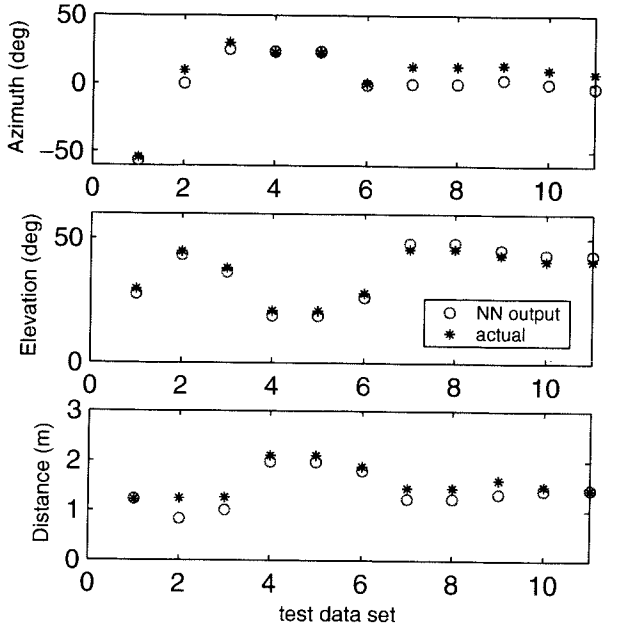



Fig. 23. ITD to source coordinate mappings after NN training.

For better accuracy, we suggest averaging the NN outputs of each source coordinate. This is to mitigate the effects of noise from the environment. An averaged result is shown in Fig. 25 where each source location is average over 10 sets of input-output data.

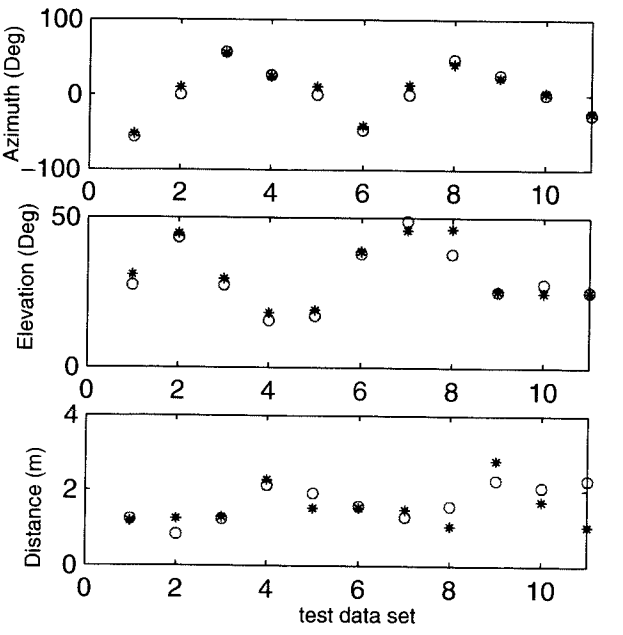



Fig. 24. ITD to source coordinate mappings after NN training.

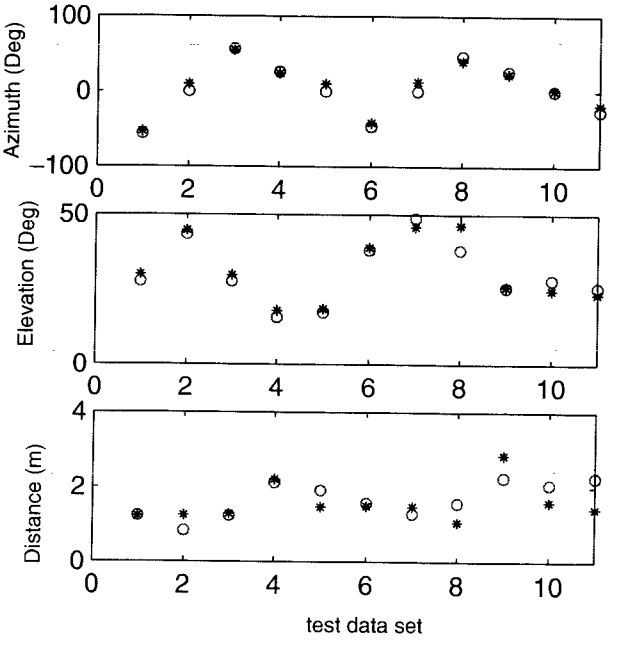



Fig. 25. Averaged source coordinates.

