2008 IEEE International Conference on
Robotics and Automation
Pasadena, CA, USA, May 19-23, 2008

Robust Estimation of Sound Direction for Robot Interface

Cheol-Taek Kim, Tae-Yong Choi, ByongSuk Choi, and Ju-Jang Lee, *Senior Member*, IEEE

***Abstract*—This paper proposes a robust method to estimate sound direction for mobile robot interface. In contrast to other methods, the proposed method estimates sound direction at intervals of short time to deal with deterioration by movement of sound source or robot. The two methods, discrete Kalman filter and time delay of arrival (TDOA) using generalized cross-correlation (GCC), are combined by quantifying reliability of TDOA. The combination prevents the deterioration by low power of signals, noise, and reverberation because discrete Kalman filter efficiently filters out such effects. The result of experiments, in which mobile robot estimates azimuth angle with three microphones, shows that the proposed method provides robustness on estimating sound direction even if the sound source or robot is moving during estimation procedure.**

I. INTRODUCTION

RECENTLY, robots have more commercial applications as it becomes more intelligent and human-like. Especially, human-like activity and interface is more important to domestic robot, and therefore human and robot interaction (HRI) has been widely researched enormously [1-6].

Among the researches, we think sound localization or sound direction estimation is very helpful for robot to interact with human. Because robots are designed to use multiple users and assumed to move automatically, the first interaction between human and robot will be the calling for attention (Fig. 1). After calling procedure, the sound direction estimation can be used to improve the speech recognition process or speaker identification as a means of spatial filter.

Three possible approaches are mainly used to estimate sound direction. 1) Maximizing the steered response power of a beamforming [6]. 2) Estimating high resolution spectral density [7]. 3) Using time delay of arrival (TDOA) [5]. The first method is very robust to noise and interference, the second is easy to interpret with probability density for multiple sound sources, and the third has low computational cost that means it can be operated in real time.

However, the state-of-the-art algorithms for estimating sound direction have not considered deteriorations by movement of sound source and robot seriously. If a mobile robot with differential drives is turning when a user calls the robot, then the movement will decrease the performance of sound direction estimation.

This work was supported by the Korea Foundation for International Cooperation of Science & Technology (KICOS) through a grant provided by the Korean Ministry of Science & Technology (MOST) in 2007. The authors are with Department of Computer Science and Electrical Engineering, Korea Advanced Institute of Science & Technology, Daejeon 305-701, Korea (e-mail:ctkim@odyssey.kaist.ac.kr)

This paper proposes a robust method to estimate sound direction for mobile robot interface. A mobile robot equipped with three microphones estimates direction of arrival (DOA) of sound at intervals of short time to deal with deterioration by movement of sound source or the robot.

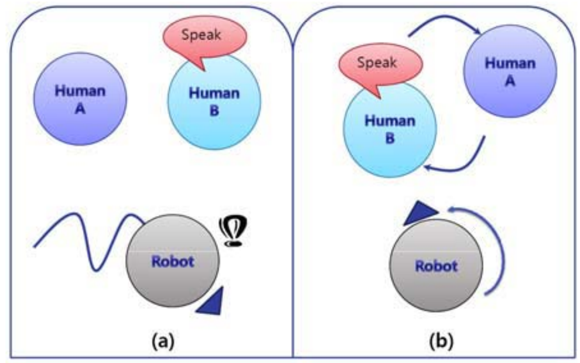

Fig. 1 (a) The calling procedure for attention of robot (b) After calling procedure, continuous tracking of speaker can improve the performance of speech processing.

As a result of estimating DOA of sound at intervals of short time, some unreliable estimation can be occurred. The main causes are low power of signals, noise, and reverberation. Practically, the unreliable DOA estimation may produce unexpected behaviors and it is the motivation of robust methods such as averaging multiple results [5] and rejecting after outlier detection [11].

We get hint from rejecting after outlier detection to address the robustness problem. Two values, the maximum peak and ratio of the values of the first and second largest peak in the GCC function, can be selectively used to detect outliers. Based on the values, we suggest the reliability network to quantify the confidence of result of TDOA.

The grid based method determines a most probable DOA using three GCC estimation results, and finally Kalman filter predicts DOA using the quantified reliability and the estimation result from the grid method. By modeling the measurement noise using quantified reliability, the Kalman filter prevents the deterioration by estimating at intervals of short time.

