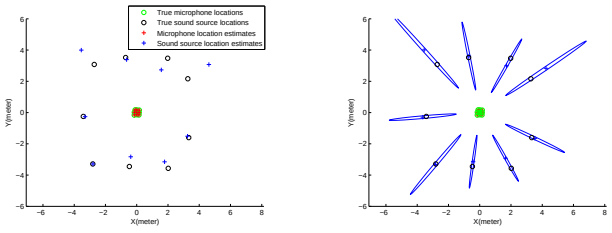
# Probabilistic Structure from Sound and Probabilistic Sound Source Localization

Chi-Hao Lin[†] and Chieh-Chih Wang[†‡]
[†]Department of Computer Science and Information Engineering
[‡]Graduate Institute of Networking and Multimedia
National Taiwan University, Taipei, Taiwan
Email: r94922124@ntu.edu.tw,bobwang@ntu.edu.tw

***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. The structure from sound (SFS) approach addresses the problem of simultaneously localizing a set of microphones and a set of acoustic events which 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 addition, a probabilistic sound source localization (PSSL) 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-shell microphones demonstrate the feasibility and performance of the proposed PSFS and PSSL approaches.**




(a) The Affine SFS results under measurement uncertainty

(b) The Probabilistic SFS results under measurement uncertainty

Fig. 1. A microphone array is located around the origin and nine sound events are generated at different locations surrounded the microphone array. Our probabilistic SFS algorithm properly deals with the measurement uncertainty. The ellipses show $2\sigma$ estimates of the sound sources.

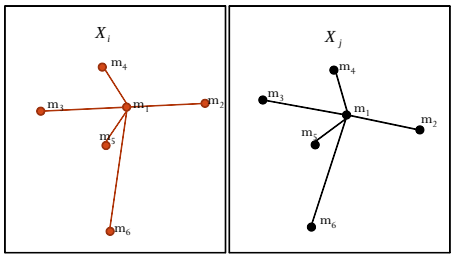
## I. Introduction

While visual perception using cameras or laser scanners has been widely addressed and discussed, only a few works in the robotics literature addressed auditory perception using microphones. 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 [1][2][3]. To deal with the issues of noise, complicated environment acoustics and microphone mismatch, Hu *et al.* [4] utilized Gaussian mixture models for detecting a speaker's position within a noisy vehicle cabinet. Valin *et al.* [5][6] demonstrated the feasibility of simultaneous multiple sound source localization. A comprehensive survey of auditory perception in robotics is available in Chapter 2 of [7]. It is shown that microphone arrays are widely used for auditory perception.

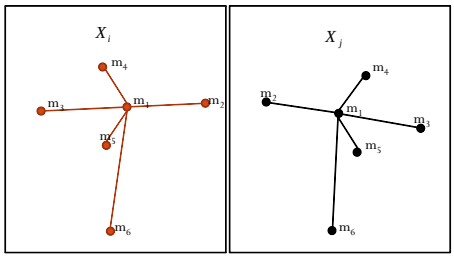
In most of the auditory perception applications, the microphone locations are usually known or calibrated. The calibration process could be tedious in which other means or equipments are required. The structure from sound (SFS) problem is to simultaneously localize a set of microphones and a set of sound sources. A solution to the SFS problem can provide a means to calibrate microphone arrays easily. Without using any additional equipments, creating sound events at different locations is sufficient to complete the calibration process. In [8], 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. Fig. 1 shows a simulation result in which the affine SFS converges to an incorrect results under measurement uncertainty.

Based on Thrun's approach, we propose a probabilistic structure from sound (PSFS) algorithm using the unscented transform [9]. 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 rotate 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 2D 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




(a) SFS output samples can be rotated around the origin (the selected microphone).




(b) SFS output samples can be flipped over around some axis.

Fig. 2. The axis inconsistency problems. The different results all satisfy the constraints.

improved with more measurements. 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 arrays using the proposed framework. Ample simulations and experiments using off-the-shell microphones verify the proposed PSFS and PSSL algorithms.

