Advanced Robotics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tadr20

Probabilistic Structure from Sound

Chieh-Chih Wang [a] , Chi-Hao Lin [b] & Jwu-Sheng Hu [c]

[a] Department of Computer Science and Information Engineering and Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 10617, Taiwan

[b] Department of Computer Science and Information Engineering and Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 10617, Taiwan

[c] Intelligent Robotics Program of Industrial Technology Research Institute, and Department of Electrical and Control Engineering at National Chiao Tung University, HsinChu 310, Taiwan

Published online: 02 Apr 2012.

To cite this article: Chieh-Chih Wang , Chi-Hao Lin & Jwu-Sheng Hu (2009) Probabilistic Structure from Sound, Advanced Robotics, 23:12-13, 1687-1702, DOI: 10.1163/016918609X12496339921975

To link to this article: http://dx.doi.org/10.1163/016918609X12496339921975

VSP

Advanced Robotics 23 (2009) 1687–1702

ADVANCED ROBOTICS
www.brill.nl/ar

Full paper

Probabilistic Structure from Sound

Chieh-Chih Wang [a,*], **Chi-Hao Lin** [a] **and Jwu-Sheng Hu** [b]

[a] Department of Computer Science and Information Engineering and Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 10617, Taiwan

[b] Intelligent Robotics Program of Industrial Technology Research Institute, and Department of Electrical and Control Engineering at National Chiao Tung University, HsinChu 310, Taiwan

Received 17 October 2008; accepted 13 January 2009

Abstract

Auditory perception is one of the most important functions for robotics applications. Microphone arrays are widely used for auditory perception in which the spatial structure of microphones is usually known. In practice, microphone array calibration can be tedious and other devices or means are required. The structure from sound (SFS) approach addresses the problem of simultaneously localizing a set of microphones and a set of acoustic events that provides a great flexibility to calibrate different setups of microphone arrays. However, the existing method does not take measurement uncertainty into account and does not provide uncertainty estimates of the SFS results. In this paper, we propose a probabilistic structure from sound (PSFS) approach using the unscented transform in which the uncertainties of the PSFS results are also available. In addition, a probabilistic sound source localization approach using the PSFS results is provided to improve sound source localization accuracy. The ample results of simulation and experiments using low-cost, off-the-shelf microphones demonstrate the feasibility and performance of the proposed PSFS approach.

Keywords

Auditory perception, microphone array, structure from sound, unscented transform

1. Introduction

While visual perception using cameras or laser scanners has been widely addressed and discussed in the robotics literature, auditory perception using microphones has attracted increasing attention over the last decade [1, 2]. To accomplish sound source localization using microphone arrays, the methods using interaural time difference, interaural phase difference, interaural level difference or fusing different cues have been demonstrated successfully [3–5]. Boll [6] proposed a method using spectral subtraction to suppress acoustic noise in speech. In addition to noise sup-

* To whom correspondence should be addressed. E-mail: bobwang@ntu.edu.tw

 DOI:10.1163/016918609X12496339921975

pression, Hu *et al.* [7] utilized Gaussian mixture models to deal with the issues of complicated environment acoustics and microphone mismatch in situations such as detecting a speaker's position within a noisy vehicle cabinet. Mak and Furukawa [8] proposed a time-of-arrival-based positioning technique to deal with non-line-of-sight situations with low-frequency acoustic signals. Sasaki *et al.* [9] described a method to localize multiple stationary and moving sound sources using a moving microphone array. Yamamoto *et al.* [2] show the capability of recognizing three simultaneous speeches. Valin *et al.* [10, 11] demonstrated the feasibility of simultaneous multiple sound source localization. In Ref. [12], microphones are distributed in the environment for acoustic robot localization in which the microphone array is well calibrated. A comprehensive survey of auditory perception in robotics is available in Chapter 2 of Ref. [13]. It is shown that microphone arrays are widely used for auditory perception.

In most auditory perception applications, the microphone locations are usually known or calibrated. The calibration process could be tedious, in which case other means or equipment are required. The structure from sound (SFS) problem is to simultaneously localize a set of microphones and a set of sound sources. A solution to the SFS problem can provide a means to calibrate microphone arrays easily. Without using any additional equipment, creating sound events at different locations is sufficient to complete the calibration process. In Ref. [14], Thrun proposed an affine SFS algorithm and demonstrated its performance using a microphone array comprised of seven Crossbow sensor motes. However, measurement uncertainty is not taken into account and the SFS estimate uncertainties are not provided. Figure 1a shows a simulation result in which the affine SFS converges to an incorrect result under measurement uncertainty.

Based on Thrun's approach, we propose a probabilistic structure from sound (PSFS) algorithm using the unscented transform [15]. Given the uncertainty estimates of interaural time differences between microphones, sample sets of time delay estimates are generated and used as inputs of the SFS algorithm. Accordingly, sample sets of estimated locations of microphones and sound sources are computed using the SFS algorithm. The location estimates of microphones and sound sources can be represented by these weighted SFS output samples. Unfortunately, as only one microphone is selected as the origin of the coordinate system in the SFS framework, the SFS output samples may suffer from the rotation effect and the mirror effect as depicted in Fig. 2. To estimate the uncertainties correctly, these axis inconsistency problems should be dealt with. In this paper, the coordinate systems of the SFS output samples in two-dimensional (2-D) cases are aligned by selecting one microphone as the origin of the coordinate system and then letting another selected microphone move only in the x-axis of this coordinate system.

In the SFS framework, the location estimates of microphones are more accurate than the sound source location estimates as more measurements or constraints are involved with microphones than with sound sources. However, given the PSFS results, sound source localization can be further improved with more measurements.