The positions measured by tape and corresponding average estimation by NN are compared in Table 3. Each row corresponds to one test position and consists of (i) position measured by tape, (ii) position estimated by NN and (iii) estimation error. The result that

Table 3. Experimental Results for Case III.

| Position measured by tape | | | Position estimation by NN | | | Estimation error | | |
|---|---|---|---|---|---|---|---|---|
| $\alpha(^\circ)$ | $\beta(^\circ)$ | $d$(mm) | $\hat{\alpha}(^\circ)$ | $\hat{\beta}(^\circ)$ | $\hat{d}$(mm) | $\alpha-\hat{\alpha}$ | $\beta-\hat{\beta}$ | $(d-\hat{d})/d\times 100\%$ |
| −56.3826 | 27.5472 | 1232.5 | −53.5528 | 29.8325 | 1224.7 | −2.8299 | −2.2853 | 0.63 |
| 0 | 43.2938 | 831.219 | 9.8216 | 44.7187 | 1243.7 | −9.8216 | −1.4249 | −49.63 |
| 56.3826 | 27.5472 | 1232.5 | 55.0631 | 29.6046 | 1286.6 | 1.3196 | −2.0574 | −4.39 |
| 26.4397 | 15.5836 | 2121.8 | 24.0203 | 17.8188 | 2214.5 | 2.4194 | −2.2352 | −4.37 |
| 0 | 17.3005 | 1916.7 | 10.2236 | 18.3987 | 1474.3 | −10.2236 | −1.0982 | 23.08 |
| −46.9525 | 37.9176 | 1578.4 | −41.7338 | 38.8681 | 1495.7 | −5.2187 | −0.9505 | 5.25 |
| 0 | 48.7723 | 1289.7 | 13.1307 | 45.781 | 1485.6 | −13.1307 | 2.9913 | −15.18 |
| 46.9525 | 37.9176 | 1578.4 | 40.6235 | 46.3213 | 1058.9 | 6.329 | −8.4037 | 32.91 |
| 26.4397 | 25.3896 | 2262.3 | 23.1425 | 25.7175 | 2855.3 | 3.2972 | −0.3279 | −26.21 |
| 0 | 27.926 | 2071.2 | 2.1904 | 24.556 | 1606.1 | −2.1904 | 3.3701 | 22.46 |
| −26.4397 | 25.3896 | 2262.3 | −24.6061 | 23.2876 | 1422.9 | −1.8336 | 2.1020 | 37.1 |

emerges from the table is that the estimation by NN is comparable to that in work [Weng & Guentchev, 2001] wherein the average angular error is ±3° and the average radial distance error is ±20%.

## 8. Conclusion

In this paper, robust sound localization has been proposed for fewer (3, 2 or 1) spatially distributed microphones than 4, the minimum required criteria for 3D sound localization in the literature, in an effort to address the robustness issues of microphone failures. It has been shown that, for three- and two-microphone systems, two and four samples of the Interaural Time Difference (ITD) measurements are required respectively. In a one-microphone system, five samples of the Interaural Intensity Difference (IID) measurement should be used instead. These investigations are important because if one microphone fails, the remaining system can still function to locate the sound source without significant impact on the localization system. In addition, our investigation has shown that 3-microphone system can estimate the azimuth and elevation simultaneously, which usually requires 4 microphones in the literature. Simulation and experimental results was presented to illustrate the performance of a three-microphone system. Results showed that the system could locate the sound source with satisfactory accuracy.

## Appendix 1. Solution to ${}^1\alpha$ and ${}^1\beta$

In this appendix, the azimuth and elevation angles of the sound source with respect to a rotated axis are derived. They are written in terms of their corresponding angles with respect to the original coordinate system and the angles of rotation of the pan tilt unit.