The rest of this paper is arranged as follows: In Section II, the affine SFS algorithm is briefly reviewed; Section III addresses the proposed PSFS algorithm in detail; Section IV describes our PSSL algorithm. The simulation and experimental results are in Section V, and the conclusion and future work are in Section VI .

## II. Affine Structure from Sound

In this section, Thrun's affine SFS algorithm [8] 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 $M \times 2$.

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

The SFS problem can be formulated as a least square problem in which $X$ and $A$ are computed by minimizing the cost function:

$$\underset{A,X}{\operatorname{argmin}} \sum_{i=2}^{N} \sum_{j=1}^{M} \left\{ \left| \begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} a_j \\ b_j \end{pmatrix} \right| - \left| \begin{pmatrix} a_j \\ b_j \end{pmatrix} \right| - \Delta_{i,j} \right\}^2 \tag{2}$$

where $\Delta_{i,j}$ denotes the difference between the distance from the $j$th sound source to the $i$th microphone and the distance from the $j$th sound source to the reference microphone. $\Delta_{i,j}$ is computed by multiplying the time delay estimate with the sound speed. 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 Eqn. 2.

Following the idea of affine structure from motion [10] in the computer vision literature, the affine SFS approach assumes that sound sources are far 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. The gradient descent method is then applied to recover the microphone and sound source locations.

## III. Probabilistic Structure from Sound

In this section, we describe the proposed PSFS approach using the unscented transform [9].

### A. The Unscented Transform

Let $x$ be a L-dimensional Gaussian with the mean $\mu_x$ and covariance matrix $\Sigma_x$. 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. 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 $f$:

$$\mathcal{Y}_i = f(\mathcal{X}_i) \qquad \text{where} \quad i = 0, \ldots, 2L \tag{3}$$

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

$$\begin{aligned} \mu_y &= \textstyle\sum_{i=0}^{2L} w_i^{(m)} \mathcal{Y}_i \\ \Sigma_y &= \textstyle\sum_{i=0}^{2L} w_i^{(c)} (\mathcal{Y}_i - \mu_y)(\mathcal{Y}_i - \mu_y)^T \end{aligned} \tag{4}$$

### B. PSFS

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

$$\mu_\Delta = \begin{bmatrix} \mu_1 & \ldots & \mu_L \end{bmatrix}^T \tag{5}$$

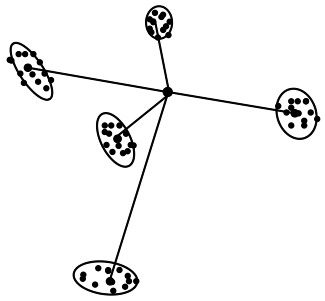


Fig. 3. The sigma points were produced from the unscented transform. The mean and the covariance can be found from the weighted sum of these sigma points.

As each element in $\mu_\Delta$ is a time delay with a variance $\sigma_i^2$. The corresponding covariance matrix is a diagonal matrix of the form:

$$\Sigma_\Delta = \text{diag}\begin{pmatrix}\sigma_1^2 & \dots & \sigma_L^2\end{pmatrix} \tag{6}$$

Now the mean $\mu_\Delta$ and covariance matrix $\Sigma_\Delta$ 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 Eqn. 4. Fig. 3 illustrates the PSFS approach using the unscented transform.