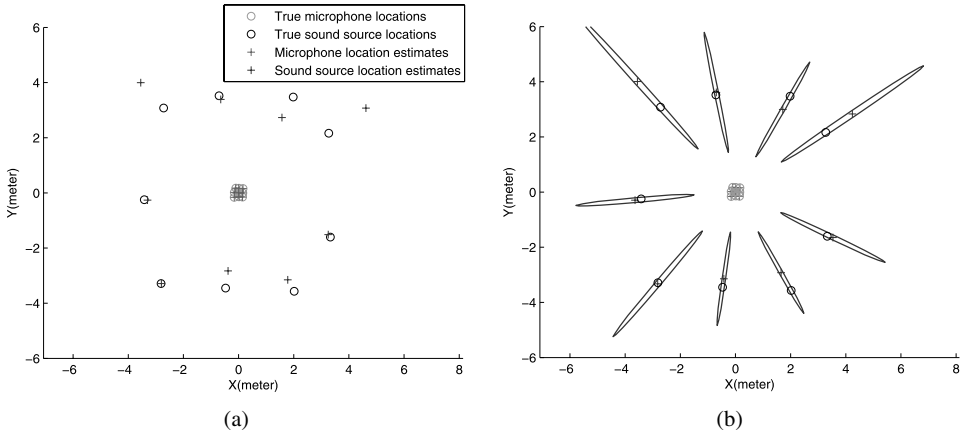


Figure 1. A microphone array is located around the origin and nine sound events are generated at different locations surrounded the microphone array. Noises are added to measurements. (a) Results using the affine SFS algorithm in which measurement uncertainty is not taken into account. (b) Results using the proposed probabilistic SFS algorithm in which the measurement uncertainty is properly dealt with. The ellipses show 2σ estimates of the sound sources.

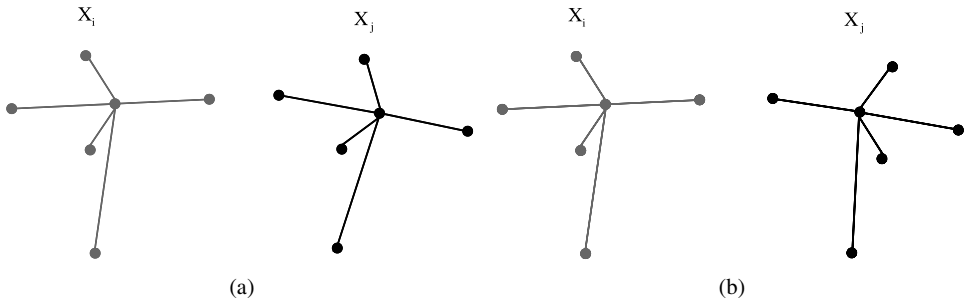


Figure 2. Axis inconsistency problems. The different results all satisfy the constraints. (a) SFS output samples can be rotated around the origin (the selected microphone). (b) SFS output samples can be flipped over around some axis.

We again utilize the unscented transform to accomplish probabilistic sound source localization (PSSL). In addition, we demonstrate that sound source localization can be further improved with a moving microphone array using the proposed framework. Ample simulations and experiments using off-the-shelf microphones verify the proposed PSFS and PSSL algorithms.

The rest of this paper is organized as follows. In Section 2, the affine SFS algorithm is briefly reviewed. Section 3 addresses the proposed PSFS algorithm in detail. Section 4 describes our PSSL algorithm. The simulation and experimental results are given in Section 5, and the conclusions and future work are given in Section 6.

2. Affine SFS

In this section, the affine SFS algorithm [14] is described briefly to provide a foundation for understanding the proposed PSFS algorithm. The SFS problem is to localize the N microphones and M sound sources simultaneously. All the sound sources are emitted from unknown locations at unknown time and all the microphones are located at unknown positions. It is assumed that all microphones are synchronized.

Let X be the microphone location matrix of size $N \times 2$ and A be the sound source location matrix of size $M \times 2$:

$$X = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_N & y_N \end{bmatrix}, \qquad A = \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \\ \vdots & \vdots \\ a_M & b_M \end{bmatrix}. \tag{1}$$

Let the first microphone be located at the origin of the coordinate system and let the second microphone only move in the x-axis of this coordinate system:

$$x_1 = 0, \qquad y_1 = 0, \qquad y_2 = 0. \tag{2}$$

The matrix of relative arrival time is defined as:

$$\Delta = c^{-1} \begin{bmatrix} d_{2,1} - d_{1,1} & d_{2,2} - d_{1,2} & \cdots & d_{2,M} - d_{1,M} \\ d_{3,1} - d_{1,1} & d_{3,2} - d_{1,2} & \cdots & d_{3,M} - d_{1,M} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N,1} - d_{1,1} & d_{N,2} - d_{1,2} & \cdots & d_{N,M} - d_{1,M} \end{bmatrix}, \tag{3}$$

where $d_{m,n}$ denotes the distance between microphone m and sound source n, and c denotes the speed of sound. The difference between the distance from the jth sound source to the ith microphone and the distance from the jth sound source to the reference microphone, $\Delta_{i,j}$, can be expressed as:

$$\Delta_{i,j} = c^{-1} \left\{ \left| \begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} a_j \\ b_j \end{pmatrix} \right| - \left| \begin{pmatrix} a_j \\ b_j \end{pmatrix} \right| \right\}. \tag{4}$$

The time delay $\Delta_{i,j}$ can be measured using the synchronized microphone array. These time delays are the only inputs of the whole SFS system. The SFS problem can be formulated as a least-squares problem in which X and A are computed by minimizing the cost function:

$$\arg\min_{A,X} \sum_{i=2}^{N} \sum_{j=1}^{M} \left\{ \left| \begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} a_j \\ b_j \end{pmatrix} \right| - \left| \begin{pmatrix} a_j \\ b_j \end{pmatrix} \right| - \Delta_{i,j} \cdot c \right\}^2. \tag{5}$$

This problem can be solved through the gradient descent method. However, a good initial guess of the locations of microphone and sound sources is critical to minimize (5). Following the idea of affine structure from motion [16] in the computer vision literature, the affine SFS approach assumes that sound sources are far

