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

Jwu-Sheng Hu, Chen-Yu Chan, Cheng-Kang Wang

Department of Electrical and Control Engineering
National Chiao-Tung University
Hsinchu, Taiwan, R.O.C.
jshu@cn.nctu.edu.tw, cychan.ece96g@g2.nctu.edu.tw,
papa.ece91@nctu.edu.tw

Chieh-Chih Wang

Department of Computer Science and Graduate
Institute of Networking and Multimedia
National Taiwan University
Taipei, Taiwan, R.O.C.
bobwang@ntu.edu.tw

Abstract- **Sound source localization is an important function in robot audition. The 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. A method to compute the far field source directions as well as the speed of sound using the estimated time delays 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 bearings-only SLAM procedure. As the source signals are not persistent and there is no identification of the signal content, data association is unknown which is solved using FastSLAM. The experimental results demonstrate the effectiveness of the proposed approaches.**

I. INTRODUCTION

Audition system is a very important feature for intelligent robot. The fundamental requirement of this system is allowing a robot to interact with humans through speech dialog. Under this requirement there have been several research issues currently active in the Robotics community. These issues include speaker localization [1][5], speech separation and enhancement [2], speech recognition and natural dialog [3], and speaker identification and multi-modal interaction [4] etc. Among them, speaker localization using either biological hearing principle [5] or microphone array [1] has drawn lots of attentions for many years [6].

The underlying principle to localize sound source using microphone array is based on 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 comparing with the distance to the source, it is unlikely to obtain accurate distance information [7]. Hence, most of the sound source localization research on mobile robot emphasized on detecting the source directions. Almost no 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]. But 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 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 [4][12], exploring the technological boundary of localization using sound measurement alone is still very important. For example, occlusion or under a sudden lighting variation could make visual recognition fail easily.

The first challenge of sound source localization is the robustness of source detection, especially under multi-source environment with reverberation. Generalized cross correlation (GCC) [9] are one of the major methods discussed for robot localization application [10]. For multiple sources, MUSIC [11] is the most popular methods for eliminating the coherence problem and it is also 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] given a simple solution for the linear equation in [13] based on the far field assumption. Yao et. al. [14] presented a source linear equation similar with [13] to estimate the source location and velocity by using least square 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 [14] is modified for direction estimate of far field sources. The distinct advantage of the method is no information about the number of sources and speed of sound are 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 are served 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 sequence 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 8-channel microphone array on a mobile robot. It is shown that the overall system effectively localize the robot and sound sources in a room environment.

II. SOUND SOURCE DIRECTION ESTIMATION

In this section, a method of estimating directions of multiple unknown sound sources using microphone array is introduced []. 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 of the number of sources in the measurement.

A. Near Field Influence Factor and Field Distance Ratio

The work in [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 comparing with the source distance in most cases. As a result, it is necessary to consider the far field scenario. Let source location be $\boldsymbol{r}_s = [x_s \quad y_s \quad z_s]$, the *i-th* sensor locations $\boldsymbol{r}_i$ and the relative time delays, $t_i - t_j$, between the *i*-th sensor and *j*-th sensor. The original equation of the delay relation (from (15) of [14]) is,

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

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

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

$\hat{\boldsymbol{r}}_s$ 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, $\rho_i << 1$. Substituting (2) to (1), we have,

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

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

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

Equation (3) can be re-written as,

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

where

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

It is straightforward to see that $f_i \geq 0$ since

$$d_i \leq |\boldsymbol{r}_i-\boldsymbol{r}_0| \tag{7}$$

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

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

Equation (8) can also be derived from plane wave propagation perspective [1]. But 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 the *near field influence factor* for their roles in the source localization using array of sensors.

