VSP

Advanced Robotics 25 (2011) 135–152

ADVANCED ROBOTICS
brill.nl/ar

Full paper

Simultaneous Localization of a Mobile Robot and Multiple Sound Sources Using a Microphone Array

Jwu-Sheng Hu [a], **Chen-Yu Chan** [a], **Cheng-Kang Wang** [a], **Ming-Tang Lee** [a,*] **and Ching-Yi Kuo** [b]

[a] Department of Electrical Engineering, National Chiao-Tung University, Hsinchu, Taiwan, ROC
[b] Robotics System Integration Technology Department, Intelligent Robotics Technology Division, Mechanical and Systems Research Laboratories, ITRI, Taiwan, ROC

Received 24 November 2009; accepted 12 February 2010

Abstract
Sound source localization is an important function in robot audition. Most existing works perform sound source localization using static microphone arrays. This work proposes a framework that simultaneously localizes the mobile robot and multiple sound sources using a microphone array on the robot. First, an eigenstructure-based generalized cross-correlation method for estimating time delays between microphones under multi-source environments is described. Using the estimated time delays, a method to compute the far-field source directions as well as the speed of sound is proposed. In addition, the correctness of the sound speed estimate is utilized to eliminate spurious sources, which greatly enhances the robustness of sound source detection. The arrival angles of the detected sound sources are used as observations in a bearing-only simultaneous localization and mapping procedure. As the source signals are not persistent and there is no identification of the signal content, data association is unknown and it is solved using the FastSLAM algorithm. The experimental results demonstrate the effectiveness of the proposed method.

Keywords
Sound source localization, microphone array, time delay estimation, bearing-only SLAM, generalized cross correlation

1. Introduction

An audition system is a very important feature for an intelligent robot. The fundamental requirement of this system allows a robot to interact with humans through speech dialog. Under this requirement, there are several research issues currently active in the robotics community. These issues include speaker localization, speech

* To whom correspondence should be addressed. E-mail: lhoney.ece95g@nctu.edu.tw

 DOI:10.1163/016918610X538525

separation and enhancement, speech recognition and natural dialog, speaker identification and multi-model interaction, etc. [1–5]. Among them, speaker localization using either the biological hearing principle [5] or a microphone array [1] has drawn a lot of attention for many years [6].

The underlying principle to localize a sound source using a microphone array is based on the time difference of arrival (TDOA) among spatially distributed microphones. For distance localization, the method of triangulation is used, and the accuracy depends on the ratio between the microphone spacing and the distance. Since the array spacing on a mobile robot is usually small compared with the distance to the source, it is unlikely to obtain accurate distance information [7]. Hence, most sound source localization research on mobile robots emphasized detecting the source directions. Hardly any work tried to solve the problem of localizing the robot and multiple sound sources simultaneously. Mobility is a unique advantage of the robot over a stationary microphone array. When moving in space, the robot effectively increases the array spacing and it is possible to compute the source distance by using the source direction information only. This is equivalent to the standard bearing-only localization problem [8]. However, it is more complicated when dealing with multiple sources as the signals are mixed together in the array measurement. Secondly, the sound source signals may not be persistent all the time. Unless the contents of source signals can be clearly identified, there will be a source association problem. The data association becomes more difficult for non-persistent and moving sources. Although other types of sensors such as vision can be incorporated, exploring the technological boundary of localization using sound measurement alone is still needed. For example, occlusion or a sudden lighting variation could make visual recognition fail easily.

The first issue of sound source localization is the robustness of source detection, especially under a multi-source environment with reverberation. Generalized cross-correlation (GCC) [9] is one of the common methods discussed for robot localization application [10]. For multiple sources, MUSIC [11] is used for eliminating the coherence problem and it has been applied to the robot audition system [12]. Walworth *et al.* [13] proposed a linear equation formulation for the estimation of the three-dimensional (3-D) position of a wave source based on the time delay values. Valin *et al.* [1] gave a simple solution for the linear equation in Ref. [13] based on the far-field assumption. Yao *et al.* [14] presented a source linear equation similar to Ref. [13] to estimate the source location and velocity by using the least-squares method. This paper presents a method of computing arrival delays of multiple sources by combining the idea of MUSIC and GCC. Further, the source linear equation of Ref. [14] is modified for direction estimation of far field sources. The distinct advantage of the method is that the information about the number of sources and speed of sound is not needed. In fact, the speed of sound is computed for each possible source and the value is used to check if it is a valid one. This greatly enhances the robustness of source detection.

The source directions obtained from the proposed method serve as the observation data for the bearing-only localization framework. Since there is no additional information about the content of the source signals, the observation data sequences require association. The problem is solved by using the FastSLAM algorithm [15], where incorrect associations of sound sources tend to possess inconsistent positions. Experiments were conducted using an eight-channel microphone array on a mobile robot. It is shown that the overall system effectively localizes the robot and sound sources in a room environment.

2. Sound Source Direction Estimation

In this section, a method of estimating directions of multiple unknown sound sources using a microphone array is introduced [15]. The novelty of this method is the ability to separate source arrival angles simultaneously without knowing the speed of sound. Further, the estimated speed of sound associated with each source is used to verify the existence of such a source. This is necessary since there is no information about the number of sources in the measurement.

2.1. Near-Field Influence Factor and Field Distance Ratio

The work in Ref. [14] provides a close form solution for estimating the source locations and sound propagation speed using multiple microphones. The accuracy depends on the aperture of the microphone geometry as well as the distance to the source. In our case, microphones are installed only on the robot. This makes the aperture relatively small compared with the source distance in most cases. As a result, it is necessary to consider the far-field scenario. Let the source location be $\mathbf{r}_s = [x_s \;\; y_s \;\; z_s]^T$, the ith sensor locations $\mathbf{r}_i$, and the relative time delays, $t_i - t_j$, between the ith sensor and jth sensor. The original equation of the delay relation (from (15) of Ref. [14]) is:

$$-\frac{(\mathbf{r}_i - \mathbf{r}_0)\cdot(\mathbf{r}_s - \mathbf{r}_0)}{v|\mathbf{r}_s - \mathbf{r}_0|} + \frac{|\mathbf{r}_i - \mathbf{r}_0|^2}{2v|\mathbf{r}_s - \mathbf{r}_0|} - \frac{v(t_i - t_0)^2}{2|\mathbf{r}_s - \mathbf{r}_0|} = (t_i - t_0), \tag{1}$$