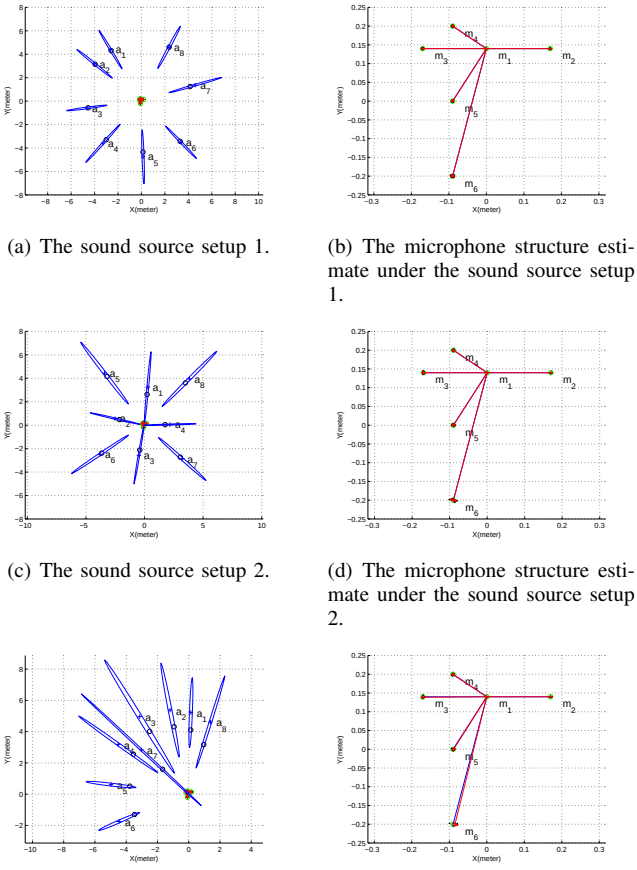
Fig. 4 shows the simulations results of PSFS using the same microphone array with different sound source configurations. Although the performances of sound source localization may depend on sound source configurations, microphone array calibration remains accurate. Noted that the microphone location estimates are more accurate than the sound source location estimates as more constrains are involved with microphones than sound sources in the SFS framework.

## IV. Probabilistic Sound Source Localization

As the PSFS framework may not provide very accurate sound source localization under uncertainty, a probabilistic sound source localization (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 proposed PSSL algorithm applies the unscented transform to utilizes the microphone array estimate and the corresponding uncertainty from PSFS with the measurements.

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

$$\Delta = \begin{bmatrix}\delta_{1,2} & \delta_{1,3} & \delta_{1,4} & \cdots & \delta_{N-1,N}\end{bmatrix}^T \tag{7}$$



(a) The sound source setup 1.

(b) The microphone structure estimate under the sound source setup 1.

(c) The sound source setup 2.

(d) The microphone structure estimate under the sound source setup 2.

(e) The sound source setup 3.

(f) The microphone structure estimate under the sound source setup 3.

Fig. 4. PSFS simulation results with different sound source configurations.

where $\delta_{i,j}$ denotes the time delay between the $i$th microphone and the $j$th 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 $\Delta$ are combined and reformed as:

$$\mu = \begin{bmatrix}x_1 & y_1 & \cdots & x_N & y_N & \delta_{1,2} & \cdots & \delta_{N-1,N}\end{bmatrix}^T \tag{8}$$

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

The covariance matrix of the input can be computed by:

$$\Sigma = \begin{bmatrix}\Sigma_{m_1} & & & & \mathbf{0} \\ & \ddots & & & \\ & & \Sigma_{m_N} & & \\ & & & \sigma_{1,2}^2 & \\ \mathbf{0} & & & & \ddots \\ & & & & & \sigma_{N-1,N}^2\end{bmatrix} \tag{9}$$

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

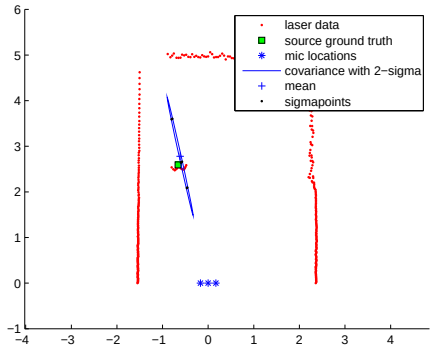



Fig. 5. Experimental result of PSSL using three microphones.

of the form:

$$\chi_i = \left[x_1^{(l)} \quad y_1^{(l)} \quad \cdots \quad x_N^{(l)} \quad y_N^{(l)} \quad \delta_{1,2}^{(l)} \quad \cdots \quad \delta_{N-1,N}^{(l)}\right]^T \quad \text{where} \quad l = 0,\ldots,2L \tag{10}$$