B. Least Square Solutions

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

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

where

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

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

and $\boldsymbol{b}=[t_1-t_0 \quad t_2-t_0 \quad \cdots \quad t_{M-1}-t_0]^T$ (12)

It is therefore easy to estimate the speed of sound:

$$v=\frac{1}{|\boldsymbol{w}_s|}=\frac{1}{\left|(\mathbf{A}_s{}^{\mathrm{T}}\mathbf{A}_s)^{-1}\mathbf{A}_s{}^{\mathrm{T}}\boldsymbol{b}\right|} \tag{13}$$

And the sound source direction can be given by:

$$\hat{\boldsymbol{r}}_s=\frac{\boldsymbol{w}_s}{|\boldsymbol{w}_s|}=\frac{(\mathbf{A}_s{}^{\mathrm{T}}\mathbf{A}_s)^{-1}\mathbf{A}_s{}^{\mathrm{T}}\boldsymbol{b}}{\left|(\mathbf{A}_s{}^{\mathrm{T}}\mathbf{A}_s)^{-1}\mathbf{A}_s{}^{\mathrm{T}}\boldsymbol{b}\right|} \tag{14}$$

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

$$\hat{\boldsymbol{r}}_s=[\cos\theta_S\sin\phi_S \quad \sin\theta_S\sin\phi_S \quad \cos\phi_S] \tag{15}$$

where θ_S and ϕ_S are azimuth and elevation angle respectively. It is straightforward to verify that $\mathbf{A}_s$ reduces rank if the vectors constructed by sensor pairs do not span the 3-D space (i.e., planar array), meaning the delay relation is satisfied by more than one source directions. Secondly, (8) is actually an approximation by considering plane wave propagation. Please refer to [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 certain accuracy. For multiple sources, the estimation becomes more difficult as the signals are mixed together in the measurements. In next section, an eigenstructure-based generalized cross-correlation (GCC) method is presented to cope with this issue.

C. Delay Estimation of Multiple Sources

Consider an array with M microphones on a mobile robot. The received signal of the *m*-th microphone which contains d sources can be described by SFT (Short-term Fourier Transform) as:

$$X_m(\omega_f,k)=\sum_{p=1}^{d}a_{mp}S_p(\omega_f,k)e^{-j\omega_l\tau_{mp}}+N_m(\omega_f,k) \qquad (16)$$

$$f=1,2,\ldots,F$$

where a_{mp} is the amplitude from the *p*-th sound source to the *m*-th 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 (1) in matrix form:

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

where

$$\boldsymbol{X}^{\mathrm{T}}(\omega_f,k)=\left[X_1(\omega_f,k),\ \cdots,\ X_M(\omega_f,k)\right]$$

$$\boldsymbol{N}^{\mathrm{T}}(\omega_f,k)=\left[N_1(\omega_f,k),\ \cdots,\ N_M(\omega_f,k)\right]$$

$$\boldsymbol{S}^{T}(\omega_f,k)=\left[S_1(\omega_f,k),\ \cdots,\ S_d(\omega_f,k)\right]$$

$$\mathbf{A}(\omega_f)=\begin{bmatrix} a_{11}e^{-j\omega_f\tau_{11}} & \cdots & a_{1d}e^{-j\omega_f\tau_{1d}} \\ \vdots & & \vdots \\ a_{M1}e^{-j\omega_f\tau_{M1}} & \cdots & a_{Md}e^{-j\omega_f\tau_{Md}} \end{bmatrix}$$

The received signal correlation matrix with eigenvalue decomposition can be described as:

$$\begin{aligned}\mathbf{R}_{xx}(\omega_f)&=\frac{1}{N}\sum_{k=1}^{N}\boldsymbol{X}(\omega_f,k)\boldsymbol{X}^{\mathrm{H}}(\omega_f,k)\\ &=\sum_{i=1}^{M}\lambda_i(\omega_f)\boldsymbol{V}_i(\omega_f)\boldsymbol{V}_i^{H}(\omega_f)\end{aligned} \qquad (18)$$

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

$$\boldsymbol{V}_1(\omega_f)=\left[V_{11}(\omega_f)\quad V_{12}(\omega_f)\quad \cdots\quad V_{1M}(\omega_f)\right] \qquad (19)$$

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:

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

The *f*-th 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 *i*-th and *j*-th microphone can be represented as:

$$R_{x_ix_j}(\tau)=\int_{\omega_1}^{\omega_F}V_{1i}(\omega)V_{1j}(\omega)e^{j\omega\tau}d\omega \qquad (21)$$