where $j = 0$ without loss of generality and v is the speed of sound. Define $\hat{\mathbf{r}}_s$ and ρ_i as:

$$\hat{\mathbf{r}}_s = \frac{\mathbf{r}_s - \mathbf{r}_0}{|\mathbf{r}_s - \mathbf{r}_0|} \quad \text{and} \quad \rho_i = \frac{|\mathbf{r}_i - \mathbf{r}_0|}{|\mathbf{r}_s - \mathbf{r}_0|}, \tag{2}$$

where $\hat{\mathbf{r}}_s$ represents the unit vector in the source direction, and ρ_i means the ratio of the array size (aperture) to the distance between the array and source, i.e. for far-field sources, $\rho_i \ll 1$. Substituting (2) to (1), we have:

$$-(\mathbf{r}_i - \mathbf{r}_0)\frac{\hat{\mathbf{r}}_s}{v} + \frac{|\mathbf{r}_i - \mathbf{r}_0|}{v}\frac{\rho_i}{2} - \frac{1}{v}\frac{v^2(t_i - t_0)^2}{|\mathbf{r}_i - \mathbf{r}_0|}\frac{\rho_i}{2} = (t_i - t_0). \tag{3}$$

The term $v(t_i - t_0)$ means the difference between the sound source to the ith and the 0th microphones. Let the difference be d_i, i.e.:

$$d_i = v(t_i - t_0) = |\mathbf{r}_s - \mathbf{r}_i| - |\mathbf{r}_s - \mathbf{r}_0|. \tag{4}$$

Equation (3) can be rewritten as:

$$-\frac{(\mathbf{r}_i - \mathbf{r}_0)}{v} \cdot \hat{\mathbf{r}}_s + f_i \frac{\rho_i}{2} = (t_i - t_0), \tag{5}$$

where:

$$f_i = \frac{|\mathbf{r}_i - \mathbf{r}_0|}{v} - \frac{|d_i|}{v} \frac{|d_i|}{|\mathbf{r}_i - \mathbf{r}_0|}. \tag{6}$$

It is straightforward to see that $f_i \geqslant 0$ since:

$$d_i \leqslant |\mathbf{r}_i - \mathbf{r}_0|. \tag{7}$$

Also, f_i achieves its maximum of $|\mathbf{r}_i - \mathbf{r}_0|/v$ when $d_i = 0$ (i.e., when the source is located along the line passing through the midpoint of the segment connecting microphone i and 0, and perpendicular to them). This also means that f_i has an order of magnitude less than or equal to the vector $(\mathbf{r}_i - \mathbf{r}_0)/v$. Therefore, from (5), it is clear that for far-field sources ($\rho_i \ll 1$), the delay relation approaches:

$$-\frac{(\mathbf{r}_i - \mathbf{r}_0)}{v} \cdot \hat{\mathbf{r}}_s = (t_i - t_0). \tag{8}$$

Equation (8) can also be derived from the plane wave propagation perspective [1]. However, the derivation above can clearly explain the far-field term and near-field influence of the delay relation on the left-hand side of (5). We define ρ_i as the field distance ratio and f_i as the near-field influence factor for their roles in source localization using an array of sensors.

2.2. Least-Squares Solutions

For an array of M sensors, (8) becomes a system of linear equations:

$$\mathbf{A}_s \mathbf{w}_s = \mathbf{b}, \tag{9}$$

where:

$$\mathbf{w}_s \equiv [\, w_1 \quad w_2 \quad w_3 \,]^T = \frac{\mathbf{r}_s}{v|\mathbf{r}_s|} = \frac{\hat{\mathbf{r}}_s}{v} \tag{10}$$

$$\mathbf{A}_s \equiv [\, -(\mathbf{r}_1 - \mathbf{r}_0) \quad -(\mathbf{r}_2 - \mathbf{r}_0) \quad \cdots \quad -(\mathbf{r}_{M-1} - \mathbf{r}_0) \,]^T \tag{11}$$

$$\mathbf{b} \equiv [\, t_1 - t_0 \quad t_2 - t_0 \quad \cdots \quad t_{M-1} - t_0 \,]^T. \tag{12}$$

It is, therefore, easy to estimate the speed of sound:

$$v = \frac{1}{|w_s|} = \frac{1}{|(\mathbf{A}_s^T \mathbf{A}_s)^{-1} \mathbf{A}_s^T \mathbf{b}|}. \tag{13}$$

The sound source direction can be given by:

$$\hat{\mathbf{r}}_{\mathrm{s}} = \frac{w_{\mathrm{s}}}{|w_{\mathrm{s}}|} = \frac{(\mathbf{A}_{\mathrm{s}}^{\mathrm{T}}\mathbf{A}_{\mathrm{s}})^{-1}\mathbf{A}_{\mathrm{s}}^{\mathrm{T}}\mathbf{b}}{|(\mathbf{A}_{\mathrm{s}}^{\mathrm{T}}\mathbf{A}_{\mathrm{s}})^{-1}\mathbf{A}_{\mathrm{s}}^{\mathrm{T}}\mathbf{b}|}. \tag{14}$$

As a result, the bearings of the source to the sensors can be computed by:

$$\hat{\mathbf{r}}_{\mathrm{s}} = [\cos\theta_{\mathrm{s}}\sin\varphi_{\mathrm{s}} \quad \sin\theta_{\mathrm{s}}\sin\varphi_{\mathrm{s}} \quad \cos\varphi_{\mathrm{s}}]^{\mathrm{T}}, \tag{15}$$

where θ_{s} and φ_{s} are azimuth and elevation angle, respectively. It is straightforward to verify that $\mathbf{A}_{\mathrm{s}}$ reduces rank if the vectors constructed by sensor pairs do not span the 3-D space (i.e., a planar array), meaning the delay relation is satisfied by more than one source direction. Secondly, (8) is actually an approximation by considering plane wave propagation. Please refer to Ref. [15] for detailed analysis of the approximation errors and array geometry issues.

The solutions of (13) and (14) are useful only when the delay among microphones can be estimated within a certain accuracy. For multiple sources, the estimation becomes more difficult as the signals are mixed together in the measurements. In the next section, an eigenstructure-based (ES)-GCC method is presented to cope with this issue.