Recall that the original source position is ${}^0p = [{}^0x_a, {}^0y_a, {}^0z_a]^T$ in cartesian coordinates or $[{}^0\alpha, {}^0\beta, d_0]^T$ in spherical coordinates with respect to $O_aX_aY_aZ_a$. Subsequently after two rotations through ${}^1\delta_\alpha$ about the $Z_0$ axis and ${}^1\delta_\beta$ about the $Y_a$ axis, the new position of the sound source denoted, ${}^1p = [{}^1x_a, {}^1y_a, {}^1z_a]^T$ or $[{}^1\alpha, {}^1\beta, {}^1d_0]^T$ is obtained as

$$\begin{aligned} {}^1p &= R_{Y_a}({}^1\delta_\beta)R_{Z_a}({}^1\delta_\alpha){}^0p \\ &= {}^1R({}^1\delta_\alpha, {}^1\delta_\beta){}^0p, \end{aligned} \tag{22}$$

where $R_{Z_a}({}^1\delta_\alpha)(R_{Y_a}({}^1\delta_\beta))$ is the rotation matrix with respect to $Z_a(Y_a)$ by ${}^1\delta_\alpha({}^1\delta_\beta)$. These rotations leave $d_0$ unchanged because there is no translation of the origin. The relationship between ${}^0\alpha$, ${}^0\beta$, ${}^1\alpha$ and ${}^1\beta$ are derived as follows. Equation (22) gives

$$\begin{bmatrix} {}^1x_a \\ {}^1y_a \\ {}^1z_a \end{bmatrix} = \begin{bmatrix} -\cos{}^1\delta_\beta \cos{}^1\delta_\alpha & -\cos{}^1\delta_\beta \sin{}^1\delta_\alpha & \sin{}^1\delta_\beta \\ -\sin{}^1\delta_\alpha & \cos{}^1\delta_\alpha & 0 \\ -\sin{}^1\delta_\beta \cos{}^1\delta_\alpha & -\sin{}^1\delta_\beta \sin{}^1\delta_\alpha & \cos{}^1\delta_\beta \end{bmatrix} \times \begin{bmatrix} {}^0x_a \\ {}^0y_a \\ {}^0z_a \end{bmatrix}. \tag{23}$$

Since

$$
\begin{aligned}
&{}^0x_a = d_0 \cos{}^0\beta \cos{}^0\alpha;\ {}^0y_a = d_0 \cos{}^0\beta \sin{}^0\alpha;\\
&{}^0z_a = d_0 \sin{}^0\beta,
\end{aligned} \tag{24}
$$

and the origin is unchanged, we have

$$
\begin{aligned}
\sin{}^1\alpha &= \frac{{}^1y_a}{\sqrt{({}^1x_a)^2 + ({}^1y_a)^2}};\\
\sin{}^1\beta &= \frac{\sqrt{d_0^2 - ({}^1x_a)^2 - ({}^1y_a)^2}}{d_0}.
\end{aligned} \tag{25}
$$

Since ${}^1\delta_\alpha = 0$ and ${}^1\delta_\beta = 0$, ${}^1\alpha = {}^0\alpha$, ${}^1\beta = {}^0\beta$, we have the solution from (23)

$$
\begin{aligned}
&\sin{}^1\beta = \cos{}^1\delta_\beta \sin{}^0\beta - \cos{}^0\beta \cos({}^0\alpha - {}^1\delta_\alpha) \sin{}^1\delta_\beta,\\
&\sin{}^1\alpha = \frac{\cos{}^0\beta \sin({}^0\alpha - {}^1\delta_\alpha)}{\cos{}^1\beta},\\
&\qquad {}^1\alpha \in [-\pi, \pi] \quad \text{and} \quad {}^1\beta \in [-\pi/2, \pi/2].
\end{aligned} \tag{26}
$$

Therefore,

$$
\begin{aligned}
{}^1\beta &= \arcsin(\cos{}^1\delta_\beta \sin{}^0\beta\\
&\quad - \cos{}^0\beta \cos({}^0\alpha - {}^1\delta_\alpha) \sin{}^1\delta_\beta),\\
{}^1\alpha &= \begin{cases}
\arcsin\left\{\frac{\cos{}^0\beta \sin({}^0\alpha - {}^1\delta_\alpha)}{\cos{}^1\beta}\right\}, & \text{if } {}^1x_a \geq 0\\
\pi - \arcsin\left\{\frac{\cos{}^0\beta \sin({}^0\alpha - {}^1\delta_\alpha)}{\cos{}^1\beta}\right\}, & \text{if } {}^1y_a \geq 0\\
-\pi - \arcsin\left\{\frac{\cos{}^0\beta \sin({}^0\alpha - {}^1\delta_\alpha)}{\cos{}^1\beta}\right\}, & \text{if } {}^1y_a < 0
\end{cases}.
\end{aligned} \tag{27}
$$

## References


Aarabi, P. [2002] "Self-localization dynamic microphone arrays," *IEEE Transactions on Systems, Man and Cybernatics — Part C: Applications and Reviews* **32**(4), 474–484.

