Evaluating Real-time Audio Localization Algorithms for Artificial Audition in Robotics

Anthony Badali* , Jean-Marc Valin†, François Michaud‡, and Parham Aarabi*

*University of Toronto
Dept. of Electrical &
Computer Engineering

†Octasic Semiconductor

‡Université de Sherbrooke
Dept. of Electrical &
Computer Engineering

***Abstract*— Although research on localization of sound sources using microphone arrays has been carried out for years, providing such capabilities on robots is rather new. Artificial audition systems on robots currently exist, but no evaluation of the methods used to localize sound sources has yet been conducted. This paper presents an evaluation of various real-time audio localization algorithms using a medium-sized microphone array which is suitable for applications in robotics. The techniques studied here are implementations and enhancements of steered response power - phase transform beamformers, which represent the most popular methods for time difference of arrival audio localization. In addition, two different grid topologies for implementing source direction search are also compared. Results show that a direction refinement procedure can be used to improve localization accuracy and that more efficient and accurate direction searches can be performed using a uniform triangular element grid rather than the typical rectangular element grid.**

***Index Terms*— Localization, Beamformer, Direction search, Robot sensing systems**

I. Introduction

The localization of sound sources using microphone arrays is a well studied problem [1][2][3][4]. Some of the most common applications of this technology include intelligent environments and teleconferencing. Recently, microphone arrays have also become popular within the area of robotics where they have been employed to track users [5][6][7] and as the basis for speech interfaces [8][9]. For example, in [6] a microphone array is used to simultaneously track multiple users interacting with the Sparticus robot. In [8] the Honda Asimo robot is used as a referee for rock-paper-scissors sound games. Auditory systems significantly enhance the interaction between robots and humans, resulting in much more natural and intuitive experiences for the users. In systems with speech interfaces, the ability to localize speakers in the environment is crucial [10]. For example, the performance of automatic speech recognition systems can be significantly improved when speaker position is known [11][12].

Within the domain of artificial audition for robotics, localization must be performed with limited processing power, thus the implementations studied are computationally efficient enough to be executed in real-time on general purpose processors. Audio localization techniques are generally based on time delay of arrival estimation (TDOA) and delay-and-sum beamforming (SRP, for Steered Response Power of a beamformer). These techniques are particularly appealing because they can be easily implemented to execute in real-time [1][5][13][14]. However, with the existence of different sound source localization methods and because robotic applications have intrinsic integration issues (e.g., real-time performance, mobile base, changing conditions) that first had to be addressed to demonstrate feasibility, working systems do not yet present clear demonstrations that the methods they use are the best ones.

Therefore, this paper investigates the accuracy of different TDOA audio localization implementations in the context of artificial audition for robotic systems. The experiments are performed using an array which has been used with mobile robotic platforms [5][6][8][9][13][15] and can be mounted on a wide range of medium to large sized robots. Algorithms considered here include the standard SRP-PHAT (Phase Transform) beamformer [2], enhancements developed by Valin et al. [13][15], and a simplification of the SRP-PHAT algorithm used in [14]. Two alternative topologies for direction search grids are also compared. The main contribution of this work is an empirical evaluation of these algorithms and search grid topologies, to outline which one works best and under which conditions to provide artificial audition on robotic systems.

The paper is organized as follows. Section II provides some background on TDOA estimation and sound localization. Section III describes the experiments and implementation details. The results are presented in Section IV, and Section V concludes the paper.

II. Sound Localization Background

The localization algorithms considered here are all based on TDOA estimation using modifications of a standard SRP-PHAT beamformer. The main elements of these techniques are time difference estimation and direction search. This section presents background into these aspects of the localization procedure and describes variations of the SRP-PHAT algorithm which are studied in these experiments.