The remainder of this paper is organized as follows. Section II reviews the principle of the DOA estimation between two microphones and discrete Kalman filter to give backgrounds for the paper. Section III provides the description of the proposed method. Section IV reports the experimental results of the proposed method which shows the robust characteristics in the estimation of the DOA. Finally, the concluding remark is given in section V.

978-1-4244-1647-9/08/$25.00 ©2008 IEEE.

II. BACKGROUND

This section describes two techniques to provide the backgrounds of the paper. The first is estimating DOA of sound using two spatially separated microphones. The second is discrete Kalman filter which are widely used to estimate unknown state with noisy data. More detailed descriptions can be found in the literatures [5], [12].

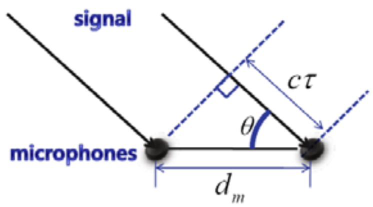

Fig. 2 The relationship between DOA θ and time lag τ: The DOA θ can be directly calculated by the time lag τ.

A. DOA Estimation

Fig. 2 describes the DOA estimation of sound with two spatially separated microphones. In the case, TDOA method (estimating time delay between two microphones), which has very long history in signal processing, is generally used and the DOA of sound is estimated by directly calculating following trigonometrical function

$$\theta = \cos^{-1}\left(\frac{c\tau}{d_m}\right), \tag{1}$$

where c is the sound velocity and d_m is the distance between microphones [5], [8], [9].

The cross-correlation [8] between two signals is mainly used to determine the estimated time delay

$$R_{s_1 s_2}(\tau) = E[s_1(t)s_2(t-\tau)], \tag{2}$$

where $s_1(t)$ and $s_2(t)$ are signals emanating from same remote source, and E denotes expectation.

Considering the realistic finite observation time, the cross correlation can be estimated by

$$\hat{R}_{s_1 s_2}(\tau) = \frac{1}{T-\tau}\int_{\tau}^{T} s_1(t)s_2(t-\tau)dt, \tag{3}$$

where T represents the observation interval. Generally, in order to improve the accuracy of the delay estimate $\hat{\tau}$, it is desirable to prefilter $s_1(t)$ and $s_2(t)$ prior to the integration in (3). This generalized method is called as generalized cross-correlation (GCC) and defined as follows

$$\hat{R}^{(g)}_{s_1 s_2}(\tau) = \int_{-\infty}^{\infty} \psi_g(f)\hat{G}_{s_1 s_2}(f)e^{j2\pi f\tau}df, \tag{4}$$

where, $\hat{G}_{s_1 s_2}(f)$ is the estimated cross power spectral density and $\psi_g(f)\hat{G}_{s_1 s_2}(f)$ is the cross power spectrum between the filtered signals.

In the literature, GCC with PHAT weight filter provides the best performance in the noisy and reverberant environment [9].

$$\psi_g(f) = \frac{1}{\left|\hat{G}_{s_1 s_2}(f)\right|} \tag{5}$$

Therefore, the DOA can be calculated by (1) with time delay $\hat{\tau}$ with the maximum peak in the GCC-PHAT function

$$\hat{\tau} = \arg\max_{\tau} R^{(g)}_{s_1 s_2}(\tau), \tag{6}$$

where $\psi_g(f)$ is given by (5).

B. Discrete Kalman Filter

The discrete Kalman filter is widely used to address the general problem of trying to estimate the state $x \in \Re^n$ with the measurement $z \in \Re^m$ using following relationship

$$\begin{aligned} x_k &= Ax_{k-1} + Bu_{k-1} + w_{k-1} \\ z_k &= Hx_k + v_k \end{aligned}, \tag{7}$$

where, w and v are the process noise and measurement noise respectively, and assumed to have normal probability distribution $N(0,Q)$ and $N(0,R)$ respectively. The matrix A specifies how the state evolves in a step, matrix B relates to the control input u to the state x, and matrix H relates the state to the measurement.

The discrete Kalman filter algorithm minimizes the effect of w and v in recursive manner and predicts the state using (7). The detailed procedure is very similar to recursive least square procedure as described in following paragraphs.