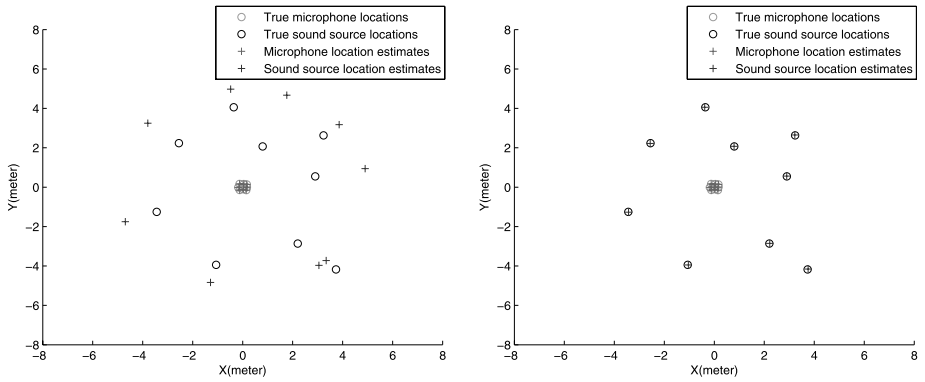


Figure 3. A simulation of the affine SFS algorithm. (a) Results under the far field approximation. The affine solution is then used as the initial guess for minimizing (5) using the gradient descent method. (b) Final result that converges to the correct locations. Note that no measurement uncertainty is added in this simulation.

away from the microphones and the incoming sound wave hits each microphone at the same incident angle. The SFS problem is simplified to recover the incident angles of the sound sources. This assumption is used to get a reasonable initial guess of the locations of microphones and sound sources for minimizing (5). The gradient descent method is then applied to recover the microphone and sound source locations. Figure 3 shows the results of a simulation. Note that no measurement uncertainty is added in this simulation. The affine SFS algorithm often converges to incorrect results if measurements are uncertain.

3. PSFS

In this section, we describe the proposed PSFS approach using the unscented transform [15].

3.1. Unscented Transform

Let x be a L-dimensional Gaussian with the mean μ_x and covariance matrix Σ_x. Let $y = f(x)$ be a nonlinear transformation from x to y. In the unscented transform, the mean and covariance of x can be presented by the $2L + 1$ sigma points. The $2L + 1$ sigma points are generated according to the following rule:

$$\begin{aligned} \chi_0 &= \mu, \\ \chi_i &= \mu + \left(\sqrt{(L+\lambda)\Sigma}\right)_i \quad \text{for } i = 1, \ldots, L \\ \chi_i &= \mu - \left(\sqrt{(L+\lambda)\Sigma}\right)_i \quad \text{for } i = L+1, \ldots, 2L, \end{aligned} \tag{6}$$

where $\lambda = \alpha^2(L + k) - L$. α and k are scaling parameters that determine the spread of the sigma points from the mean. Each sigma point $\mathcal{X}_i$ has two weights associated

with it. The first one, $w_i^{(m)}$, is used to recover the mean and the second one, $w_i^{(c)}$, is used to recover the covariance. These sigma points are passed through the the function f:

$$\mathcal{Y}_i = f(\mathcal{X}_i) \quad i = 0, \ldots, 2L. \tag{7}$$

The corresponding sigma point $\mathcal{Y}_i$ can be computed. Finally, the mean μ_y and covariance Σ_y can be calculated by:

$$\begin{aligned} \mu_y &= \sum_{i=0}^{2L} w_i^{(m)} \mathcal{Y}_i \\ \Sigma_y &= \sum_{i=0}^{2L} w_i^{(c)} (\mathcal{Y}_i - \mu_y)(\mathcal{Y}_i - \mu_y)^{\mathrm{T}}. \end{aligned} \tag{8}$$

3.2. PSFS

As the relative arrival time matrix Δ of size $(N-1) \times M$ is the input of the nonlinear SFS process, the sigma points can be computed given Δ and the corresponding covariance matrix. To apply the formula of the unscented transform, Δ is reformed as a long $L = (N-1) \times M$-dimensional random vector:

$$\mu_\Delta = [\, \mu_1 \quad \cdots \quad \mu_L \,]^{\mathrm{T}}. \tag{9}$$

As each element in μ_Δ is a time delay with a variance σ_i^2, the corresponding covariance matrix is a diagonal matrix of the form:

$$\Sigma_\Delta = \mathrm{diag}\,(\, \sigma_1^2 \quad \cdots \quad \sigma_L^2 \,). \tag{10}$$

Now the mean μ_Δ and covariance matrix Σ_Δ can be used to extract the sigma points using the unscented transform. These sigma points are reformed as matrices of size $(N-1) \times M$, which are passed through the standard SFS procedure. The location mean and covariance of each microphone and each sound source are recovered with the weighted combination of each corresponding sigma point using (8). Figure 4 shows the simulations results of PSFS using the same microphone array with different sound source configurations. Although the performances of sound source localization may depend on sound source configurations, microphone array calibration remains accurate. The experimental results using low cost, off-the-shelf microphones are shown in Section 5.

4. PSSL

As the PSFS framework may not provide very accurate sound source localization under uncertainty, a PSSL algorithm to improve accuracy and performance of sound source localization is proposed and addressed in this section. The process we apply to solve the SSL problem is similar to SFS. The SSL problem can be formalized as an optimization problem in which we need to find a sound event location to minimize the quadratic difference between the predicted and real measurements. The