A. TDOA Estimation

Sound localization is commonly performed by TDOA techniques. Using the observed time differences between audio signals arriving at a pair of microphones, the position of a speaker can be constrained to lie on a hyperboloid in

space. A point estimate of speaker position can be computed by solving the intersection of multiple hyperboloids from different microphone pairs at known positions. The generalized cross-correlation (GCC) is one of the most popular TDOA estimation algorithms [16]. Denoting the Fourier transform signal received at microphone i as $X_i(\omega)$, the GCC estimate $\hat{\tau}$ between microphone i and j can be computed as,

$$\hat{\tau} = \arg\max_{\beta} \int_{\omega} W(\omega) X_i(\omega) \overline{X_j(\omega)} e^{-j\omega\beta} d\omega \qquad (1)$$

where $W(\omega)$ defines a weighting function which is commonly selected to be the PHAT given by

$$W_{PHAT}(\omega) = \frac{1}{|X_i(\omega)||X_j(\omega)|} \qquad (2)$$

The PHAT is popular for sound localization due to its robustness in noisy and reverberant environments [17].

B. Beamforming Search

In general, localization can be performed by applying iterative re-weighted least squares to solve for the speaker position [3]. However, noise in the TDOA estimates can cause the system to be unstable, leading to poor solutions [15]. The most common and successful audio localization techniques are based on the steered response power (SRP) or beamformer energy [4]. Using the GCC, a likelihood L is assigned to each position x as follows,

$$L(x) = \sum_{i<j} \int_{\omega} W(\omega) X_i(\omega) \overline{X_j(\omega)} e^{-j\omega\tau_{ij}(x)} d\omega \qquad (3)$$

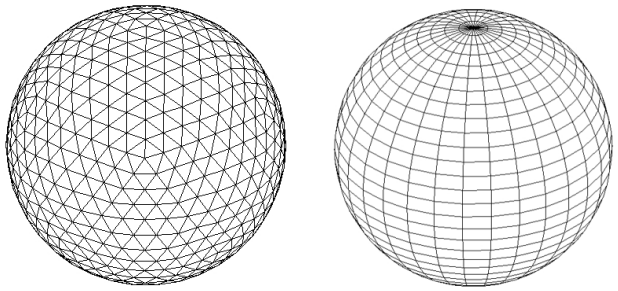
where the i and j index over all microphone pairs and $\tau_{ij}(x)$ denotes the TDOA between microphones i and j corresponding to a source at position x. This function is often referred to as the spatial likelihood function (SLF) [1]. Intuitively, if a source is located at position $\hat{x}$ then each integral term in (3) should be maximized at $\tau_{ij}(\hat{x})$, yielding a maximal likelihood at $L(\hat{x})$. In practice, noise introduces errors in the estimates of time delays, in which case the beamformer in (3) is much more robust than the simple GCC estimate in (1). Real-time position estimation can be achieved by using discrete search-based techniques [1][13][15][18]. These algorithms can be efficiently implemented by precomputing the expected time delays $\tau_{ij}(x)$ at a set of source positions on a grid. The position with the largest total beamformer energy summed across all pairs is selected as the speakers location [5][13][18]. The above procedure is used in conjunction with the PHAT frequency weighting function, and is commonly referred to as the SRP-PHAT technique. All techniques considered here will use a PHAT weighting function, thus it will not be explicitly mentioned when describing the algorithms.

C. Far-field Assumption

For robotic applications in unconstrained environments, a search within a large 3D grid can become computationally expensive. It is also difficult to accurately estimate position when the distance to the sound source is larger than the microphone array dimensions. As an alternative, the far-field approximation [15] can be used to estimate the direction of the sound source. This reduces the search space to 2-dimensions and avoids making unreliable estimates of speaker distance. Under the assumption that the direction of the sound source is the same for all microphones, the time difference of arrival between each microphone pair can be approximated by

$$\tau_{ij} = \frac{1}{c} (\vec{p}_i - \vec{p}_j) \cdot \vec{u} \qquad (4)$$

where p_i and p_j are the position vectors of microphone i and j respectively, $\vec{u}$ is the source direction vector, and c is the speed of sound in air. An empirical evaluation using a rectangular microphone array found a mean approximation error under 4° for sources at distances comparable to the array dimensions which converge below 1° for larger distances [15]. This makes the far-field assumption particularly appropriate for small arrays which are found on many robotic platforms.