2.3. Delay Estimation of Multiple Sources

Consider an array with M microphones on a mobile robot. The received signal of the mth microphone that contains d sources can be described by short-term Fourier transform (STFT) as:

$$X_m(\omega_f, k) = \sum_{p=1}^{d} a_{mp} S_p(\omega_f, k) \mathrm{e}^{-j\omega_f \tau_{mp}} + N_m(\omega_f, k), \quad f = 1, 2, \ldots, F, \tag{16}$$

where a_{mp} is the amplitude from the pth sound source to the mth microphone, τ_{mp} is the associated delay, $N_m(\omega_f, k)$ is the interference, ω_f is the frequency band and k is the frame number. Rewrite (16) in matrix form:

$$\mathbf{X}(\omega_f, k) = \mathbf{A}(\omega_f)\mathbf{S}(\omega_f, k) + \mathbf{N}(\omega_f, k), \tag{17}$$

where:

$$\mathbf{X}^{\mathrm{T}}(\omega_f, k) = [X_1(\omega_f, k), \ldots, X_M(\omega_f, k)]$$
$$\mathbf{N}^{\mathrm{T}}(\omega_f, k) = [N_1(\omega_f, k), \ldots, N_M(\omega_f, k)]$$
$$\mathbf{S}^{\mathrm{T}}(\omega_f, k) = [S_1(\omega_f, k), \ldots, S_d(\omega_f, k)]$$
$$\mathbf{A}(\omega_f, k) = \begin{bmatrix} a_{11}\mathrm{e}^{-j\omega_f\tau_{11}} & \cdots & a_{1d}\mathrm{e}^{-j\omega_f\tau_{1d}} \\ \vdots & & \vdots \\ a_{M1}\mathrm{e}^{-j\omega_f\tau_{M1}} & \cdots & a_{Md}\mathrm{e}^{-j\omega_f\tau_{Md}} \end{bmatrix}.$$

The received signal correlation matrix with eigenvalue decomposition can be de-

scribed as:

$$\mathbf{R}_{xx}(\omega_f) = \frac{1}{N}\sum_{k=1}^{N}\mathbf{X}(\omega_f,k)\mathbf{X}^H(\omega_f,k) = \sum_{i=1}^{M}\lambda_i(\omega_f)\mathbf{V}_i(\omega_f)\mathbf{V}_i^H(\omega_f), \tag{18}$$

where $\lambda_i(\omega_f)$ and $\mathbf{V}_i(\omega_f)$ are eigenvalues and corresponding eigenvectors with $\lambda_1(\omega_f) \geqslant \lambda_2(\omega_f) \geqslant \cdots \geqslant \lambda_M(\omega_f)$, and $\mathbf{V}_1(\omega_f)$ is the principal component vector of the sound source at frequency ω_f, which is defined as:

$$\mathbf{V}_1(\omega_f) = [\, V_{11}(\omega_f) \quad V_{12}(\omega_f) \quad \cdots \quad V_{1M}(\omega_f)\,]^{\mathrm{T}}. \tag{19}$$

The principal component vector contains the directional information of the principal sound sources at each frequency. As a result, the principal component matrix at each frequency can be established as:

$$\mathbf{E}_1 = \begin{bmatrix} V_{11}(\omega_1) & V_{11}(\omega_2) & \cdots & V_{11}(\omega_F) \\ V_{12}(\omega_1) & V_{12}(\omega_2) & \cdots & V_{12}(\omega_F) \\ \vdots & \vdots & & \vdots \\ V_{1M}(\omega_1) & V_{1M}(\omega_2) & \cdots & V_{1M}(\omega_F) \end{bmatrix}. \tag{20}$$

The fth column can be considered as the distribution vector of the received signal on M microphones at frequency ω_f. Hence, the eigenstructure-based GCC function between the ith and jth microphone can be represented as:

$$R_{x_i x_j}(\tau) = \int_{\omega_1}^{\omega_F} V_{1i}(\omega)V_{1j}(\omega)\mathrm{e}^{j\omega\tau}\,\mathrm{d}\omega. \tag{21}$$

The time delay can be estimated by finding the peaks of the ES-GCC function:

$$\hat{\tau}_{\text{ES-GCC}} = \arg\max_x R_{x_i x_j}(\tau). \tag{22}$$

2.4. Direction Estimation for Multiple Sources

For multiple sources, there will be multiple peaks in the GCC function of (21) for each pair of microphones and multiple delays are obtained at each STFT frame. The question is how to combine these delays among microphone pairs to form the vector $\mathbf{b}$ of (12). Denote τ_{jk} as the kth delay of the microphone pair $(j, 0)$, $k = 1, \ldots, q_j$, where q_j is the total number of delays (peaks) of this pair. Note that q_j may be different for different pairs (depending on the threshold level of the peak value). For M microphones, there will be $(q_1 \times q_2 \times \cdots \times q_{M-1})$ number of possible combinations of the vector $\mathbf{b}$. However, since the minimum number of microphone pairs to solve (9) is 3, we can sort out the combination by starting from three pairs and iteratively adding additional pairs. Without loss of generality, assume the indices of microphone pairs are arranged in the order such that $q_1 \geqslant q_2 \geqslant q_3 \geqslant q_4 \geqslant \cdots \geqslant q_{M-1}$. Then the delay vector of each source can be found

by minimizing the error between the associated sound speed estimation and the nominal one (e.g., 340 m/s). Specifically, a set of possible sound sources can be found as:

$$S = \{(l, m, n) \mid |e_{lmn}| \leqslant \bar{e} \text{ for } 1 \leqslant l \leqslant q_1, 1 \leqslant m \leqslant q_2, 1 \leqslant n \leqslant q_3\}, \tag{23}$$

where

$$e_{lmn} = \frac{1}{|(\mathbf{A}_3^{\mathrm{T}}\mathbf{A}_3)^{-1}\mathbf{A}_3^{\mathrm{T}}\mathbf{b}_{lmn}|} - \bar{v} \tag{24}$$

$$A_i = \begin{bmatrix} x_1 - x_0 & y_1 - y_0 & z_1 - z_0 \\ \vdots & \vdots & \vdots \\ x_i - x_0 & y_i - y_0 & z_i - z_0 \end{bmatrix}. \tag{25}$$

$\mathbf{b}_{lmn} = [\tau_{1l} \ \ \tau_{2m} \ \ \tau_{3n}]^{\mathrm{T}}$ and $\bar{v}$ is the nominal speed of sound. Note that the error bound $\bar{e}$ is imposed so that some of the delay vectors with unreasonable speed of sound can be eliminated. This is the advantage of the proposed method comparing with classical methods like MUSIC to screen out sources that are not real (e.g., electronic noise). Secondly, the possible number of sound sources can be greater than q_1 since multiple sources could result in the same delay for a microphone pair. Next, the delays from microphone pair 4 to pair $(M - 1)$ can be added similarly. The process is quite straightforward and the explanation is omitted here. Laboratory experience showed that a correct number of sources can be obtained repeatedly for the error bound $\bar{e} = 15$ m/s [15].