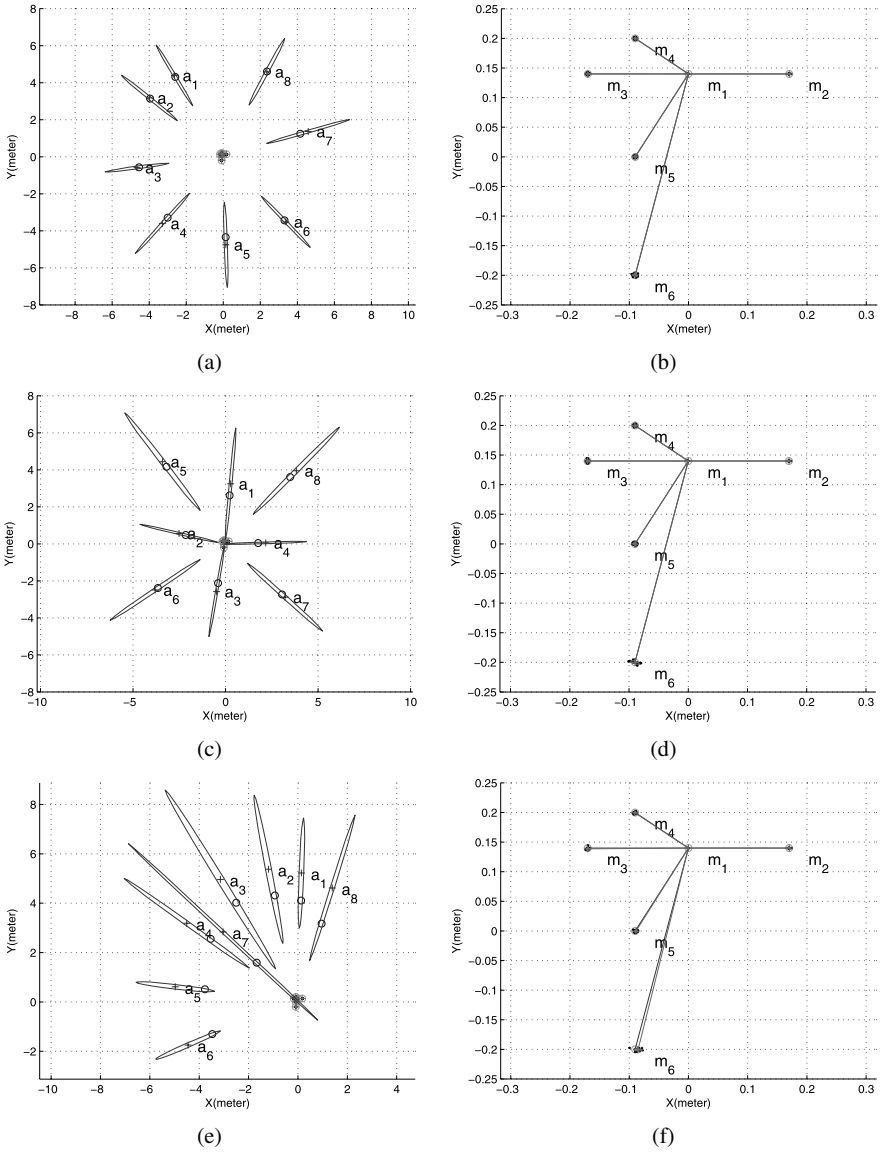


Figure 4. Simulation results of the PSFS algorithm with uncertain measurements. Different sound source configurations were tested to show the generality of the proposed approach: (a), (c) and (e) show the PSFS results, and (b), (d) and (f) are the enlargements of (a), (c) and (e). The centers of the circles show the true locations of microphones and sound sources, respectively. (a, c, e) Crosses and ellipses show the means and the 2σ bounds of the sound source estimates. (b, d, f) Crosses and ellipses show the means and the 2σ bounds of the microphone estimates. (a) The sound source configuration 1. Sound sources were equally distributed with a constant distance around the microphone array. (b) The enlargement of (a) to show the microphone structure estimates. (c) The sound source configuration 2. (d) The enlargement of (c) to show the microphone structure estimates. (e) The sound source configuration 3. Sound sources were only distributed on one-side of the the microphone array. (f) The enlargement of (e) to show the microphone structure estimates.

proposed PSSL algorithm applies the unscented transform to utilizes the microphone array estimate and the corresponding uncertainty from PSFS with the new measurements.

Recall that PSFS provides a location mean m_i and covariance matrix Σ_{m_i} for each microphone i. For any new measurement of a sound source/event, we can compute a vector of time delays between microphones:

$$\Delta = [\,\delta_{1,2} \quad \delta_{1,3} \quad \delta_{1,4} \quad \cdots \quad \delta_{N-1,N}\,]^{\mathrm{T}}, \tag{11}$$

where $\delta_{i,j}$ denotes the time delay between the ith microphone and the jth microphone. This data vector is of dimension $C_2^N \times 1$. We can also have the variance of each time delay estimate $\sigma_{i,j}^2$.

To extract sigma points in the unscented transform, the microphone estimate X and the time delay estimates Δ are combined and reformed as:

$$\mu = [\,x_1 \quad y_1 \quad \cdots \quad x_N \quad y_N \quad \delta_{1,2} \quad \cdots \quad \delta_{N-1,N}\,]^{\mathrm{T}}, \tag{12}$$

where μ is of dimension $L = (2N + C_2^N) \times 1$.

The covariance matrix of the input can be computed by:

$$\Sigma = \begin{bmatrix} \Sigma_{m_1} & & & & & & \\ & \ddots & & & \mathbf{0} & & \\ & & \Sigma_{m_N} & & & & \\ & & & \sigma_{1,2}^2 & & & \\ & \mathbf{0} & & & \ddots & & \\ & & & & & \sigma_{N-1,N}^2 \end{bmatrix}. \tag{13}$$

The sigma points can be computed straightforwardly using the unscented transform formulation. These sigma points are of the form:

$$\chi_i = \begin{bmatrix} x_1^{(l)} & y_1^{(l)} & \cdots & x_N^{(l)} & y_N^{(l)} & \delta_{1,2}^{(l)} & \cdots & \delta_{N-1,N}^{(l)} \end{bmatrix}^{\mathrm{T}} \quad l = 0, \ldots, 2L. \tag{14}$$

Each sigma point is then passed through a least-squares problem solver:

$$A^{(l)*} = \arg\min_{A} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left\{ \left| \begin{pmatrix} x_i^{(l)} \\ y_i^{(l)} \end{pmatrix} - \begin{pmatrix} a^{(l)} \\ b^{(l)} \end{pmatrix} \right| - \left| \begin{pmatrix} x_j^{(l)} \\ y_j^{(l)} \end{pmatrix} - \begin{pmatrix} a^{(l)} \\ b^{(l)} \end{pmatrix} \right| - \delta_{i,j}^{(l)} \cdot c \right\}^2 \quad l = 0, \ldots, 2L. \tag{15}$$

A location of the sound source $A^{(l)} = (a^{(l)}, b^{(l)})$ is found accordingly. The mean μ' and covariance Σ' of the sound source location can be recovered through the standard unscented transform procedure (8). Figure 5 shows a real experimental result of the proposed PSSL algorithm using three microphones. The microphone array is pre-calibrated using PSFS. The result demonstrates that PSSL provides a proper sound sound location estimate. The experiment details will be described in the next section.