(a.) Triangular Mesh (b.) Rectangular Mesh

Fig. 1. Spherical Search Grids, a.) Triangular sampling from recursive icosahedron subdivision (3 levels). b.) 20 x 20 spherical tessellation with rectangular sampling.

D. Localization Algorithms

This evaluation seeks to measure the performance differences between several variations of the SRP-PHAT algorithm which can be used for real-time localization on a robot platform. Four different techniques are evaluated:

1) *PEAK*: This method is a simplified beamformer using only the maxima (or peak) of the GCC between microphone pairs to compute the most likely speaker position. This is similar to the method implemented in [14] and is presented as a computationally simpler alternative to the standard beamformer.
2) *SRP*: This method corresponds to the standard beamformer as described above. The most likely speaker position is selected using (3).
3) *SW (Spectral Weighting)*: This method adaptively modifies the GCC weighting function in an attempt to assign larger weights to frequency components which have a higher signal-to-noise ratio [13]. The most likely speaker position is selected using (3). However,

additional terms are multiplied with the weighting function $W(\omega)$.

4) *DR (Direction Refinement)*: This method applies a direction estimate refinement procedure after localization in an attempt to improve accuracy [14]. This is achieved by executing a local high-resolution search without using the far-field assumption. A far-field localization algorithm such as SRP or PEAK is first executed to find an initial direction estimate. A local search grid is then centered at the estimated direction. Since the far-field assumption is not used, the time delays are a function of speaker distance and the search must be performed across both source direction and distance. However, based on the observation that the time delays quickly approach the far-field approximation as a function of speaker distance, only a few nearby distances are searched. The search distances are manually specified and fixed.

Both SW and DR are enhancements which can be used in conjunction with other techniques (PEAK, SRP) and several different combinations are considered in these experiments. Two different search grid topologies are also tested, as shown in Fig. 1:

- A spherical rectangular element grid (R) sampled in uniform degree increments.
- A triangular element grid (T) sampled at uniform distances along the surface of the sphere. The latter grid is generated by performing recursive icosahedron subdivision [13].

III. Experimental Setup

A cubical 8 microphone array from [5][6][13][15] was used for our experiments, and is shown in Fig. 2. The array dimensions are 32 cm by 32 cm by 36 cm with a single microphone placed at each vertex. Audio input was acquired using a National Instruments PCI-6071E data acquisition card which provides hardware synchronized channels to ensure accurate TDOA estimation.

The experiments were carried out for range of different source positions which are appropriate in the context of a small to medium sized robot platform. The task was localization of a single fixed source which was generated by playing prerecorded speech sequences. The test positions were placed at heights which were at level and above the array center to correspond to humans interacting with a medium to large sized robot. An overhead sketch with labeled test positions is shown in Fig. 3. At each position, the SNR (signal-to-noise ratio) was also varied by adjusting the volume of the audio playback. Experiments were performed in a laboratory setting with a reverberation time of 0.1 seconds. Two sets of experiments were conducted:

- Experiment 1 was performed over all five positions shown in Fig. 3 with only stationary background noise affecting localization accuracy. A histogram of samples from the background noise signal is shown in Fig. 4. The noise distribution is unimodal and appears roughly Gaussian.
- Experiment 2 was carried out by placing a noise source at position 5 and performing localization on sources at positions 1, 2, and 3. Complex, non-stationary noise was generated by playing vocal and instrumental classical music.