The resulting delay vectors computed through this process can be used to obtain the source directions by (14) and (15).

3. Localization of Sources and Robot

The simultaneous localization and mapping (SLAM) problem is the procedure of recognizing a set of feature landmarks and localizing the sensor odometer with respect to the landmark set. A microphone array platform carried by a two-wheeled robot was used in this paper to perform the localization of the robot and feature landmarks (sound sources). According to Section 2, the microphone array is capable of recognizing an unknown number of sound sources as the feature points and obtaining the associated angles of arrival. The angles are considered as the bearing measurements and this becomes a standard bearing-only SLAM problem [8]; the FastSLAM algorithm [16, 17] is adopted. FastSLAM estimates the robot path using a particle filter and the map feature locations are estimated using the extended Kalman filter (EKF). Each particle possesses its own set of EKFs for all feature points. Particles in FastSLAM are denoted as:

$$\mathbf{Y}_t^{[k]} = (\mathbf{X}_t^{[k]}, \boldsymbol{\mu}_{1,t}^{[k]}, \boldsymbol{\Sigma}_{1,t}^{[k]}, \boldsymbol{\mu}_{2,t}^{[k]}, \boldsymbol{\Sigma}_{2,t}^{[k]}, \ldots, \boldsymbol{\mu}_{d,t}^{[k]}, \boldsymbol{\Sigma}_{d,t}^{[k]}), \tag{26}$$

where $[k]$ is the index of the particle; $\mathbf{X}_t^{[k]} = [(x, y, \theta)^{\mathrm{T}}]_t^{[k]}$ is the pose estimate of the robot at time t, and $\boldsymbol{\mu}_{p,t}^{[k]}$ and $\boldsymbol{\Sigma}_{p,t}^{[k]}$ are the mean and covariance of the pth land-

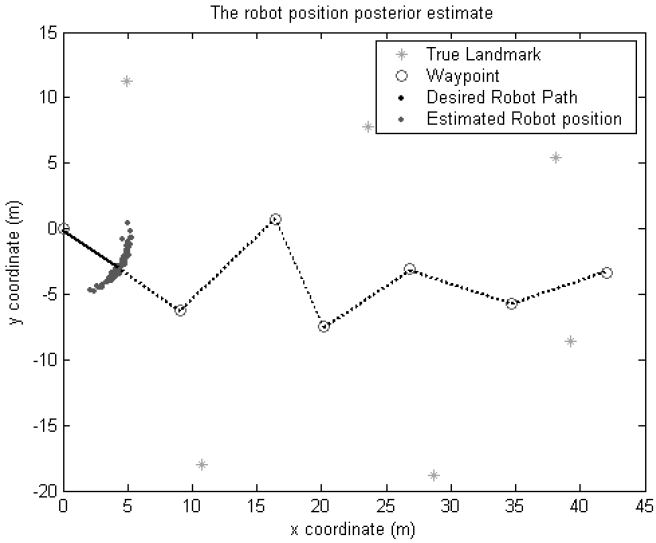

Figure 1. Robot position posterior estimation.

mark location, which is assumed to be Gaussian. The algorithm can be separated into the following three steps:

3.1. Step 1: Sampling New Pose According to Path Posterior

For each particle at time t, the control input $\mathbf{u}_t$ is used to estimate the $\mathbf{Y}_t^{[k]}$ from $\mathbf{Y}_{t-1}^{[k]}$. It samples the new robot position $\mathbf{X}_t^{[k]}$ according to the posterior:

$$\mathbf{X}_t^{[k]} \sim p(\mathbf{X}_t^{[k]} | \mathbf{X}_t^{[k]}, \mathbf{u}_t), \tag{27}$$

where $\mathbf{X}_{t-1}^{[k]}$ is the posterior estimate of robot location at time $t-1$. The sampling step can be seen graphically in Fig. 1.

3.2. Step 2: Use the Observation to Update the Feature Estimation

At this step, the posterior of the feature point is estimated. The update is stated here with the normalizer η denoted by:

$$p(\mathbf{m}|\mathbf{X}_{1:t}, \mathbf{Z}_{1:t}) = \eta \cdot p(\mathbf{Z}_t|\mathbf{X}_t, \mathbf{m}) p(\mathbf{m}|\mathbf{X}_{1:t-1}, \mathbf{Z}_{1:t-1}), \tag{28}$$

where $\mathbf{m}$ are the feature landmarks and $\mathbf{Z}_{i:j}$ is the observation from time step i to j. The probability distribution of landmarks $p(\mathbf{m}|\mathbf{X}_{1:t-1}, \mathbf{Z}_{1:t-1})$ at time $t-1$ is represented by a Gaussian distribution with mean $\boldsymbol{\mu}_{p,t-1}^{[k]}$ and covariance $\boldsymbol{\Sigma}_{p,t-1}^{[k]}$. For the new estimation, FastSLAM linearizes the perceptual model $p(\mathbf{Z}_t|\mathbf{X}_t, \mathbf{m})$ in the same way as the EKF. The measurement function h could be approximated by a Taylor expansion:

$$\begin{aligned} h(\mathbf{m}, \mathbf{X}_t^{[k]}) &\approx h(\mathbf{u}_{t-1}^{[\mathrm{k}]}, \mathbf{X}_t^{[k]}) + h'(\mathbf{X}_t^{[k]}, \mathbf{u}_{t-1}^{[\mathrm{k}]})(\mathbf{m} - \boldsymbol{\mu}_{t-1}^{[\mathrm{k}]}) \\ &= \hat{\mathbf{Z}}_t^{[\mathrm{k}]} + \mathbf{H}_t^{[\mathrm{k}]}(\mathbf{m} - \boldsymbol{\mu}_{t-1}^{[\mathrm{k}]}). \end{aligned} \tag{29}$$