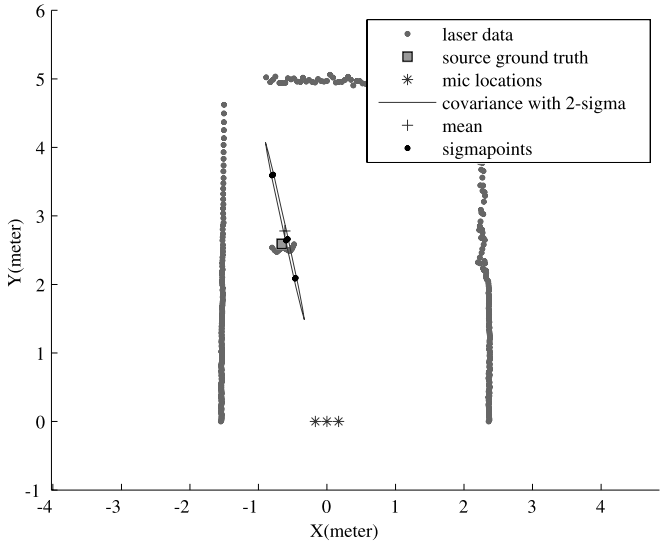


Figure 5. Experimental result of PSSL using three microphones.

5. Experimental Results

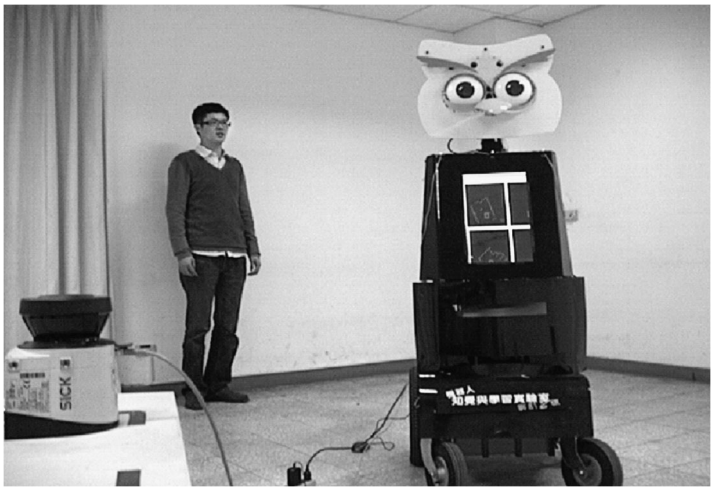
In this section, the proposed PSFS and PSSL algorithms are evaluated through experiments using real speech collected from a person. Figure 6 shows the experiment setup in which an eight-channel A/D board is used to collect sound source data and six low-cost, off-the-shelf microphones are mounted on the NTU-PAL2 robot. A SICK S200 laser scanner is used for collecting ground truth. Two types of experiments were conducted: one to calibrate the microphone array using the proposed PSFS algorithm and the other to localize the sound source with the calibrated microphone array using the proposed PSSL algorithm. We further demonstrate PSSL with a moving microphone array.

5.1. Time Delay Estimation

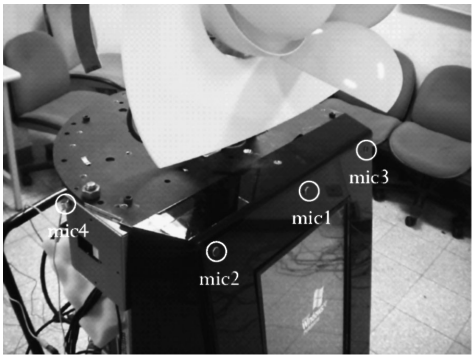
In each experiment, 10 s of speech data at different locations were collected from six microphones. All sound source signals were sampled at 44.1 kHz. The time delay of arrival (TDOA) estimation was performed using 1024 samples and 512 samples were shifted at the next frame. The generalized cross-correlation approach [17] is utilized to estimate time delays between microphones. As there are silent segments in these speeches, the TDOA estimates may be unstable. The peak of the histogram of the TDOA estimates of 10 s of speech was chosen as the input of PSFS or PSSL. Figure 7 illustrates the approach to estimate time delay between microphones.

5.2. PSFS Results

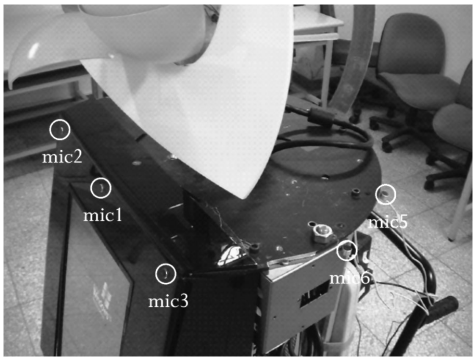
The PSFS experiments were conducted in two different environments in terms of environment sizes. The first experiment was performed in a seminar room for the



(a)



(b)



(c)

Figure 6. Experiment setup. (a) The NTU-PAL2 robot, a person and a SICK S200 laser scanner. (b) The microphone positions. (c) The microphone positions.

near-field condition. The room is about 6 m × 6 m in size. The second experiment was performed in an atrium for the far-field condition. The atrium is about 16 m × 18 m in size. The laser scanner was used to detect the speaker's location for evaluation.

5.2.1. Near-Field Condition

The experiment was conducted in the seminar room and the sound sources were about 2–3 m away from the microphones. Six speeches were collected at different locations. Figure 8 shows the experimental results of PSFS with six sound sources. The average angular error is 0.35°. The average microphone location error is 0.0081 m. The average sound source location error is 0.75 m.

5.2.2. Far-Field Condition

The experiment was conducted in the atrium and the sound sources were about 6–8 m away from the microphones. Six speeches were collected at different locations.