All the techniques were implemented in Matlab which was used as a common platform to ensure any performance differences were due to the algorithms and not different implementation platforms. The test signals were sampled at 22 kHz and localization was executed on 25 ms windows with 50% overlap. To be consistent with a real-time implementation, the time delays for each position were pre-computed in a lookup table and rounded to units of samples to allow efficient indexing and reduce the number of cross-correlations computed. The rectangular grid was tessellated at a resolution of 60x60 and the triangular grid

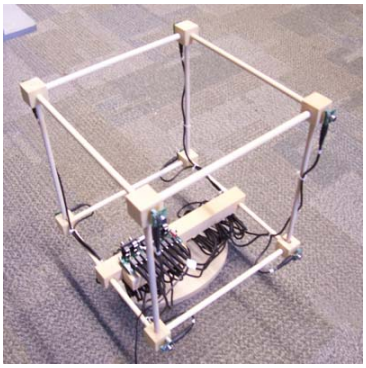


Fig. 2. The cubical microphone array used in the experiments. The array was elevated approximately 45 cm from the ground during experiments.

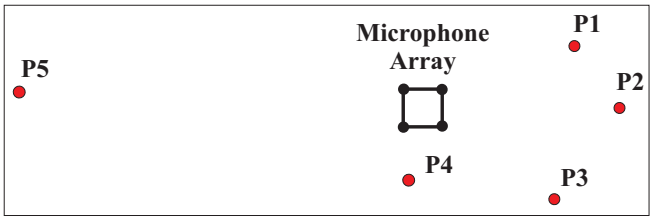


Fig. 3. A top view sketch of speaker positions used in the experiment.

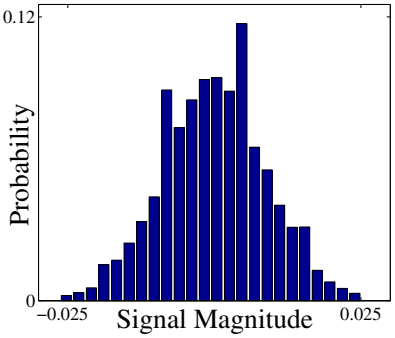


Fig. 4. Histogram of the background noise signal.

was computed by recursively sub-dividing an icosahedron to four levels. These results in direction searches over 3600 and 2562 points respectively. These particular resolutions were selected as they can both be performed in real-time using general-purpose processors [1][13]. Using icosahedron subdivision, the resolution of the triangular grid can only change by factors of four, thus the next larger grid becomes computationally expensive to search. The spectral weighting procedure also has many parameters related to estimating the noise spectrum. In general, parameter values were set in accordance with [13] and any parameters not specified there were chosen as specified in [19]. The direction refinement procedure is computed using a 196 element local rectangular grid which ranges in both directions from $-1.5\,^\circ$ to $1.5\,^\circ$ in increments of $0.5\,^\circ$ and is computed at distances of 0.5, 1.5, 3.0, and 5.0 m from the array center. The time differences were quantized to units of 0.5 samples for this smaller grid.

IV. RESULTS

The results of the experiments are shown in Tables I and II and Figures 5 and 6. The implementations using search grids created with rectangular patches are denoted with (R) and those with triangular patches with (T). The mean error is computed as the angle between the ground truth and estimated direction. To avoid large biases introduced by large errors which occur when a segment does not contain any samples from the speaker, segments with an error greater than $30\,^\circ$ are discarded from the calculation of mean error. The percent anomalies are also computed for each technique as the percentage of segments which had a localization error greater than $10\,^\circ$.

The results in Tables I and II are averaged across both SNR and speaker position. The high percentage of anomalies can be attributed to the small time windows used for localization. This value decreases significantly as window size is increased. In a separate analysis (not described here) which used the data from Experiment 1, the percent anomalies all dropped below 10% at window sizes of 100 ms, while relative performance of each algorithm stayed the same. The results in Fig. 5 depict mean error as a function of SNR, where each data point is computed by averaging results across all positions at a given SNR. In practice, the signal-to-noise ratios are slightly different across positions and an approximate SNR is computed by averaging the individual SNR estimates. Similarly, Fig. 6 show the percentage anomalies as a function of approximate SNR.