Let $\hat{x}_k^-$ denote *a priori* estimated state at step k using knowledge of the process prior to step k, and $\hat{x}_k$ denote *a posteriori* estimated state at step k given observation z_k. Then, *a priori* and *a posteriori* estimated error covariance can be defined as

$$\begin{aligned} P_k^- &= E[e_k^- e_k^{-T}], \\ P_k &= E[e_k e_k^T], \end{aligned} \tag{8}$$

where $e_k^- \equiv x_k - \hat{x}_k^-$ and $e_k \equiv x_k - \hat{x}_k$.

The Kalman filter updates *a priori* state and error covariance based on previous state, input, and error covariance

$$\begin{aligned} \hat{x}_k^- &= A\hat{x}_{k-1} + Bu_{k-1} \\ P_k^- &= AP_{k-1}A^T + Q \end{aligned} \tag{9}$$

Based on *a priori* state and error covariance, the Kalman filter updates a Kalman gain K, *a posteriori* error covariance, and computes *a posteriori* state $\hat{x}_k$

$$K = P_k^- H^T (HP_k^- H^T + R)^{-1}$$
$$\hat{x}_k = \hat{x}_k^- + K(z_k - H\hat{x}_k^-) \tag{10}$$
$$P_k = (1 - KH)P_k^-$$

By (9) and (10), the Kalman filter accomplishes minimization of *a posteriori* error covariance, and *a posteriori* state $\hat{x}_k$ is the optimal prediction for given measurements.

Note that the matrix R in (10) is the measurement noise error covariance matrix and is used to compute Kalman gain K. The Kalman gain K determines the effect of the observation z_k in prediction of *a posteriori* state, and therefore the matrix R affects the rate of update based on the current observation z_k.

III. ROBUST SOUND DIRECTION ESTIMATION SYSTEM

Fig. 3 describes the proposed method that estimates DOA using three spatially separated microphones. Three signals are extracted during time interval Δ, and three TDOAs are estimated by GCC –PHAT function. Using three TDOAs, a DOA is estimated by grid based method, and discrete Kalman filter predicts DOA based on the grid based estimation and the quantified reliability. We divide whole procedure into three major steps. First step is called as 'Time Delay Estimation', second step is called as "Direction of Arrival Estimation', and third step is called as "Robust Direction of Arrival Estimation". Briefly, we call the steps as step 1, step 2, and step 3.

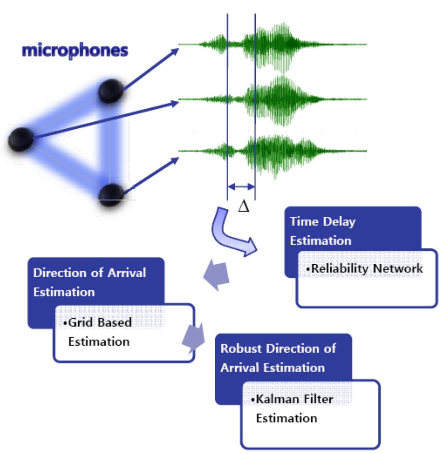

Fig. 3 The block diagram of the robust sound direction estimation system

A. Configuration of Hardware System

Before explaining three major steps, we describe the configuration of hardware system. Three microphones are used to estimate the azimuth angle for a single sound source. For geometrical advantage, the distance between each microphone pair is set to the identical value $d_m = 0.32m$, and therefore three microphones forms a regular triangle. In each microphone, the arrival sound is received in the 16bit raw data with sampling rate $f_s = 16kHz$. For each interval Δ, the estimation of time delay is processed when all signal powers exceed the threshold values.

B. Time Delay Estimation (Step 1)

The time delay estimation is basically calculated by the GCC-PHAT as described in section II. However, the method sometimes produces unreliable result when the DOA is estimated at intervals of short time. Therefore, the system should measure the confidence of the estimation from the GCC-PHAT. In the literatures, several measures have been suggested to detect outliers to suppress the effect from it [11].

We use two simple measures to quantify the reliability of GCC-PHAT result. First measure is the value of the maximum peak (m_P) in the GCC function. Second measure is the ratio (m_r) of the values of the first and second largest peak in the GCC function. These two measures have slightly different characteristics in the noisy and reverberant environment.