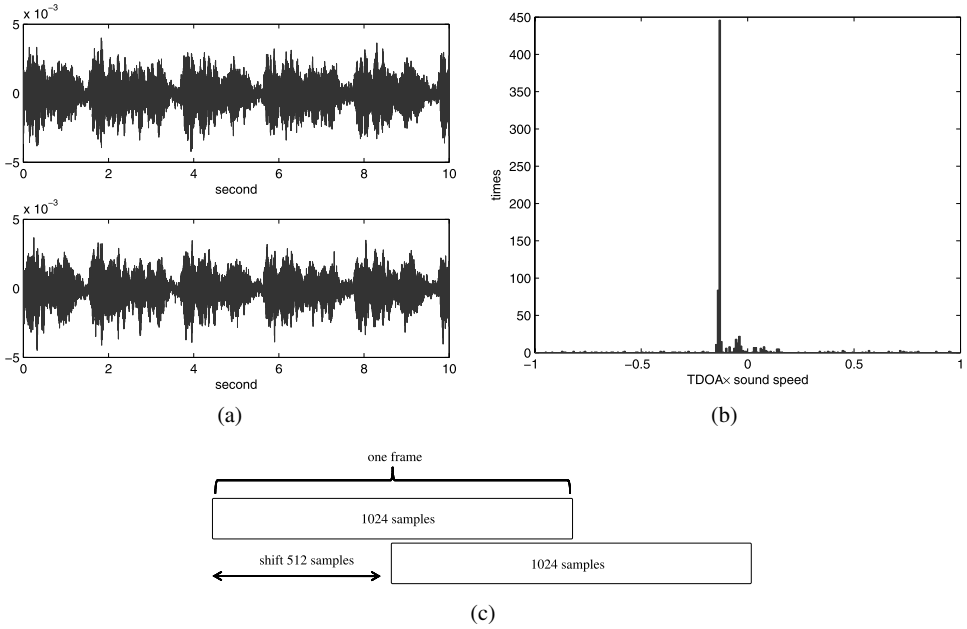


Figure 7. Example of TDOA estimation. (a) The waveforms of 10 s speech from two microphones. (b) The histogram of TDOA of 10 s speech. (c) The overlap setting.

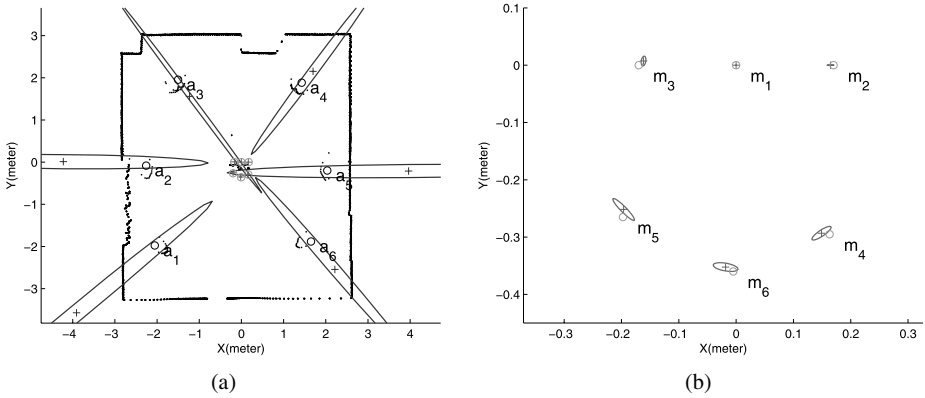
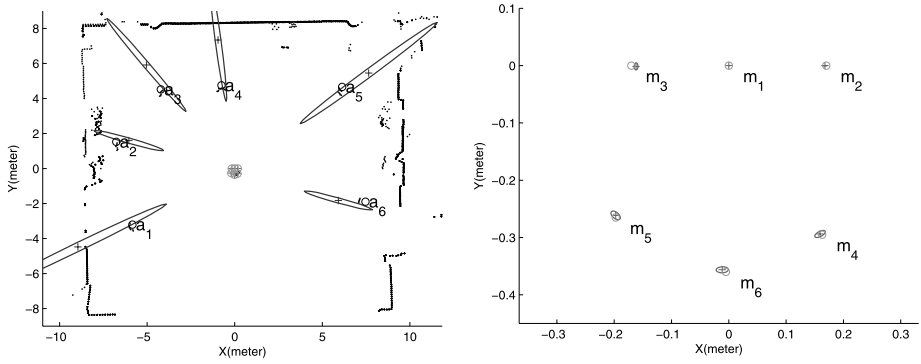


Figure 8. Near-field situation. The result of the PSFS experiment performed in the seminar room. The 2σ ellipses show that our approach provides proper uncertainty estimates of the PSFS results under the situations that measurements are uncertain. (a) The sound source estimates. (b) The microphone structure estimates.

Figure 9 shows the experimental results of PSFS with six sound sources. The average angular error is 2.0245°. The average microphone location error is 0.0041 m. The average sound source location error is 7.0773 m.

5.2.3. Evaluation

The above results were all carefully evaluated. The average sound source location errors of 0.75 m in the near-field condition and 7.0773 m in the far-field condition



(a) (b)

Figure 9. Far-field situation. The results of the PSFS experiment performed in the atrium. (a) The sound source estimates. (b) The microphone structure estimates.

are reasonable results given that the baselines of the microphones are very short. Figures 8 and 9 demonstrate that our approach provides proper uncertainty estimates of the PSFS results under the situations that measurements are uncertain.

5.3. PSSL Results

With the PSFS results, two experiments of PSSL were conducted. One was to localize the sound source using the static robot with more measurements. Figure 10 shows the PSSL results in a series of measurement updates. As both the microphone array and the sound source are stationary, the PSSL results are roughly the same as indicted by dark grey ellipses. Light grey ellipses become smaller after three measurement updates, which demonstrates that the estimates are more accurate and certain with more measurement updates.

The other experiment was to localize the sound source using a moving robot. Figure 11 shows that the PSSL results can be greatly improved. The robot movement was estimated by scan matching using laser scanner data. As odometry can also provide good robot movement estimates locally, similar performance can be achieved using inexpensive odometry.