For Experiment 1, the simplification of using only maximum peaks of the GCC for localization (PEAK) performs worse than the others in terms of mean error across all SNR values (Fig. 5 left). This result demonstrates the benefits of using a complete delay-and-sum beamformer (SRP), and we observe that both under simple and complex noise conditions the beamformer improves accuracy. The spectral weighting (SW) procedure generally did not improve the localization accuracy, and the SRP beamformer performed better across all SNR values. The decrease in performance can be attributed to difficulty estimating the noise spectrum, where the signal spectrum is mistakenly being estimated as noise spectrum. Also, given the wideband nature of this noise, we do not expect a frequency re-weighting algorithm to improve performance. The higher percent anomalies of the SW techniques also suggest signal spectrum being mistaking estimated as noise.

The direction refinement was applied to the both the rectangular and triangular search grids and had a significant effect on localization accuracy. In both cases, the mean errors were at least as good in all experiments. Again, the results for the DR procedure applied to the PEAK technique are not shown here as the effect will be similar to the SRP algorithm but with a lower accuracy. In general, the improvement in mean error from using DR is limited by the angular range of the local search grid, and for this reason the DR procedure does not significantly reduce the number of anomalies. In these experiments, differences in the percent anomalies can be attributed to frames which have localization errors around $10\,^\circ$ for which the refinement is adapting the direction estimates to be slightly inside or slightly outside the $10\,^\circ$ threshold used to classify anomalies. Overall, the best performer in this experiment is the SRP+DR(T) technique.

For Experiment 2, despite significantly different sources of noise, the relative performance of the techniques was similar to Experiment 1. However, the performance of each individual algorithm does not always decrease with approximate SNR in Experiment 2. This is because noise in this experiment is highly dynamic and there is a large variance in SNR during the audio playback. To maximize the different types of interactions between the noise and source signals, they were not synchronized across experiments, but this caused the SNR at key points during playback to vary significantly from the average SNR. Regardless of this effect, the relative performance of each technique remained consistent. The mean error performance of spectral weighting (SW) improves in Experiment 2 because the noise source is colored, however this improvement only appears at higher SNR when the noise spectrum and signal spectrum can be more easily distinguished. The dynamic nature of the music makes spectrum estimation particular difficult in the lower SNR tests. Once again, SRP+DR(T) has the best results averaged across position and SNR (Table II) and although it's accuracy is slightly lower than SR+DR+SW(T) in the higher SNR tests (Fig. 5 right), it still performs the best overall.

It is also important to note that none of the test positions were at the distances used in the direction refinement search, and the results clearly show that the DR procedure improves accuracy even in these cases.

With respect to search grid, we observe that the triangular (T) search grid outperforms it's rectangular (R) counterpart in both Experiments. For all the techniques (PEAK, SRP, SRP+DR), the triangular mesh implementations perform better in terms of mean error percent anomalies across all SNR values and in terms of the overall averages. Although not shown here, the triangular search grid will improve accuracy of the PEAK procedure in a similar manner to

TABLE I

AVERAGE RESULTS FOR EXPERIMENT 1

	PEAK(R)	PEAK(T)	SRP(R)	SRP(T)	SRP+DR(R)	SRP+DR(T)	SRP+SW(T)	SRP+SW+DR(T)
Mean Error (deg)	5.16	3.61	3.93	2.79	2.54	2.03	3.37	2.55
Percent Anomalies	31	30.5	28.1	27.1	28.1	27.2	33.4	33.4

TABLE II

AVERAGE RESULTS FOR EXPERIMENT 2

	PEAK(R)	PEAK(T)	SRP(R)	SRP(T)	SRP+DR(R)	SRP+DR(T)	SRP+SW(T)	SRP+SW+DR(T)
Mean Error (deg)	6.26	5.29	4.29	3.55	3.49	3.18	3.87	3.42
Percent Anomalies	48.5	48.4	46.7	46.4	46.7	46.4	48.5	48.5