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

$$\hat{\tau}_{ES-GCC}=\arg\max_{\tau}R_{x_ix_j}(\tau) \qquad (22)$$

D. 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 SFT frame. The question is how to combine these delays among microphone pairs to form the vector $\boldsymbol{b}$ of (12). Denote τ_{jk} as the *k*-th delay of the microphone pair (*j*, 0), k = 1~q_j where q_j is the total number of delays (peaks) of this pair. Note that q_j maybe different for different pairs (depending on the threshold level of the peak value). For M microphones, there will be (q_1×q_2···×q_{M-1}) number of possible combinations of the vector $\boldsymbol{b}$. However, since the minimum number of microphone pairs to solve (9) is 3, we can sort out the combination by starting from 3 pairs and iteratively adding additional pair. Without loss of generality, assume the indices of microphone pairs are arranged in the order such that $q_1\geq q_2\geq q_3\geq q_4\cdots\geq q_{M-1}$. 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/sec). Specifically, a set of possible sound sources can be found as the following:

$$S=\left\{(l,m,n)\middle|\left|e_{lmn}\right|\leq\overline{e}\ \text{for}\ 1\leq m\leq q_1,1\leq m\leq q_2,1\leq n\leq q_3\right\} \qquad (23)$$

where

$$e_{lmn}=\frac{1}{\left|\left(\mathbf{A}_3{}^{\mathrm{T}}\mathbf{A}_3\right)^{\mathbf{-1}}\mathbf{A}_3{}^{\mathrm{T}}\boldsymbol{b}_{lmn}\right|}-\overline{v} \qquad (24)$$

$$\mathbf{A_i}=\begin{bmatrix} x_1-x_0 & y_1-y_0 & z_1-z_0 \\ \vdots & \vdots & \vdots \\ x_i-x_0 & y_i-y_0 & z_i-z_0 \end{bmatrix} \qquad (25)$$

$\boldsymbol{b}_{1mn}=\left[\tau_{1l}\quad \tau_{2m}\quad \tau_{3n}\right]^{T}$ and $\overline{v}$ the nominal speed of sound. Note that the error bound $\overline{e}$ 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 which 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 of microphone pair 4 can be added similarly until the pair M-1. 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 $\bar{e} = 15m/s$ [15].

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

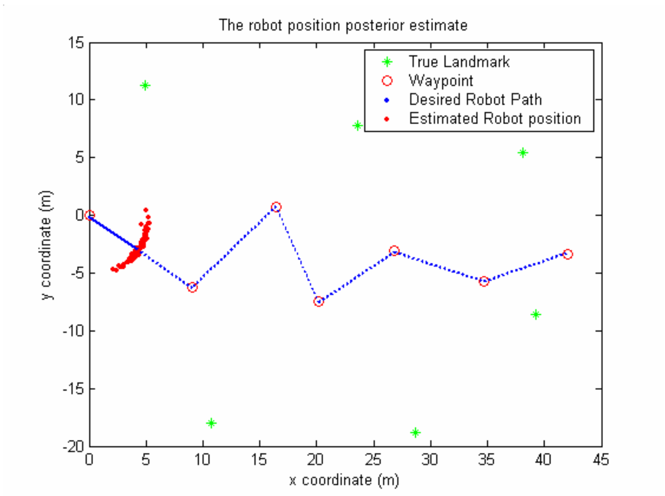

Fig. 1. The robot position posterior estimation

III. Localization of Sources and Robot

Simultaneous Localization and Mapping (SLAM) problem is the procedure of recognizing a set of feature landmarks $(\mu_1, \Sigma_1, \mu_2, \Sigma_2, \cdots, \mu_d, \Sigma_d)$ and localizing the sensor odometer $X = (x_r, y_r, \theta_r)$ with respect to the landmark set. A microphone array platform carried by a two-wheel robot was used in this paper to perform the localization of the robot and landmarks. According to section II, the microphone array is capable of recognizing unknown number of sound source as the feature points and obtaining associated angle of arrival. The angles are considered as the bearing measurements and this becomes a standard bearing-only SLAM problem [8]. Since the localization problem of this paper is no different from others, the FastSLAM algorithm [16] is adopted here. FastSLAM estimate the robot path using a particle filter and the map feature locations are estimated using EKF. Each particle possesses its own set of EKFs for all feature point. Particles in FastSLAM are denoted as

$$Y_t^{[k]} = (X_t^{[k]}, \mu_{1,t}^{[k]}, \Sigma_{1,t}^{[k]}, \mu_{2,t}^{[k]}, \Sigma_{2,t}^{[k]}, \cdots, \mu_{d,t}^{[k]}, \Sigma_{d,t}^{[k]}) \tag{26}$$

where $[k]$ is the index of the particle; $X_{r,t}^{[k]}$ is the path estimate of the robot, and $\mu_{p,t}^{[k]}$ and $\Sigma_{p,t}^{[k]}$ are the mean and covariance of the Gaussian distribution indicating the p-th landmark location. The algorithm can be separated into the following three steps:

A. Sampling New Pose according to path posterior

For each particle at time t, the control input u_t is used to estimate the $Y_t^{[k]}$ from $Y_{t-1}^{[k]}$. It samples the new robot position $X_t^{[k]}$ according to the posterior,

$$X_t^{[k]} \sim p(X_t^{[k]} \mid X_{t-1}^{[k]}, u_t) \tag{27}$$

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

B. 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

$$p(m \mid X_{1:t}, Z_{1:t}) = \eta\, p(Z_t \mid X_t, m)\, p(m \mid X_{1:t-1}, Z_{1:t-1}) \tag{28}$$

where m is the landmark and $Z_{i:j}$ is the observation from time step i to j. The probability distribution of landmarks $p(m \mid X_{1:t-1}, Z_{1:t-1})$ at time t-1 is represented by a Gaussian distribution with mean $\mu_{p,t-1}^{[k]}$ and covariance $\Sigma_{p,t-1}^{[k]}$. For the new estimation, FastSLAM linearizes the perceptual model $p(Z_t \mid X_t, m)$ in the same way as EKF. The measurement function h could be approximated by a Taylor expansion:

$$\begin{aligned} h(m, X_t^{[k]}) &\approx h(\mu_{t-1}^{[k]}, X_t^{[k]}) + h'(X_t^{[k]}, \mu_{t-1}^{[k]})(m - \mu_{t-1}^{[k]}) \\ &= \hat{Z}_t^{[k]} + H_t^{[k]}(m - \mu_{t-1}^{[k]}) \end{aligned} \tag{29}$$

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

$$K_t^{[k]} = \Sigma_{t-1}^{[k]} H_t^{[k]} (H_t^{[k]T} \Sigma_{t-1}^{[k]} H_t^{[k]} + R_t)^{-1} \tag{30}$$

$$\mu_t^{[k]} = \mu_{t-1}^{[k]} + K_t^{[k]}(Z_t - \hat{Z}_t^{[k]}) \tag{31}$$

$$\Sigma_t^{[k]} = (I - K_t^{[k]} H_t^{[k]T}) \Sigma_{t-1}^{[k]} \tag{32}$$

After repeating step *A*. and *B*. *M* times, the temporary set of *M* particles is created.

C. Resampling

In the final step, FastSLAM resample the set of the *M* particles. First we'll calculate the importance factor of each particle. The factor is given by

$$w_t^{[k]} \approx \eta \left| 2\pi Q_t^{[k]} \right|^{-\frac{1}{2}} e^{\{-\frac{1}{2}(z_t - \hat{z}_t^{[k]})^T Q_t^{[k]-1} (z_t - \hat{z}_t^{[k]})\}}$$

with the covariance

$$Q_t^{[k]} = H_t^{[k]T} \Sigma_{p,t-1}^{[k]} H_t^{[k]} + R_t$$

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 these weighting.

One of the key features of FastSLAM is that as long as s mall subset if the particles are 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 position, which increase the probability that will be sampled away during the resample phase of the algorithm.

IV. Experimental Results

An 8-channel microphone array is constructed using specially-made digital microphones. The digital microphone integrates an electret condenser microphone cores, analog output amplifier, and sigma-delta modulator on a single chip [15]. The digital bit-stream transmission achieves a minimum interference comparing with conventional analog microphone signals. The microphone array topology and the mobile robot for the experiment is shown in Fig. 2. Note that it is a 3-dimensional microphone array which is able to estimate the sound source elevation angle. In this experiment, however, this angle is ignored since the localization concerns with 2-D locations of the robot and the sound sources.