Let us define $\hat{\tau}$, τ^*, and e_{th} as the estimation of GCC-PHAT, real time delay, and error threshold value. Then, the success rate of GCC-PHAT is defined as $P(|\hat{\tau} - \tau^*| < e_{th})$. Consider a database that contains signals and real time delay τ^*. Using the database, new training data can be built by recoding the conditional success rate for given measure m_1 and m_2 (success rate for the data of which both m_P, and m_r exceed the given measures)

This paper suggests to use a single-hidden layer feedforward network to quantify reliability based on both m_P and m_r. The network is structured with ten sigmoid functional hidden nodes, each input neuron of it has input bias, and it has a linear output layer. The training data is newly built database that contains randomized given measures and corresponding conditional success rate. With the training data, the neural network is trained using variable projection method, which is modified Levenberg-Marquardt algorithm [13].

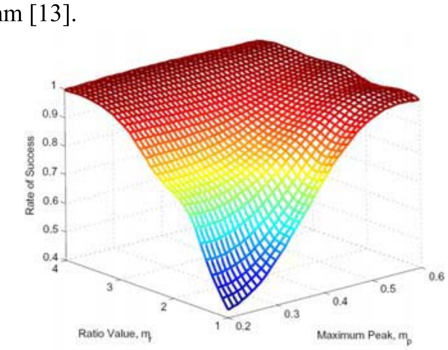

Fig. 4 The reliability estimation result with feedforward neural network.

An example of function approximation based on the trained neural network is shown in Fig. 4. This network is used to quantify reliability and named as 'reliability network'. If the m_P and m_r are both low, the success rate is under 0.5 which means the estimated time delay is unreliable. While, if the m_P and m_r are both high, the success rate goes to 1 which means the estimated time delay is reliable. In the time delay estimation step, for each microphone pair (three pairs), the estimated time delay($\hat{\tau}$) and reliability(ε) are computed. Using these variables, the DOA is estimated by geometrical interpretation.

C. *Direction of Arrival Estimation (Step 2)*

The DOA with two microphones can be calculated by (1) using time delay. However, the calculated DOA is not unique because it has always mirror DOA. Therefore, this paper suggests novel DOA estimation with three microphones.

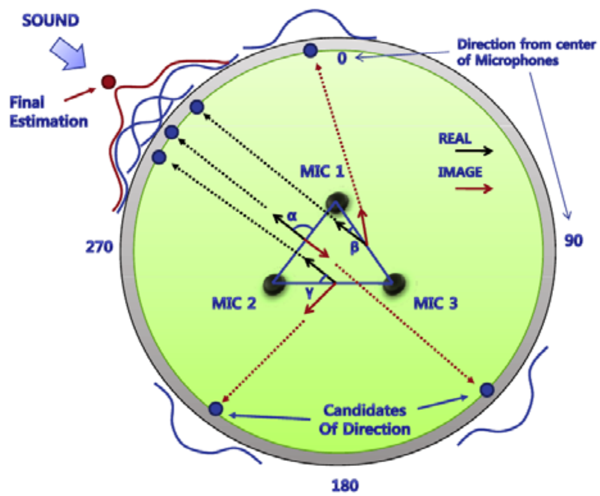

Fig. 5 direction of arrival estimation based on grid Method – stochastic approach

Fig. 5 depicts the proposed strategy of DOA estimation. The objective is to estimate the direction of sound from center of microphones with three time delay estimations. Let the calculated DOA between microphone 1-2 be α , the DOA between microphone 1-3 be β , and the DOA between microphone 2-3 be γ . Then, the six candidates of DOA from the center can be determined, because each pair of microphones has two directions of DOA.

$$\begin{aligned} \kappa_1 &= \alpha - 30, & \kappa_2 &= -\alpha - 30 \\ \kappa_3 &= \beta + 30, & \kappa_4 &= -\beta + 30 \\ \kappa_5 &= \gamma + 90, & \kappa_6 &= -\gamma + 90 \end{aligned} \tag{11}$$

As shown in Fig. 5, three candidates are located in real direction of sound. We suggest the estimation based on stochastic analysis because some estimations of time delay are unreliable. Let us assume the probability of DOA of i th candidate with the reliability ε_i has normal distribution with mean κ_i and standard deviation $\eta(1-\varepsilon_i)+\sigma$, where η is a user-defined penalty parameter for unreliability and σ represents the basic uncertainty of the system. Total 360 grids are generated to represent the estimated DOA with the resolution 1°. For the i th candidate, the probabilities of DOA in $k_i \pm 30$ grids are estimated and added to the grids. The final DOA is estimated by finding the grid with maximum probability value. The computation cost of this procedure is just 366 calculations of exponential function.