Here the derivative h' is taken with respect to the feature landmarks $\mathbf{m}$. The approximation is tangent to h at $\mathbf{X}_t^{[k]}$ and $\mathbf{u}_{t-1}^{[k]}$. The new mean and covariance could be obtained using the standard EKF measurement update.

$$\mathbf{K}_t^{[k]} = \boldsymbol{\Sigma}_{t-1}^{[k]} \mathbf{H}_t^{[k]} (\mathbf{H}_t^{[k]\mathrm{T}} \boldsymbol{\Sigma}_{t-1}^{[k]} \mathbf{H}_t^{[k]} + \mathbf{R}_t)^{-1} \tag{30}$$

$$\boldsymbol{\mu}_t^{[k]} = \boldsymbol{\mu}_{t-1}^{[k]} + \mathbf{K}_t^{[k]} (\mathbf{Z}_t - \hat{\mathbf{Z}}_t^{[k]}) \tag{31}$$

$$\boldsymbol{\Sigma}_t^{[k]} = (\mathbf{I} - \mathbf{K}_t^{[k]} \mathbf{H}_t^{[k]\mathrm{T}}) \boldsymbol{\Sigma}_{t-1}^{[k]}. \tag{32}$$

After repeating Steps 1 and 2 M times, the temporary set of M particles is created.

3.3. Step 3: Resampling

In the final step, FastSLAM resamples the set of the M particles. First, we will calculate the importance factor of each particle. The factor is given by:

$$w_t^{[k]} \approx \eta |2\pi \mathbf{Q}_t^{[k]}|^{-1/2} \mathrm{e}^{-(1/2)(\mathbf{Z}_t - \hat{\mathbf{Z}}_t^{[k]})^{\mathrm{T}} (\mathbf{Q}_t^{[k]})^{-1} (\mathbf{Z}_t - \hat{\mathbf{Z}}_t^{[k]})}, \tag{33}$$

with the covariance:

$$\mathbf{Q}_t^{[k]} = \mathbf{H}_t^{[k]\mathrm{T}} \boldsymbol{\Sigma}_{p,t-1}^{[k]} \mathbf{H}_t^{[k]} + \mathbf{R}_t, \tag{34}$$

which means the closer the particle's estimation is to the observation, the more important it is. After all the weighting is computed, the real probability distribution is described by this weighting.

One of the key features of FastSLAM is that as long as a small subset of the particles is based on the correct association, data association is not as fatal as in EKF approaches. Particles with incorrect data association tend to possess inconsistent feature positions, which increases the probability that will be sampled away during the resampling phase of the algorithm.

4. Experimental Results

An eight-channel omni-directional microphone array was constructed using digital microphones. The digital microphone integrates an electric condenser microphone core, an analog output amplifier and a sigma–delta modulator on a single chip [15]. The digital bit-stream transmission achieves minimum interference compared with conventional analog microphone signals. The microphone array topology and the mobile robot for the experiment are shown in Fig. 2. Note that it is a 3-D microphone array that is able to estimate the sound source elevation angle. In this experiment, however, this angle is ignored since the localization concerns 2-D locations of the robot and the sound sources.

The sampling rate of the microphone array is 16 kHz and each STFT frame contains 512 samples with 256 overlapping samples. The sound arrival angles are computed after collecting 20 frames, which is about 3 times per second. The program procedure for calculating ES-GCC and FastSLAM is shown in Fig. 3.

Figure 2. Digital microphone array mounted on the robot.

The algorithm needs to accumulate a certain time frame to get solid delay information between microphones. Only those delay combinations with reasonable sound speed generate a DOA estimation. Furthermore, the DOA estimation is not necessary correct, which may be eliminated in the Outlier Elimination step. For all the reasons above, the DOA estimation will not be correct when the robot is moving. Thus, we program the algorithm to update the estimation only when the robot stops.

Experiments were performed in two different cases. The room size for the two experiments was 4750 mm × 3600 mm and height of 3600 mm approximately, and its reverberation time at 1000 Hz was 0.52 s. Room temperature was approximately 20°C. In the following, we test the ES-GCC part and the whole algorithm in two different cases C1 and C2.

4.1. Reliability of Direction Estimation for Multiple Sources (C1)

In case C1, we first evaluate the time delay estimation (TDE) performance of the ES-GCC part for a single-source case and then we compare the direction estimation results for a multiple-source case.

There are six speech sources played by loudspeakers and a white noise generated from an air conditioner. The speech sources are Chinese speech (female and male). The relative locations between the sources, noise and microphone array are described in Fig. 4. The distances between these sound sources and the microphone array are all 2400 mm. Note that the microphone array in C1 is not moving. In Fig. 4, the microphone locations are the following (mm):

$$\text{Mic. } 1 = [\,200 \quad 100 \quad 0\,], \qquad \text{Mic. } 2 = [\,200 \quad -100 \quad 0\,]$$