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

$$A^{(l)*} = \text{argmin}_A \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left\{ \left| \begin{pmatrix} x_i^{(l)} \\ y_i^{(l)} \end{pmatrix} - \begin{pmatrix} a^{(l)} \\ b^{(l)} \end{pmatrix} \right| - \left| \begin{pmatrix} x_j^{(l)} \\ y_j^{(l)} \end{pmatrix} - \begin{pmatrix} a^{(l)} \\ b^{(l)} \end{pmatrix} \right| - \delta_{i,j}^{(l)} \right\}^2 \quad \text{where} \quad l = 0,\ldots,2L \tag{11}$$

A location of the sound source $A^{(l)} = (a^{(l)}, b^{(l)})$ is found accordingly. The mean $\mu'$ and covariance $\Sigma'$ of the sound source location can be recovered through the standard unscented transform procedure (Eqn. 4). Fig. 5 shows a real experimental result of the proposed PSSL algorithm using three microphones. The microphone 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.

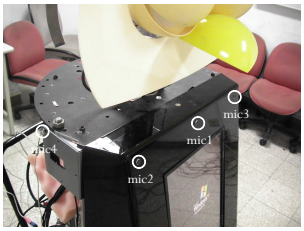
## V. Experimental Results

In this section, the proposed PSFS and PSSL algorithms are evaluated through experiments using real speeches collected from a person. Fig. 6 shows the experiment setup in which a 8 channel A/D board is used to collect sound source data and six low-cost, off-the-shell 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 is to calibrate the microphone array using the proposed PSFS algorithm and the other is to localize the sound source with the calibrated microphone array using the proposed PSSL algorithm. We further demonstrate PSSL with a moving microphone array.

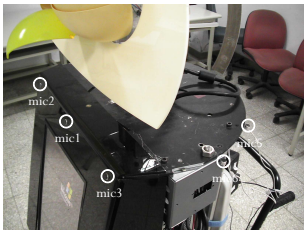
### A. Time Delay Estimation

In each experiment, 10 seconds of speech data at different locations were collected from six microphones. All sound source signals were sampled at 44.1 kHz. The time delay

(a) The NTU-PAL2 robot, a person and a SICK S200 laser scanner.

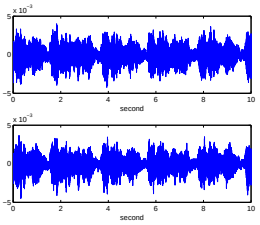



(b) The microphone positions.

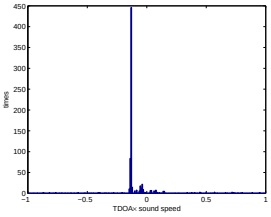



(c) The microphone positions.

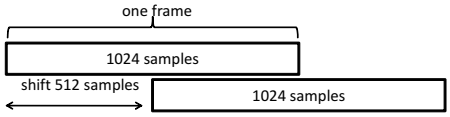
Fig. 6. The experiment setup.




(a) The waveforms of 10s speech from two microphones




(b) The histogram of TDOA of 10s speech.



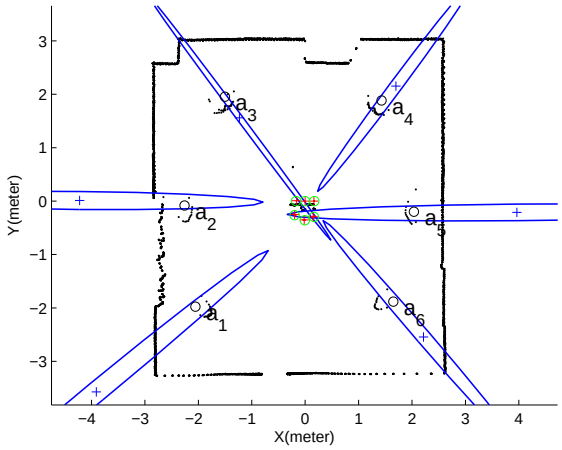

(c) The overlap setting.

Fig. 7. The example of the TDOA estimation.