The reliability of the final DOA is just assessed by the mean value of the two largest reliability values dropping the lowest reliability. In the direction of arrival estimation step, the estimated DOA ($\hat{\theta}$) and reliability(ε) are computed. Using these data, the robust DOA estimation is performed.

D. *Robust Direction of Arrival Estimation (Step 3)*

The robust DOA estimation uses the Kalman filter with the quantified reliability to reduce the effect of unreliable DOA. Let us assume that a robot or a human moves in constant angular velocity in the form of the Kalman state equation (7)

$$\begin{aligned} \begin{bmatrix} x_k \\ y_k \end{bmatrix} &= \begin{bmatrix} 0 & \Delta \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{k-1} \\ y_{k-1} \end{bmatrix} + \begin{bmatrix} w_{k-1}^{(1)} \\ w_{k-1}^{(2)} \end{bmatrix}, \\ z_k &= \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_k \\ y_k \end{bmatrix} + v_k, \end{aligned} \tag{12}$$

where x , y , and Δ are DOA, angular velocity of DOA, and estimation interval respectively. The process noise terms $w_k^{(1)}$ and $w_k^{(2)}$ are set to constant value, and therefore the processing error covariance Q remains fixed.

The robust DOA estimation starts with initial values of x_0 , y_0 , and P_0 that are set to

$$\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad P_0 = \begin{bmatrix} v_0^2 & v_0 \\ v_0 & 1 \end{bmatrix}, \tag{13}$$

where v_0 is set to large value because the initial error is unknown in the starting of estimation. First, the system processes the prediction step (9) with the state equation described in (12). After the step, the measurement error covariance R is set as follows

$$R = [\lambda(1-\varepsilon)]^2, \tag{14}$$

where λ is a user-defined penalty parameter for unreliability. Using (14), the measurement update step (10) is performed.

As noted in section II, the covariance matrix R is related to Kalman gain K . If the matrix R is updated by (14), the system estimates the DOA based on the prediction $\hat{x}_k^-$ in (9) when the reliability ε is small, while the system estimates the DOA based on the measurement when the reliability ε is large. If the design parameters η , σ , and λ are appropriately set, the system may estimate the DOA robustly based on the reliability measurement.

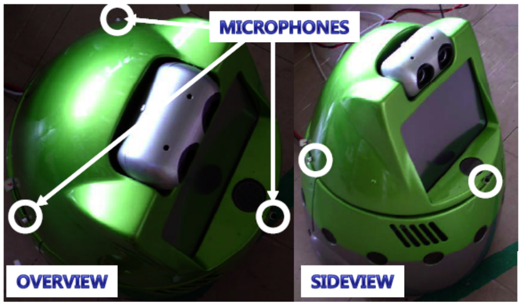

Fig. 6 Mobile Robot Platform for the Experiment

IV. Experimental Results

The performance of the proposed robust DOA estimation system is described in this section. It is applied to small size mobile robot as shown in Fig. 6. The robot has 667MHz Via-Eden CPU based hardware and windows based software. The robot has DC-motor based differential drive and Pen/Tilt motion system.

The performance is measured in terms of Field Of View (FOV) which represents the absolute difference between true and estimated sound direction. In the robot society, generally from 10° to 15° FOV is considered as a reasonable error.

To illustrate numerical results, the proposed system is tested for two cases. One is the static case and the other is the dynamic case. The user-defined parameters η , σ , and λ are set to 40, 10, and 60, which are found by intuitive trial and error method for all simulations. For convenience, the DOA estimation is called 'step 2', and the robust DOA estimation procedure is called 'step 3'. Because the step 2 can be shown as a simple extension of conventional GCC-PHAT, the performance of the proposed system can be compared to the result in the step 2.

The various estimation intervals Δ s are checked to show the relationship between Δ and the performance of the system. The initial parameter values of the Kalman filter are set to as follows

$$P=\begin{bmatrix} 8100 & 90 \\ 90 & 1 \end{bmatrix}, \quad Q=\begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}$$

A. Static Case

To build up the static database, a robot is fixed in the center $(0,0)$ of the room and a human speaks to the robot without movement during short time (about 0.5 second). The speaker is located on $(\chi\cos(\zeta),\chi\sin(\zeta))$, where χ is the distance from the center (1, 2, 3, 4, and 5 meter) and ζ is the angle from the baseline (0, 45, -45, 90, -90, 135, -135, 180) as shown in Fig. 7. The database is built up with three speakers.