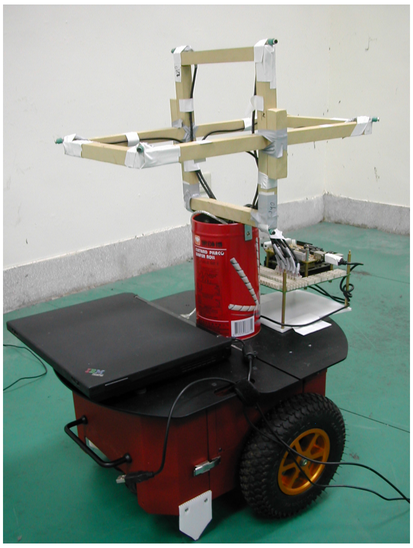

Fig. 2. Digital microphone array mounted on the robot

The room size is 4750 mm × 3600 mm and height of 3600 mm approximately. Let the origin be at 1500 mm from two sides of the wall (where the robot starts to move). Three loudspeakers at the height of 400 mm are installed at (-900,2380), (1500,2350) and (3320,625). Female and male voices are broadcasted simultaneously through these speakers to simulate the sources. These speech signals contain silence periods so the number of active sources varies in no particular order. The microphone array sampling rate is 16 KHz and each SFT frame contains 512 samples with overlapping 256 samples. The sound arrival angles are computed after collecting 20 frames which is about 3 times per second. The robot stops at 10 waypoints to record the acoustic data. Method in section II is used to calculate the sound sources arrival angles. Without knowing the data association of each calculated angle, the FastSLAM algorithm of unknown data association in section III to estimate the locations of robot and sound sources.

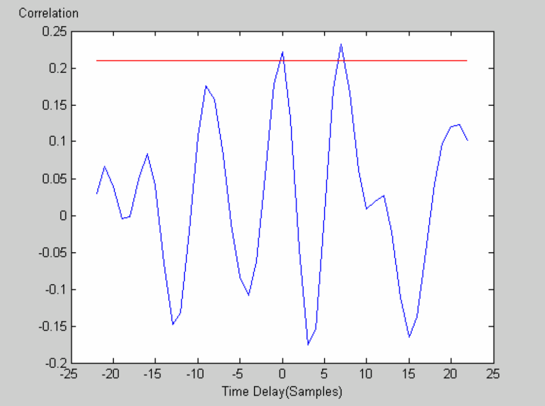

Fig. 3 A typical eigen-structure based GCC values

Fig. 3 shows a typical result of the eigen-structure based GCC value (Eq.(21)) from a pair of microphones as a function of delay. The graph exhibits several peaks which indicate existence of multiple sources. The threshold for valid delays in this paper is set as the 90% level of the largest peak (e.g., the red line in Fig. 3). Different microphone pairs may result in different number of possible sources. The procedure of section II.D is implemented to sources with unreasonable estimates of sound speed. For this experiment, the range of sound speed is set as 300 m/s to 400 m/s. This range will give more sound source candidates to test if the particle filter is able to eliminate spurious sources.

The path recording result is shown in Fig 4, where the blue dots stand for the ground truth measured by the laser range finder. The path recorded by the mobile platform (plotted in red) 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 yellow dot is the position estimation 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 yellow dots is because of the robot will stop at these points to perform the method in section II. It'll stop for around 5 second 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 waypoint.

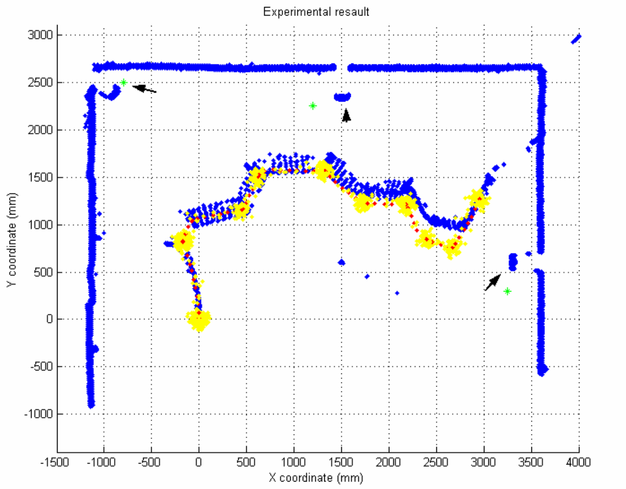

Fig. 4. Experimental result of the FastSLAM

Another important effect of FastSLAM is that it simultaneously estimates the sound source location using EKF. The green stars in Fig 3 are the estimated mean of the three sound landmarks. And the ground truth of the sound source is pointed by the black arrows. Although the path estimation contains biasing due to the encoder error, the landmark estimation is reasonable. Table 1 shows the sound source locations estimates from FastSLAM and the mean locations computed from the laser range finder's data.