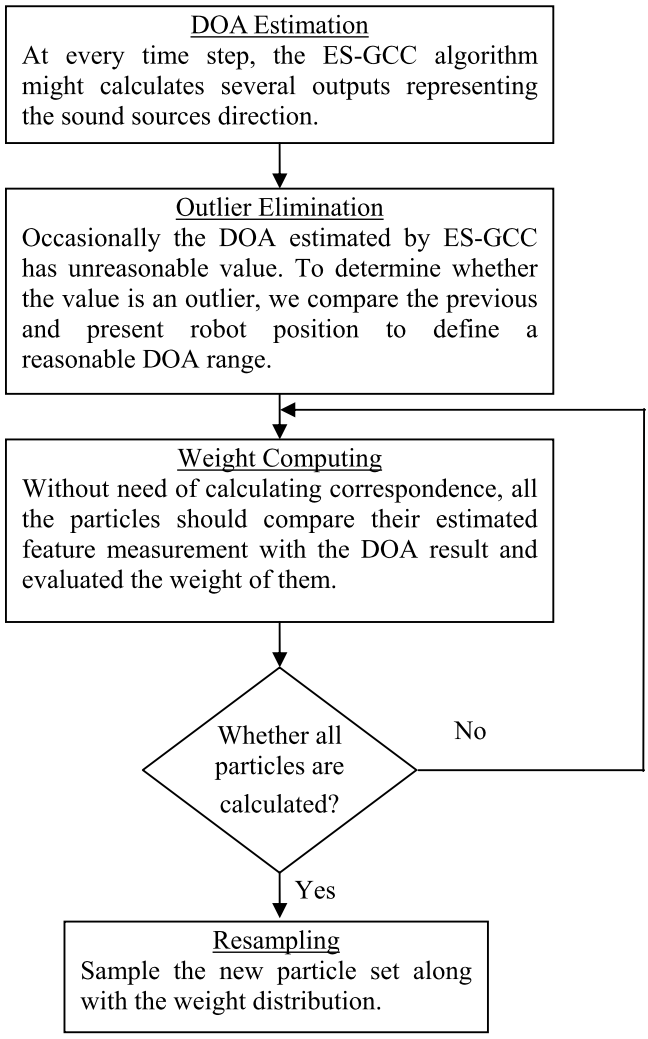

Figure 3. Algorithm procedure.

$$\text{Mic. } 3 = [-200 \quad -100 \quad 0], \qquad \text{Mic. } 4 = [-200 \quad 100 \quad 0]$$
$$\text{Mic. } 5 = [0 \quad 100 \quad 100], \qquad \text{Mic. } 6 = [0 \quad 100 \quad -100]$$
$$\text{Mic. } 7 = [0 \quad -100 \quad 100], \qquad \text{Mic. } 8 = [0 \quad -100 \quad -100].$$

The air conditioner, which is 4000 mm from the first microphone, is turned on during this experiment (noise in Fig. 4).

The TDE of the ES-GCC part is compared with two GCC-based algorithms, GCC-PHAT and GCC-ML [9]. The microphone array is divided into seven pairs to implement the GCC-based algorithms (i.e., (Mic. 1, Mic. 2), ..., (Mic. 1, Mic. 8)), and the TDEs are calculated separately. Here, we use a performance index, root

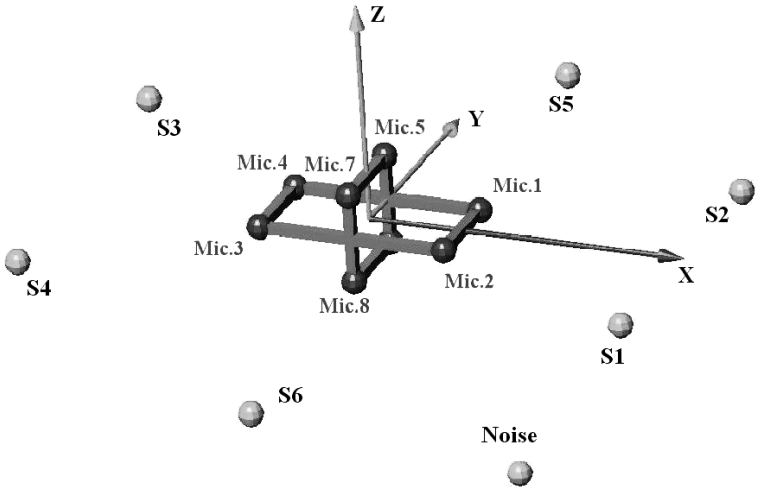

Figure 4. Sound source locations relative to the microphone array for C1.

mean square error (RMSE), to evaluate the performance of the proposed method, which is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_\mathrm{T}} \sum_{i=1}^{N_\mathrm{T}} (\hat{D}_i - D_i)^2}, \tag{35}$$

where N_T is the total number of estimations (i.e., total number of estimated time delays for six sources and seven microphone pairs), $\hat{D}_i$ is the ith time delay estimation and D_i is the ith correct delay sample. Figure 5 shows the RMSE results as a function of SNR for three different TDE algorithms. The total number of estimation N_T is 300. The SNR is defined as the average energy ratio between each speech source and the noise:

$$\mathrm{SNR} = \frac{1}{6L} \sum_{i=1}^{6} \sum_{l=0}^{L-1} \left(\frac{S_i^2(l)}{N^2(l)} \right), \tag{36}$$

where L is the number of samples. As seen from Fig. 5, GCC-PHAT yields better TDE performance than GCC-ML at higher SNR. This is because the experimental environment is reverberant and GCC-ML suffers significant performance degradation under reverberation.

Compared to GCC-ML, GCC-PHAT is more robust with respect to reverberation. However, the GCC-PHAT method neglects the noise effect and, hence, it begins to exhibit dramatic performance degradation as the SNR is decreased. Unlike GCC-PHAT, GCC-ML does not exhibit this phenomenon since it has *a priori* knowledge about the noise power spectra that can help the estimator to cope with distortion. ES-GCC achieves the best performance. This is because the ES-GCC method does not focus on the weighting function process of the GCC-based method and it directly

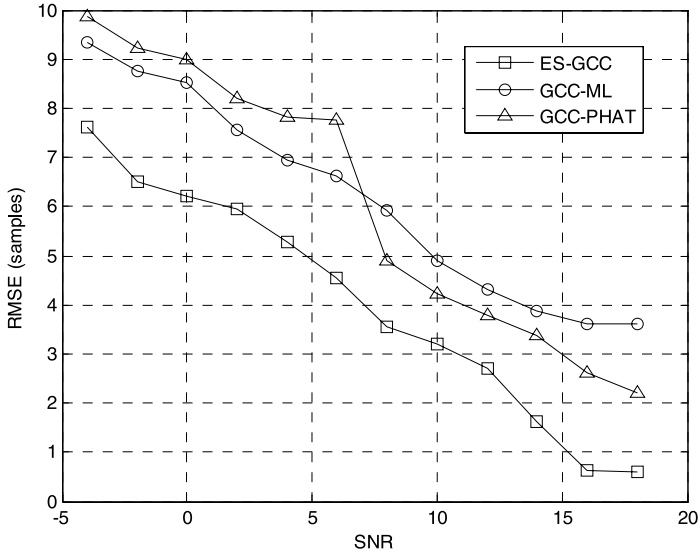

Figure 5. TDE RMSE results versus SNR.

takes the principal component vector as the microphone received signal for further signal processing.