To build up the reliability network, total 1773 training pairs are made using the signals extracted from 20ms cases. The trained network produces 3.2514e-4 10-fold cross validation error and is used for static and dynamic cases.

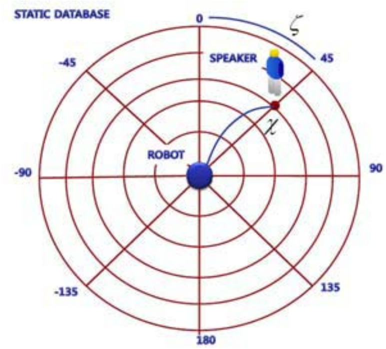

Fig. 7 Static Database Acquisition

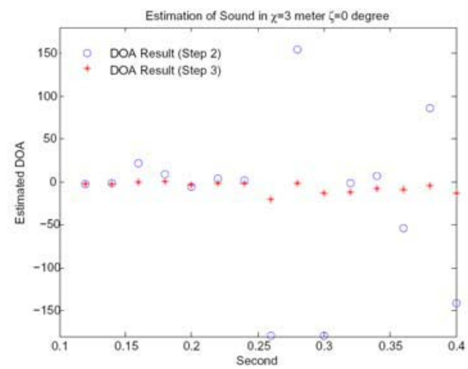

Fig. 8 The DOA estimation result from step 2 and step 3 in the condition $\Delta=20ms,\chi=3,\zeta=0$.

All speech signals are sampled with 16 kHz sampling rate using 16 bit analog-to-digital converter. Because of existence of degenerative effects such as low power, noise, and reverberation, the estimated time delay becomes unreliable in some interval Δ .

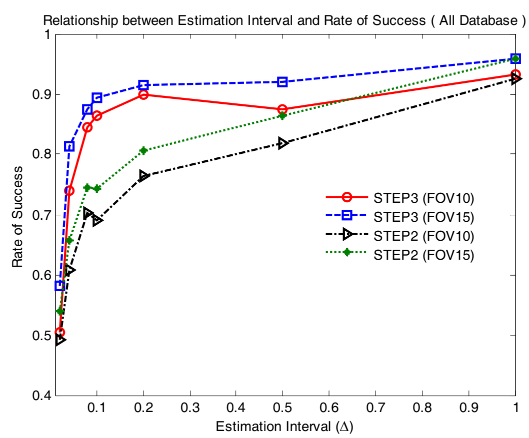

Fig. 9 The relationship between the estimation interval Δ and the performance of the system in all databases

Fig. 8 shows the result in static case when $\Delta = 20ms$. The proposed method (step 3) controls the outliers well. The relationship between estimation interval Δ and the performance of the system is illustrated in Fig. 9. The figure shows that the robust DOA estimation by the Kalman filter is extremely helpful when the estimation interval Δ is small. To get the robust performance for FOV10, the estimation interval Δ should be larger than 0.2.

B. Dynamic Case

For the dynamic case, the database is built up with four different motions of the robot while a human talks to the robot in a fixed point as shown in Fig. 10. The values from encoder sensors are recorded during the motion and used for the ground truth of the DOA.

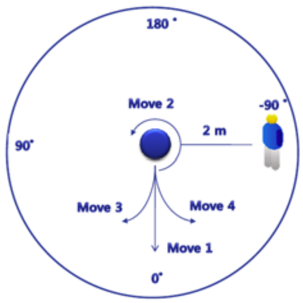

Fig. 10 Building up the database for the dynamic case

An example of the DOA estimation result (move 2) is shown in Fig. 11. The time interval is 40ms. The robot turns clockwise about 150° during about 4 seconds and the DOA estimation is successful even though the ground truth angle is continuously changed. The performance of the dynamic case is summarized in Table I. The proposed system provides the robust performance for four difference motions using the short interval estimation strategy. Even if the DOA is estimated at intervals of short time, the signal is not degenerated significantly by continuous delays.