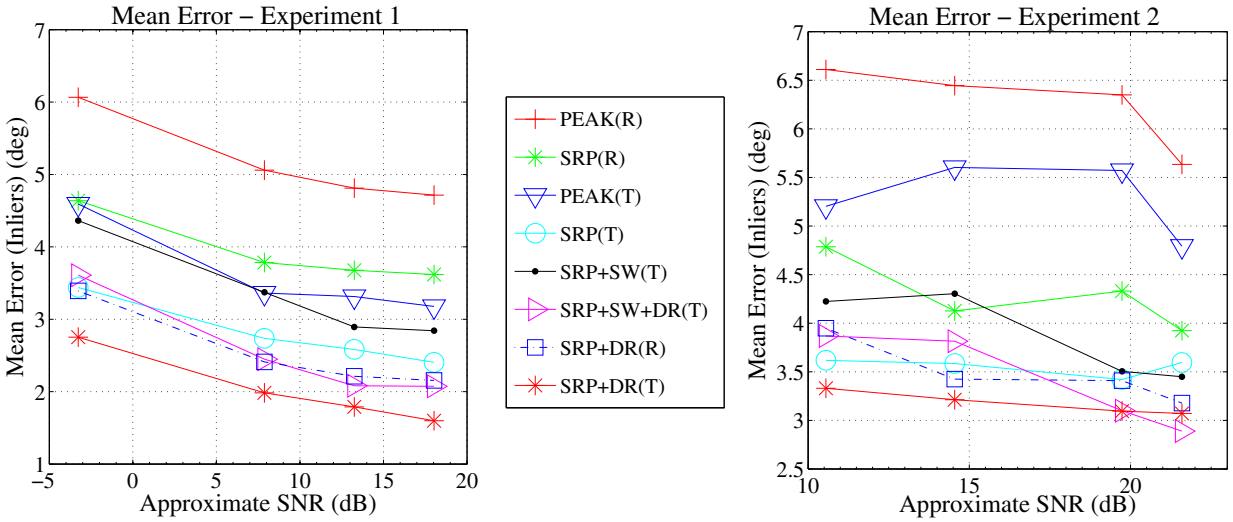


Fig. 5. Mean Localization Error Results averaged across all positions for Experiments 1 and 2.

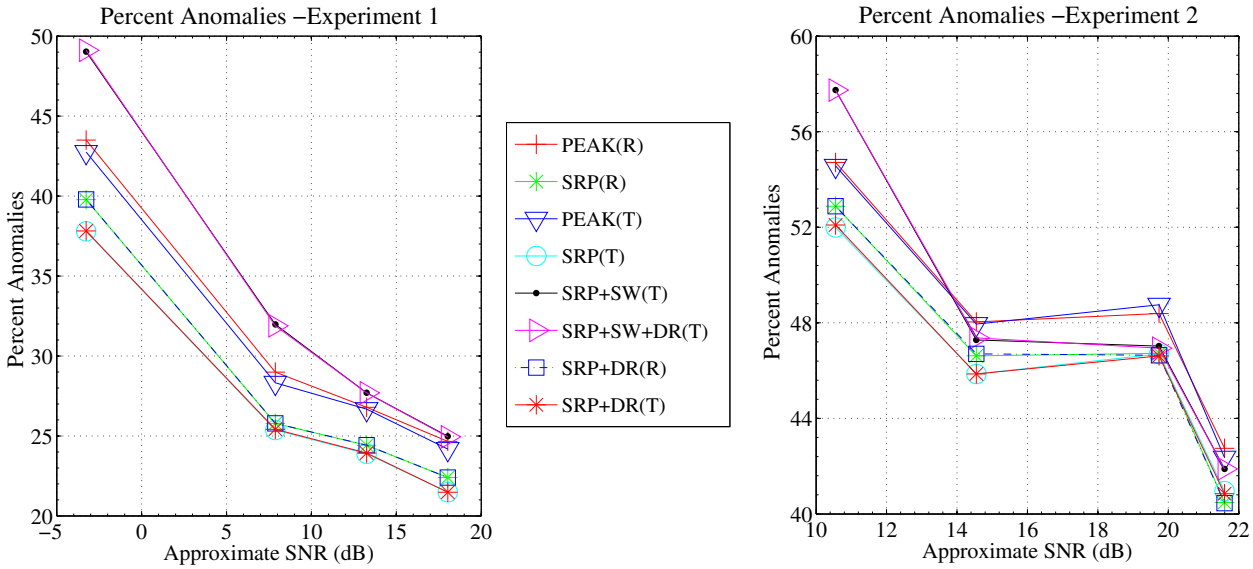


Fig. 6. Percentage Anomaly Results averaged across all positions for Experiments 1 and 2.