Brandstein, M. S. [1995] "A framework for speech source localization using sensor arrays," PhD dissertation, Brown University.

Brandstein, M. S., Adcock, J. E. and Silverman, H. F. [1997] "A closed-from location estimator for use with room environment microphone arrays," *IEEE Transactions on Speech and Audio Processing* **5**(1), 45–50.

Brandstein, M. S. and Silverman, H. F. [1997] "A practical methodology for speech source localization with microphone arrays," *Computer Speech and Language* **11**(2), 91–126.

Brandstein, M. and Ward, D. [2001] *Microphone Arrays — Signal Processing Techniques and Applications*, New York, Springer.

Buchner, H. and Kellermann, W. [2001] "An acoustic human-machine interface with multi-channel sound reproduction," in *Proceedings of IEEE Fourth Workshop on Multimedia Signal Processing*, Cannes, France, October 3–5, pp. 359–364.

Champagne, B., Bedard, S. and Stephenne, A. [1996] "Performance of time-delay estimation in the presence of room reverberation," *IEEE Transactions on Speech and Audio Processing* **4**(2), 148–152.

Chan, Y. T. and Ho, K. C. [1994] "A simple and efficient estimator for hyperbolic location," *IEEE Transactions on Signal Processing* **42**(8), 1905–1915.

Chen, J. C., Hudson, R. E. and Yao, K. [2002] "Maximum-likelihood sound localization and unknown sensor location estimation for wideband signals in the near-field," *IEEE Transactions on Signal Processing* **50**(8), 1843–1854.

Chen, J., Jiang, L. and Ser, W. [2000] "A framework for speaker tracking using microphone array and camera," in *Proceedings of 5th International Conference on Signal Processing*, Vol. 2, Beijing, China, August 21–25, pp. 1384–1387.

Cook, P. [1999] *Music, Cognition, and Computerized Sound — An Introduction to Psychoacoustics*, Cambridge, Massachusetts: The MIT Press.

Cui, Y. J. and Ge, S. S. [2003] "Autonomous vehicle positioning with gps in urban canyon environments," *IEEE Transactions on Robotics and Automation* **19**(1), 15–25.

Desloge, J. G., Rabinowitz, W. and Zurek, P. [1997] "Microphone-array hearing aids with binaural output — Part 1: Fixed-processing systems," *IEEE Transactions on Speech and Audio Processing* **5**(6), 529–542.

Ferguson, B. G., Criswick, L. G. and Lo, K. W. [2002] "Locating far-field impulsive sound sources in air by triangulation," *Journal of the Acoustical Society of America* **111**(1), 104–116.

Ferguson, B. G. and Lo, K. W. [2002] "Passive ranging errors due to multipath distortion of deterministic transient signals with application to the localization of small arms fire," *Journal of the Acoustical Society of America* **111**(1), 117–128.

Gazor, S. S. and Grenier, Y. [1995] "Criteria for positioning of sensors for a microphone array," *IEEE Transactions on Speech and Audio Processing* **3**(4), 294–303.

Gazor, S. S. and Grenier, Y. [1996] "An algorithm for multisource beamforming and multitarget tracking," *IEEE Transactions on Signal Processing* **44**(6), 1512–1522.

Ge, S. S., Hang C., Lee T. and Zhang, T. [2001] *Stable Adaptive Neural Network Control*, ser. The Kluwer International Series on Asian Studies in Computer and Information Science, Norwell, USA, Kluwer Academic, **13**.

Ge, S. S., Lee, T. and Harris, C. [1998] *Adaptive Neural Network Control of Robotic Manipulators*, River Edge, NJ, World Scientific.

Guivant, J. E. and Nebot, E. M. [2003] "Solving computational and memory requirements of feature-based simultaneous localization and mapping algorithms," *IEEE Transactions on Robotics and Automation* **19**(4), 749–755.

Gustafsson, T., Rao, B. D. and Trivedi, M. [2003] "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Transactions on Speech and Audio Processing* **11**(6), 791–803.

Huang, J., Ohnishi, N. and Sugie, N. [1995] "A biomimetic system for localization and separation of multiple sound sources," *IEEE Transactions on Instrumentation and Measurement* **44**, 733–738.