Secondly, the direction estimation results of the proposed algorithm are compared to the wideband incoherent MUSIC algorithm [11] with arithmetic mean. Ten major frequencies, ranging from 0.1 to 3.4 kHz, were adopted for the MUSIC algorithm. The RMSE measurements of the sound source direction estimations are shown in Fig. 6. For fair comparison, the RMSE is calculated when the sound source number estimation is correct. Figure 6 shows that the MUSIC algorithm becomes worse as the sound source number is increased since the MUSIC algorithm is sensitive to coherent signals, especially when the environment has multiple sound sources and reverberant. The MUSIC algorithm assumes the sound source number is known. However, the sound source number is usually unknown in practical environments and the incorrect sound source number for the MUSIC algorithm would cause the worse performance. In our method, we use FastSLAM to solve the unknown data association.

4.2. Localization Accuracy (C2)

In the second case (C2) (depicted in Fig. 7), we combine ES-GCC and FastSLAM algorithms to estimate the locations of the sound sources and the robot. The robot (P3-DX) is moving through a specifically designed path to avoid some improper situations for DOA estimation and bearing-only SLAM. Let the origin be at 1500 mm from two sides of the wall (where the robot starts to move). Three loudspeakers at a height of 400 mm are installed at $(-900, 2380)$, $(1500, 2350)$ and $(3320, 625)$. Female and male voices are broadcast simultaneously through these speakers to simulate the sources. These speech signals contain silence periods so the number

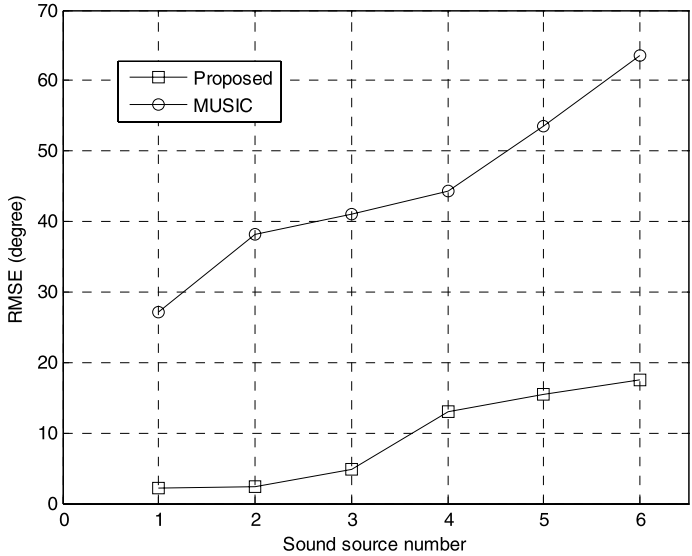

Figure 6. Sound source directions estimation result.

Figure 7. Spatial relation of the speaker, robot and laser range finder for C2.

of active sources varies in no particular order. The robot stops at 10 waypoints to record the acoustic data. The method in Section 2 is used to calculate the sound source arrival angles.

The path recording result is shown in Fig. 8, where the dark gray dots indicate the ground truths measured by the laser range finder. The path recorded by the mobile platform (plotted in triangles) is considered as the input of the particle filter. There will be a biasing error between the encoder data and the real ground truth. The light grey dots are the position estimations of the robot performed by FastSLAM. The estimated path is more likely to follow the path of the encoder data, since it was considered as the real input of the filter. The clustering result of the light grey dots is because the robot stopped at these points to perform the method in Section 2. It stopped for around 5 s to ensure the calculation of the sound emitting angle is stable. Also, the filter will perform only the predict phase while it is moving between the clusters. The update phase is performed at the waypoints.

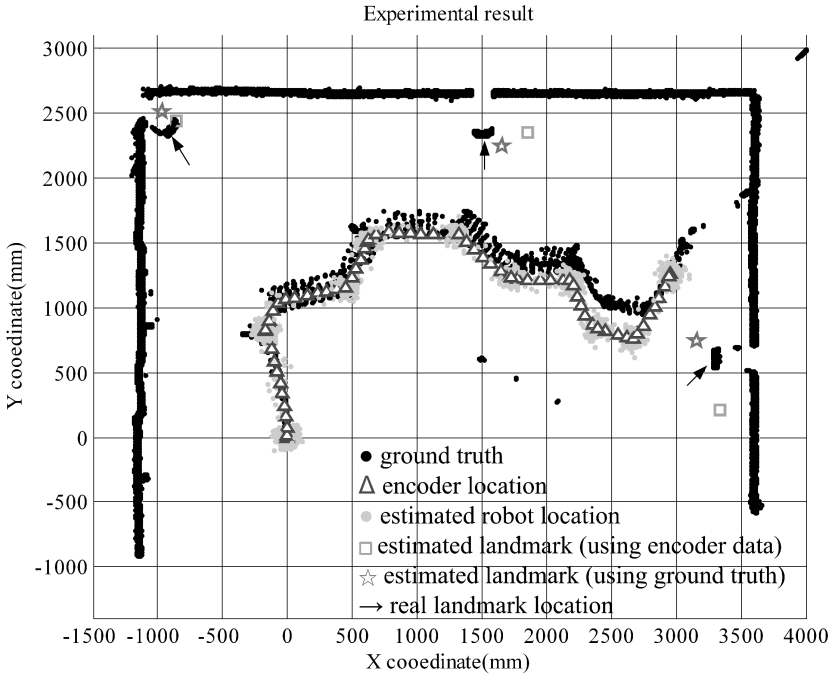

Figure 8. Experimental result of FastSLAM.

Table 1.
Localization result of the sound sources

Source	Laser range data (mm)	EKF estimates			
		Using ground truth path		Using encoder data	
		Estimated location (mm)	Distance error (mm)	Estimated location (mm)	Distance error (mm)
1	(−900, 2380)	(−854.2, 2443)	77.89	(−969, 2511.5)	148.5
2	(1500, 2350)	(1851, 2352.6)	351.01	(1662, 2242.9)	194.2
3	(3320, 625)	(3336.7, 214.3)	411.04	(3157.3, 741.7)	200.23

Another important effect of FastSLAM is that it simultaneously estimates the locations of sound sources using the EKF. The squares in Fig. 8 are the estimated mean of the three sound landmarks using encoder data and the stars are the estimations using the true path. The ground truths of the sound source are pointed by the black arrows. Figure 8 shows that the encoder data may cause a bias error on the estimations of the landmarks. Table 1 shows the estimated distances between sound sources and the laser range finder, and compares them with the mean distances computed from the laser range finder's data.