6. Conclusions

Microphone arrays are widely used for auditory perception. However, microphone array calibration can be tedious in practice and other devices or means are required. The existing SFS framework provides a nice approach to simultaneously calibrate microphones and sound sources without using any other devices. Unfortunately, SFS does not take time delay estimate uncertainty into account. In this paper, we proposed the PSFS approach using the unscented transform to deal with this issue. The uncertainty estimates of the PSFS results are also available in our framework. We have shown that the estimates of sound sources are more uncertain than microphones in both SFS and PSFS. Accordingly, we proposed the PSSL approach

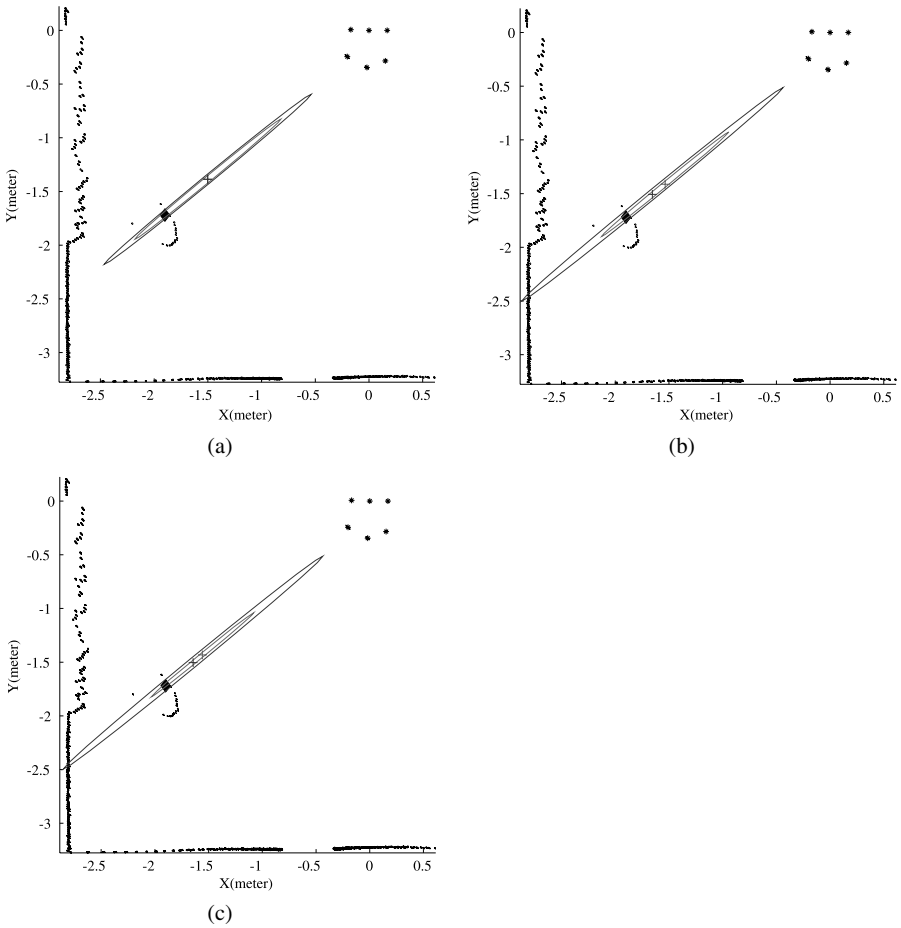


Figure 10. PSSL results with a series of measurement updates. Black stars indicate the locations of microphones. The black solid diamond is the location of the sound source. Ellipses show the estimate uncertainty (2σ) of the PSSL result. As both the microphone array and the sound source are stationary, the PSSL results are roughly the same as indicted by dark grey ellipses. Light grey ellipses show the estimate uncertainty (2σ) of the PSSL result after the measurement update. The light grey ellipse in (c) is smaller than in (a) and (b). (a) Update 1. (b) Update 2. (c) Update 3.

to improve the accuracy of SSL with more measurements. We also demonstrated that the accuracy of PSSL can be greatly improved with a moving microphone array/robot. The simulation and experimental results verify the proposed PSFS and PSSL algorithms.

As SFS could converge to an incorrect result under measurement uncertainty, PSFS may provide an incorrect estimate as well. Detecting SFS failures by analyzing time delay estimates between microphones could be a feasible approach to deal with these issues. In addition, dealing with the issues of diffraction/reflection of sound by robot itself, walls and obstacles, and applying particle filters to PSFS and PSSL should be of our interests.