Huang, J., Ohnishi, N. and Sugie, N. [1997] "Sound localization in reverberant environment based on the model of the precedence effect," *IEEE Transactions on Instrumentation and Measurement* **46**, 842–846.

Huang, J., Supaongprapa, T., Terakura, I., Ohnishi, N. and Sugie, N. [1997] "Mobile robot and sound localization," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 2, Grenoble, France, September 7–11, pp. 683–689.

Huang, Y. T., Benesty, J., Elko, G. W. and Mersereau, R. M. [2001] "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Transaction on Speech and Audio Processing* **9**(8), 943–956.

Katkovnik, V. and Gershman, A. B. [2000] "A local polynomial approximation based beamforming for source localization and tracking in nonstationary environments," *IEEE Signal Processing Letters* **7**(1), 3–5.

Knapp, C. H. and Carter, G. C. [1976] "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing* **ASSP-24**(4), 320–327.

Morimoto, M. [2001] "The contribution of two ears to the perception of vertical angle in sagittal planes," *Journal of the Acoustical Society of America* **109**(4), 1596–1603.

Pu, C. J., Harris, J. and Principe, J. C. [1997] "A neuromorphic microphone for sound localization," in *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, Computational Cybernetics and Simulation*, Vol. 2, Orlando, USA, pp 1469–1474.

Rucci, M., Edelman, G. M. and Wray, J. [1999] "Adaptation of orienting behavior: From the barn owl to a robotic system," *IEEE Transactions on Robotics and Automation* **15**(1), 96–110.

Shinn-Cunningham, B. G., Santarelli, S. and Kopco, N. [2000] "Tori of confusion: Binaural localization cues for sources within reach of a listener, *Journal of the Acoustical Society of America* **107**(3), 1627–1636.

Tabrikian, J. and Messer, H. [1996] "Three-dimensional source localization in a waveguide," *IEEE Transactions on Signal Processing* **44**(1), 1–13.

Tanaka, M. and Kaneda, Y. [1993] "Performance of sound source direction estimation methods under reverberant conditions," *Journal of the Acoustical Society of Japan E* **14**(4), 291–292.

Wang, Q. H., Ivanov, T. and Arari, P. [2004] "Acoustic robot navigation using distributed microphone arrays," *Information Fusion* **5**(2), 131–140.

Welker, D., Greenberg, J., Desloge, J. and Zurek, P. [1997] "Microphone-array hearing aids with binaural output. (ii) A two-microphone adaptive system," *IEEE Transactions on Speech and Audio Processing* **5**(6), 543–551.

Weng, J. Y. and Guentchev, K. Y. [2001] "Three-dimensional sound localization from a compact non-coplanar array of microphones using tree-based learning," *Journal of the Acoustical Society of America* **110**(1), 310–323.

## Biography

**Shuzhi Sam Ge** received the BSc degree from Beijing University of Aeronautics and Astronautics (BUAA) in 1986, and the PhD degree as well as the Diploma of Imperial College (DIC) from Imperial College of Science, Technology and Medicine, in 1993. He is in the Department of Electrical and Computer Engineering, National University of Singapore since 1993. He has authored and co-authored over 200 international journal and conference papers, three monographs and co-invented three patents. He is a member of the Technical Committee on Intelligent Control since 2000, an associate editor of IEEE Transactions on Control Systems Technology since 1999, an associate editor of IEEE Transactions on Automatic Control since 2004, a corresponding editor for Asia and Australia, IEEE Control Systems Magazine since 2004, and an associate editor of IEEE Transactions on Neural Networks, since 2004. His current research interests include the control of nonlinear systems, neural/fuzzy systems, robotics, hybrid systems, sensor fusion, and system development.

**Ai Poh Loh** received the BEng (Electrical 1st class) from the University of Malaya, Kuala Lumpur, in 1983 and the DPhil degree in control from Oxford University in 1986.

She is currently an Associate Professor and Deputy Head (Academic) at the Department of Electrical and Computer Engineering at the National University of Singapore. From 1994 to 1997, she was a visiting lecturer at MIT. Her research interests include auto-tuning, fault detection, signal processing and nonlinear adaptive control.

**Feng Guan** received the BEng from Wuhan University of Hydraulics and Electrics, Wuhan, China, in 1997 and MEng from Shanghai Jiaotong University, Shanghai, China, in 2001, respectively.

During 2001 to 2005, he is a Research Scholar working toward the PhD degree in the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His current research interests are in the fields of sound localization and sensor fusion.