A very important feature of FastSLAM is that it will filter out unreasonable data in the resample state. Once the data (particle) is associated with the wrong landmark index, the importance factor of that particle will be diminished and cause particle elimination. Note that there has not been an algorithm that estimates the locations of the sound sources and the robot using only a microphone array on the robot under the multiple sound sources case. The microphone array on the robot in Ref. [4] is only utilized for estimating the directions of the sound sources and the localization is mainly done by the room microphone array. In this paper, we try to solve the multiple sound source localization problem by using the DOA estimation with a microphone array on the robot and the moving information of the robot.

5. Conclusions

This work proposes a method that is able to simultaneously localize a mobile robot and unknown number of multiple sound sources in the environment. The method is based on a combinational algorithm of DOA estimation and bearing-only SLAM. The DOA estimation using delay information is able to estimate the speed of sound as well as the far-field source direction. While the emitting angles are estimated, they are considered as the observation of a particle filter. The FastSLAM algorithm is used to solve the bearing-only SLAM problem for unknown data association. Experimental results show the effectiveness of the proposed method.

References

1. J. M. Valin, F. Michaud, J. Rouat and D. Létourneau, Robust sound source localization using a microphone array on a mobile robot, in: *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Las Vegas, NV, vol. 2, pp. 1228–1233 (2003).
2. J. M. Valin, J. Rouat and F. Michaud, Enhanced robot audition based on microphone array source separation with post-filter, in: *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Sendai, vol. 3, pp. 2123–2128 (2004).
3. S. Lauria, G. Bugmann, T. Kyriacou and E. Klein, Mobile robot programming using natural language, *Robotics Autonomous Syst.* **38**, 171–181 (2002).
4. K. Nakadai, K. I. Hidai, H. G. Okuno and H. Kitano, Real-time multiple speaker tracking by multi-modal integration for mobile robots, in: *Proc. 7th Eur. Conf. on Speech Communication and Technology*, Aalborg, pp. 1193–1196 (2001).
5. J. Hörnstein, M. Lopes, J. Santos-Victor and F. Lacerda, Sound localization for humanoid robots — building audio-motor maps based on the HRTF, in: *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Beijing, pp. 1170–1176 (2006).
6. M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, Berlin (2001).
7. J. M. Valin, F. Michaud and J. Rouat, Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering, *Robotics Autonomous Syst.* **55**, 216–228 (2007).
8. K. E. Bekris, M. Glick and L. E. Kavraki, Evaluation of algorithms for bearing-only SLAM, in: *Proc. IEEE Int. Conf. on Robotics and Automation*, Orlando, FL, pp. 1937–1943 (2006).

9. C. H. Knapp and G. C. Carter, The generalized correlation method for estimation of time delay, *IEEE Trans. Acoust. Speech Signal Process.* **24**, 320–327 (1976).
10. Q. H. Wang, T. Ivanov and P. Aarabi, Acoustic robot navigation using distributed microphone arrays, *Inform. Fusion* **5**, 131–140 (2004).
11. M. Wax, T. Shan and T. Kailath, Spatio-temporal spectral analysis by eigenstructure methods, *IEEE Trans. Acoust. Speech Signal Process.* **32**, 817–827 (1984).
12. H. Isao, A. Futoshi, A. Hideki, O. Jun, I. Naoyuki, K. Yoshihiro, K. Fumio, H. Hirohisa and Y. Kiyoshi, Robust speech interface based on audio and video information fusion for humanoid HRP-2, in: *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Sendai, pp. 2402–2410 (2004).
13. M. Walworth and A. Mahajan, 3D position sensing using the difference in the time-of-flights from a wave source to various receivers, in: *Proc. IEEE Int. Conf. on Advanced Robotics*, Seoul, pp. 91–94 (2001).
14. K. Yao, R. E. Hudson, C. W. Reed, D. Chen and F. Lorenzelli, Blind beamforming on a randomly distributed sensor array system, *IEEE J. Select. Areas Commun.* **16**, 1555–1567 (1998).
15. C. K. Wang, Multiple sound source direction estimation and sound source number estimation, *Master thesi*s, National Chiao-Tung University (2008).
16. S. Thrun, W. Burgard and D. Fox, *Probabilistic Robotics*. MIT Press, Cambridge, MA (2005).
17. M. Montemerlo, S. Thrun, D. Koller and B. Wegbreit, FastSLAM: a factored solution to the simultaneous localization and mapping problem, in: *Proc. AAAI Nat. Conf. on Artificial Intelligence*, Edmonton, pp. 593–598 (2002).

About the Authors

Jwu-Sheng Hu received the BS degree from the Department of Mechanical Engineering, National Taiwan University, Taiwan, in 1984, and the MS and PhD degrees from the Department of Mechanical Engineering, University of California at Berkeley, in 1988 and 1990, respectively. He is currently a Professor in the Department of Electrical and Control Engineering, National Chiao-Tung University, Taiwan, ROC. His current research interests include microphone array signal processing, active noise control, intelligent mobile robots, embedded systems and applications.

Chen-Yu Chan received the BS degree in Electrical Engineering and Computer Science Honors Program, and the MS degree in Electrical and Control Engineering from National Chiao-Tung University, Taiwan, in 2007 and 2009, respectively. His current research interests include simultaneous localization and mapping, microphone array signal processing, and intelligent mobile robots.

Cheng-Kang Wang received the BS and the MS degrees in Electrical and Control Engineering from National Chiao-Tung University, Taiwan, in 2006 and 2008, respectively.

Ming-Tang Lee received the BS and MS degrees in Electrical and Control Engineering from National Chiao-Tung University, Taiwan, in 2007 and 2008, respectively. He is currently a PhD candidate in the Department of Electrical and Control Engineering at National Chiao-Tung University. His research interests include sound source localization, speech enhancement, microphone array signal processing and adaptive signal processing.

Ching-Yi Kuo received the MS and PhD degrees in Industrial Engineering and Engineering Management from National Tsing Hua University, Hsinchu, Taiwan, in 2000 and 2008, respectively. She is currently a Researcher of the Robotics System Integration Technology Department, Mechanical and Systems Research Laboratories in the Industrial Technology Research Institute (ITRI). Her research interests include soft computing, applications of artificial intelligence/machine intelligence and robotics.