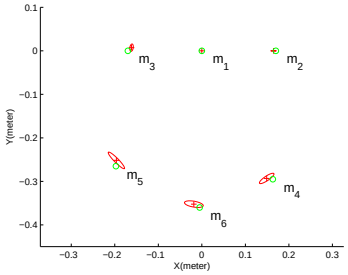
of arrival (TDOA) estimation was performed using 1024 samples and 512 samples was shifted at the next frame. The generalized cross-correlation approach [11] is utilized to estimate time delays between microphones. As there are silent segments in these speeches, the TDOA estimates may be unstable. The peak of the histogram of the TDOA estimates of 10s speech was chosen as the input of PSFS or PSSL. Fig. 7 illustrates the approach to estimate time delay between microphones.

### B. PSFS Results

The PSFS experiments were conducted in two different environments in terms of environment size. The first ex-

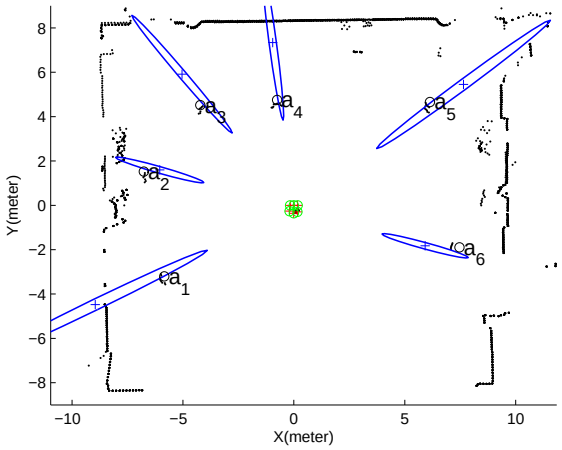



(a) The sound source estimates

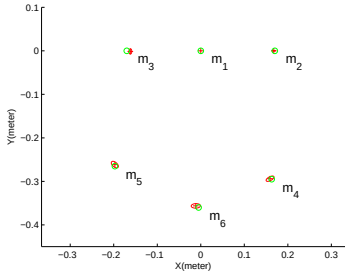



(b) The microphone structure estimates

Fig. 8. The near field situation. The result of the PSFS experiment was performed in a seminar room. The blue circles show the ground truth of the microphone locations and the blue squares show the true locations of the sound sources.




(a) The sound source estimates




(b) The microphone structure estimates

Fig. 9. The far-field situation. The results of the PSFS experiment performed in the atrium. The blue circles show the ground truth of the microphone locations and the blue squares show the ground truth of the sound sources.

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

*1) The Near Field Condition:* The experiment was conducted in the seminar room and the sound sources were away from the microphones about 2-3 meters. Six speeches were collected at different locations. Fig. 8 shows the experimental results of PSFS with 6 sound sources. The average angular error is 0.35 degree. The average microphone location error is 0.0081 m. The average sound source location error is 0.75 m.

*2) The Far Field Condition:* The experiment was conducted in the atrium and the sound sources were away from the microphones about 6-8 meters. Six speeches were collected at different locations. Fig. 9 shows the experimental results of PSFS with 6 sound sources. The average angular error is 2.0245 degree. The average microphone location error is 0.0041 m. The average sound source location error is 7.0773 m.

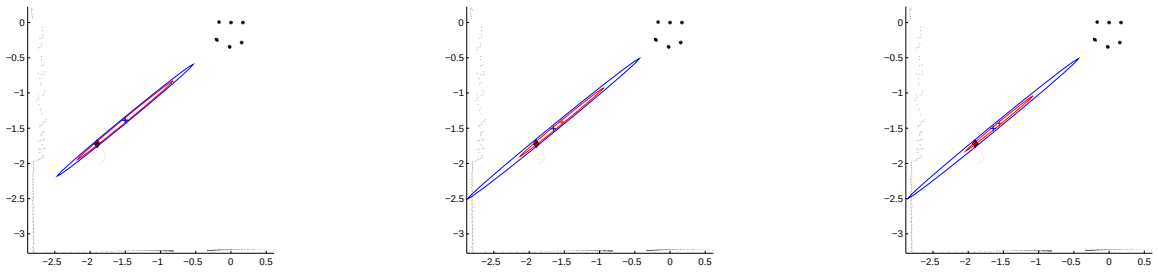
### C. PSSL Results

With the PSFS results, two experiments of PSSL were conducted. One is to localize the sound source using the static robot with more measurements. Fig. 10 shows the PSSL results in a series of measurement updates. It is shown that the estimates are more accurate and certain with more measurement updates.