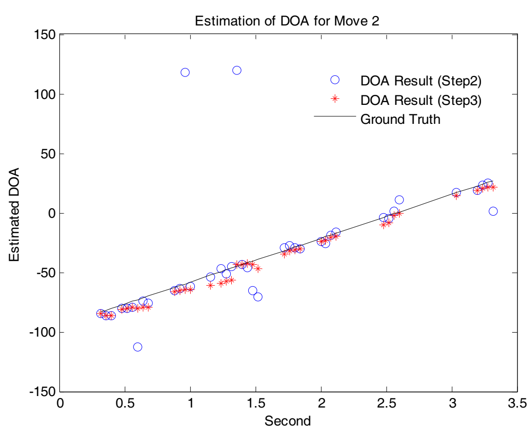

Fig. 11 An example of DOA estimation when the estimation interval is 40ms (Move 2)

TABLE I
THE PERFORMANCE OF DYNAMIC CASES ($\Delta = 40ms$)

STEP (FOV)	MOVE1	MOVE2	MOVE3	MOVE4
Step3 (10)	**0.8704**	**0.8952**	**0.9359**	**0.9638**
Step3 (15)	0.8827	0.9758	0.9680	0.9783
Step2 (10)	0.7654	0.8145	0.7500	0.7826
Step2 (15)	0.8025	0.8468	0.8205	0.8189

V. CONCLUSION

A robust sound direction estimation system is proposed in the paper using the Kalman filter based on reliability measure. The proposed technology is extremely useful when it finds and tracks a sound source at a distance within two meters as illustrated in the experimental result.

The novelty of the paper can be found in the calculation of the measurement error covariance matrix to filter out unreliable DOA estimations in Kalman filter process. The proposed DOA estimation tracks very fast for the reliable signals such as loud or noiseless voice whereas track very slowly for the unreliable signals such as noise.

REFERENCES

[1] M. N. Nicolescu, M. J. Mataric, "Learning and Interacting in human-robot domains", *IEEE Trans. System, Man and Cybernetics, Part A.*, vol. 31, no. 5, pp. 419-430, Sep. 2001.

[2] D. Kim, J. Lee, H. –S. Yoon, and E. –Y. Cha, "A Non-Cooperative User Authentication System in Robot Environment", *IEEE Trans. Consumer Electron.*, vol. 53, no. 2, pp. 804-811, May., 2007.

[3] J. Casper and R. R. Murphy, "Human-Robot Interactions During the Robot-Assisted Urban Search and Rescue Response at the World Trade Center", *IEEE Trans. System, Man and Cybernetics, Part B.*, vol. 33, no. 3, pp. 367-385, June., 2003.

[4] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, T. Poggio, "Robust Object Recognition with Cortex-Like Mechanisms", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 411-426, March, 2007.

[5] M. Omologo and P. Svaizer, "Use of the Crosspower-Spectrum Phase in Acoustic Event Location", *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 3, pp. 288-292, May, 1997.

[6] D. N. Notkin and R. Duraiswami, "Accelerated Speech Source Localization via a Hierarchical Search of Steered Response Power", *IEEE Trans. Speech and Audio Processing,* vol. 12, no. 5, pp. 499-508, Sep., 2004.

[7] P. Stoica and R. L. Moses, *Introduction to Spectral Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1997.

[8] C. H. Knapp, G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay", *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-24, no. 4, Aug., 1976.

[9] D. Hertz, "Time Delay Estimation by Combining Eficient Algorithms and Generalized Cross-Correlation Methods", *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-34, no. 1, Feb., 1986.

[10] J. –M. Valin, F. Michaud, J. Rouat, "Robust 3D Localization and Tracking of Sound Sources Using Beamforming and Particle Filtering", *Proc. IEEE Intl. Conf. on Acoustic, Speech, and Signal Processing*, pp. 1520-6149, 2006.

[11] D. Bechler, M. S. Schlosser, and K. Kroschel, "System for Robust 3D Speaker Tracking Using Microphone Array Measurements", *Proc. IEEE/RSJ, Intl. Conf. on Intelligent Robots and Systems*, pp. 2117-2122, 2004.

[12] G. Welch and G. Bishop, "An Introduction to the Kalman Filter", Dept. Comp. Sci. Uinv. North Carolina, Chapel Hill, TR95-041.

[13] C. –T. Kim, J. –J. Lee, H. Kim, "Variable Projection Method and Levenberg-Marqaurdt Algorithm for Neural Network Training", *IEEE 32nd Annual Conf. Industrial Elec.*, pp. 4492-4497, 2006.