Table 1 The localization result of the sound source

	Laser Range Data (mm)	EKF estimate (mm)	Distance error (mm)
Source 1	2980.0	2993.3	13.3 (0.45%)
Source 2	1747.0	1677.1	-69.9 (-4%)
Source 3	1820.0	1775.5	-44.5 (-2.4%)

Table 2 The bearing result of the sound source

	Laser Bearing Data (°)	EKF estimate (°)	Distance error (°)
Source 1	143.4	140.4	-3
Source 2	90.5	100.3	9.8
Source 3	0.0	-9.7	-9.7

A very important benefit 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 shrink down and cause particle elimination. So, the FastSLAM algorithm is robust to unknown data association.

V. CONCLUSION

This work estimate unknown number of sound sources using eigenstructure-based generalized cross correlation. And it is also 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 able to solve the bearing-only SLAM problem for unknown data association.

ACKNOWLEDGMENT

This work was supported in part by the National Science Council of Taiwan, ROC under grant NSC 95-2218-E-009 - 008 and the Ministry of Economic Affairs under grant 95-EC-17-A-04-S1-054.

REFERENCES

[1] Valin, J. M., Michaud, F., Rouat, J., and Létourneau, D., "Robust sound source localization using a microphone array on a mobile robot," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1228-1233, 2003.

[2] Valin, J.-M., Rouat, J., and Michaud, F., "Enhanced robot audition based on microphone array source separation with post-filter," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol.3, pp. 2123- 2128, 2004.

[3] Lauria, Stanislao, Bugmann,,Guido, Kyriacoua, Theocharis and Kleinb, Ewan , "Mobile robot programming using natural language," *Robotics and Autonomous Systems*, Volume 38, Issues 3-4, 31 March 2002, Pages 171-181

[4] Nakadai, Kazuhiro, Hidai, Ken-ichi, Okuno, Hiroshi G., Kitano, Hiroaki, "Real-Time Multiple Speaker Tracking by Multi-Modal Integration for Mobile Robots," *7th European Conference on Speech Communication and Technology*, Aalborg, Denmark, September 3-7, 2001

[5] Hörnstein, Jonas, Lopes, Manuel, Santos-Victor, José, and Lacerda, Francisco, "Sound localization for humanoid robots – building audio-motor maps based on the HRTF," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1170-1176, Oct., 2006.

[6] Brandstein, Michael and Ward, Darren, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, June 15, 2001.

[7] Valin, Jean-Marc, Michaudb, Francois, and Rouatb, Jean, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," Robotics and Autonomous Systems 55 (2007) 216–228.

[8] Bekris, K.E., Glick, M., and Kavraki, L.E., "Evaluation of algorithms for bearing-only SLAM," *Proceedings 2006 IEEE International Conference on Robotics and Automation*, May 15-19, 2006, pp. 1937- 1943.

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

[10] Q. H. Wang, T. Ivanov, and P. Aarabi, "Acoustic robot navigation using distributed microphone arrays," *Information Fusion*, Information Fusion 5, vol. 5, pp. 131-140, June 2004

[11] Wax, M., Shan, T., and Kailath, T., "Spatio-Temporal spectral analysis by eigenstructure methods," *IEEE Transactions on Acoustic, Speech, Signal Processing*, vol. 32, pp. 817-827, Aug 1984.

[12] Isao H., Futoshi A., Hideki A., Jun O., Naoyuki I., Yoshihiro K., Fumio K., Hirohisa H., Kiyoshi Y., " Robust Speech Interface Based on Audio and Video Information Fusion for Humanoid HRP-2," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.2402-2410, 2004.

[13] Walworth, M., and Mahajan, A., "3D position sensing using the difference in the time-of-flights from a wave source to various receivers," *Proc. IEEE International Conference on Advanced Robotics*, pp.611-616, 1997.

[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*., vol. 16, pp. 1555-1567, Oct. 1998

[15] Wang, Cheng-Kang, "Multiple Sound Source Direction Estimation and Sound Source Number Estimation," *Master Thesis*, National Chiao-Tung University, Taiwan, 2008.

[16] S. Thrun, W. Burgard, D. Fox, 2005. Probabilistic Robotics [hardcover]. pp.444-449