The other experiment is to localize the sound source using the moving robot. Fig. 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, the similar performance can be achieved using inexpensive odometry.

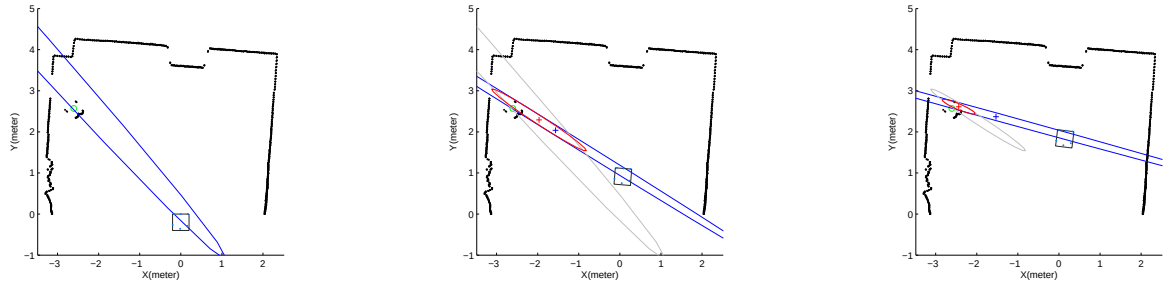
## VI. Conclusion

Microphone arrays are widely used for auditory perception. However, microphone array calibration could 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



(a) Update 1 (b) Update 2 (c) Update 3

Fig. 10. The PSSL results with a series of measurement updates. Blue ellipse shows the estimate uncertainty ($2\sigma$ ) of the PSSL result. Red ellipse shows the estimate uncertainty ($2\sigma$) of the PSSL result after the measurement update.




(a) Update 1 (b) Update 2 (c) Update 3

Fig. 11. The PSSL result with a moving robot. Green circle shows the true location of the sound source. Blue ellipse shows the estimate uncertainty ($2\sigma$ ) of the PSSL result. Red ellipse shows the estimate uncertainty ($2\sigma$) of the PSSL result after the measurement update.

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 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 this issues. In addition, applying particle filters to PSFS and PSSL should be of our interest.

## ACKNOWLEDGMENTS

The authors would like to thank Jwu-Sheng Hu for fruitful discussions on the work. This work was partially supported by grants from Taiwan NSC (#96-2628-E-002-251-MY3, #96-2218-E-002-035, #96-2218-E-002-008); Excellent Research Projects of National Taiwan University (#95R0062-AE00-05); Taiwan DOIT TDPA Program (#95-EC-17-A-04-S1-054); Taiwan ITRI, III; and Intel.

## REFERENCES

[1] S. T. Birchfield and R. Gangishetty, "Acoustic localization by interaural level difference," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[2] W. Cui, Z. Cao, and J. Wei, "Dual-microphone source location method in 2-d space," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

[3] K. C. Ho and M. Sun, "Passive source localization using time differences of arrival and gain ratios of arrival," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 464–477, February 2008.

[4] J.-S. Hu, C.-C. Cheng, and W.-H. Liu, "Robust speaker's location detection in a vehicle environment using gmm models," *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics*, vol. 36, no. 2, pp. 403–412, April 2006.

[5] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, March 2007.

[6] 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 Transactions on Robotics*, vol. 23, no. 4, pp. 742–752, August 2007.

[7] E. B. Martinson, "Acoustical awareness for intelligent robotic action," Ph.D. dissertation, Georgia Institute of Technology, November 2007.

[8] S. Thrun, "Affine structure from sound," in *Proceedings of Conference on Neural Information Processing Systems (NIPS)*. Cambridge, MA: MIT Press, 2005.

[9] S. Julier and J. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[10] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *International Journal of Computer Vision*, vol. 9, no. 2, pp. 137–154, February 1992.

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