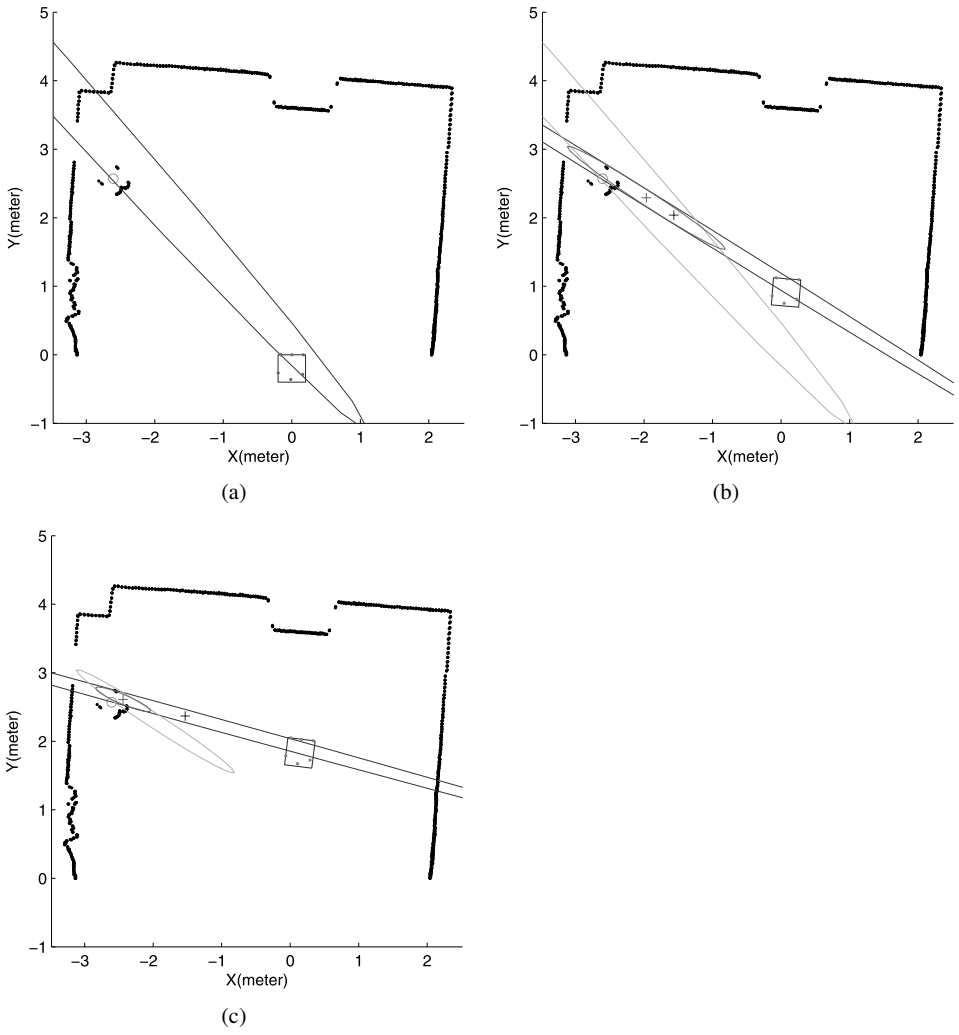


Figure 11. PSSL result with a moving robot. Circles show the true location of the sound source. Dark grey ellipses show the estimate uncertainty (2σ) of the PSSL result. Light grey ellipses show the estimate uncertainty (2σ) of the PSSL result after the measurement update. (a) Update 1. (b) Update 2. (c) Update 3.

Acknowledgements

The authors would like to acknowledge the valuable suggestions from anonymous reviewers. This work was partially supported by grants from Taiwan NSC (96-2628-E-002-251-MY3, 97-2218-E-002-017, 98-2623-E-002-017-D); Excellent Research Projects of the National Taiwan University (95R0062-AE00-05); Taiwan DOIT TDPA Program (95-EC-17-A-04-S1-054); Taiwan ITRI, CCI, MSI; and Intel.

References

1. K. Nakadai, D. Matsuura, H. G. Okuno and H. Kitano, Applying scattering theory to robot audition system: robust sound source localization and extraction, in: *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Las Vegas, NV, pp. 1147–1152 (2003).
2. S. Yamamoto, K. Nakadai, H. Tsujino and H. G. Okuno, Asssessment of general applicability of robot audition system by recognizing three simultaneous speeches, in: *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Sendai, pp. 2111–2116 (2004).
3. S. T. Birchfield and R. Gangishetty, Acoustic localization by interaural level difference, in: *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, pp. IV-1109–IV-1112 (2005).
4. W. Cui, Z. Cao and J. Wei, Dual-microphone source location method in 2-D space, in: *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Toulouse, pp. IV-845–IV-848 (2006).
5. K. C. Ho and M. Sun, Passive source localization using time differences of arrival and gain ratios of arrival, *IEEE Trans. Signal Process.* **56**, 464–477 (2008).
6. S. F. Boll, A spectral subtraction algorithm for suppression of acoustic noise in speech, *IEEE Trans. Acoust. Speech Signal Process.* **27**, 131–140 (1979).
7. J.-S. Hu, C.-C. Cheng and W.-H. Liu, Robust speaker's location detection in a vehicle environment using GMM models, *IEEE Trans. Syst. Man Cybernet. B* **36**, 403–412 (2006).
8. L. C. Mak and T. Furukawa, A time-of-arrival-based positioning technique with non-line-of-sight mitigation using low-frequency sound, *Adv. Robotics* **22**, 507–526 (2008).
9. Y. Sasaki, S. Kagami and H. Mizoguchi, Multiple sound source mapping for a mobile robot by self-motion triangulation, in: *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Beijing, pp. 380–385 (2006).
10. 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).
11. J.-M. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai and H. G. Okuno, Robust recognition of simultaneous speech by a mobile robot, *IEEE Trans. Robotics* **23**, 742–752 (2007).
12. Q. H. Wang, T. Ivanov and P. Aarabi, Acoustic robot navigation using distributed microphone arrays, *Inform. Fusion* **5**, 131–140 (2004).
13. E. B. Martinson, Acoustical awareness for intelligent robotic action, *PhD Dissertation*, Georgia Institute of Technology (2007). Also available at http://www.cc.gatech.edu/ai/robot-lab/online-publications/MartinsonThesis2007.pdf
14. S. Thrun, Affine structure from sound, in: *Proc. Neural Information Precessing Systems Conf.*, Vancouver, pp. 1353–1360 (2006).
15. S. J. Julier and J. K. Uhlmann, Unscented filtering and nonlinear estimation, *Proc. IEEE* **92**, 401–422 (2004).
16. C. Tomasi and T. Kanade, Shape and motion from image streams under orthography: a factorization method, *Int. J. Comput. Vision* **9**, 137–154 (1992).
17. 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).

About the Authors

Chieh-Chih Wang received the BS and MS degrees from National Taiwan University in 1994 and 1996, respectively. He earned his PhD in Robotics from the School of Computer Science, Carnegie Mellon University, in 2004. During his graduate study, he worked with the Bayesian Vision Group at the NASA Ames Research Center and at Z + F Inc. in Pittsburgh, PA, USA. From 2004 to 2005, he was an Australian Research Council (ARC) Research Fellow of the ARC Centre of Excellence for Autonomous Systems and the Australian Centre for Field Robotics at the University of Sydney. In 2005, he joined the Department of Computer Science and Information Engineering at National Taiwan University where he is an Assistant Professor, and is pursuing his academic interests in robotics, machine perception and machine learning. He received the Best Conference Paper Award at the 2003 IEEE International Conference on Robotics and Automation.

Chi-Hao Lin received the BS degree in Computer Information Science and Business Mathematics from Soochow University, in 2005, and the MS degree in Computer Science and Information Engineering from the National Taiwan University, in 2008. His research interests include sound source localization, and simultaneous localization and mapping.

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. From 1991 to 1993, he was an Assistant Professor in the Department of Mechanical Engineering, Wayne State University, Detroit, MI, USA, where he received the Research Initiation Award from the National Science Foundation. In 1993, he joined the Department of Electrical and Control Engineering, National Chiao Tung University, Taiwan and became a Full Professor, in 1998. He served as the Vice-Chairman of the Department since 2006 and from 2008 he has worked in-part at the Industrial Technology Research Institute of Taiwan where he serves as the Advisor for the Intelligent Robotics program. His current research interests include robotics, microphone arrays, active noise control and embedded systems.