the other techniques. The rectangular grid, which is sampled uniformly in angle, has a large density of points in a very small neighborhood around the poles. This significantly reduces sampling density across the rest of the sphere. In this analysis, poles were located at positions which are not likely for sources detected by a (mobile) robot (directly above and below the array). Although the poles can be oriented in a direction where speakers are most likely to be located, the

dense sampling at these points requires the units of time delays to be quantized much more finely because search points are so close together. This would significantly increase the computation time because the GCC must be computed for many more time delays. It would also be inefficient to have such a fine resolution for the more widely spaced search points which make up most of the rectangular mesh.

V. Conclusion

This paper presents an evaluation of various implementations and modifications of real-time SRP-PHAT based localization systems. Results suggest that the direction refinement procedure presented in [13][5] (SRP+DR) improves localization accuracy, even when the source is not at the radii used for the direction search. It should be noted that this procedure requires additional computation as a second search is performed and additional GCC sums need to be computed, but it is still possible to reach real-time performance as the local grids are relatively small. In addition, the triangular-patch search topology yielded higher accuracy than the rectangular patch topology for all algorithms. Uniform distances between search directions are also more appropriate for computing quantized TDOA lookup tables used to perform quick direction searches. For the algorithms considered here, a standard SRP-PHAT beamformer using the direction refinement procedure (DR) with no spectral weighting (SW) and a (isotropic) triangular patch search grid is the best solution for real-time audio localization.

Future research related to these experiments includes a theoretical analysis of the effects of search grid resolution between the triangular and rectangular tessellations. The results obtained from the direction refinement procedure also suggest that a coarse-to-fine search is worth investigating. This may potentially improve both speed and accuracy of the localization procedures. With respect to artificial audition for robotic systems, future work includes evaluating the robustness of localization methods in complex situations involving simultaneous tracking of multiple users or relative motion between the robot and sound sources.

VI. Acknowledgments

François Michaud holds the Canada Research Chair on Mobile Robotics and Autonomous Intelligent Systems and Parham Aarabi holds Canada Research Chair in Internet Video, Audio, and Image Search. This project is funded by the Natural Sciences and Engineering Research Council of Canada, through its NSERC/Canada Council for the Arts New Media Initiative.

References

[1] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," *EURASIP Journal of Applied Signal Processing*, vol. 2003, pp. 338–347, 2003.

[2] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 1997, pp. 375–378.

[3] F. Gustafsson and F. Gunnarsson, "Positioning using time-difference of arrival measurements," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[4] J. DiBiase, H. Silverman, and M. Brandstein, *Microphone Arrays: Signal Processing Techniques and Applications*, chapter Robust Localization in Reverberant Rooms, pp. 216–228, Springer-Verilag, 2007.

[5] J-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robot. Auton. Syst.*, vol. 55, pp. 216–228, 2007.

[6] F. Michaud, C. Côté, D. Létourneau, Y. Brosseau, J.M. Valin, E. Beaudry, C. Raïevsky, A. Ponchon, P. Moisan, P. Lepage, Y. Morin, F. Gagnon, P. Giguère, M.A. Roux, S. Caron, P. Frenette, and F. Kabanza, "Spartacus attending the 2005 AAAI conference," *Auton. Robots*, vol. 22, no. 4, pp. 369–383, 2007.

[7] Y. Tamai, S. Kagami, Y. Amemiya, Y. Sasaki, H. Mizoguchi, and T. Takano, "Circular microphone array for robot's audition," in *Proceedings of IEEE Sensors*, Oct. 2004, pp. 565–570.

[8] K. Nakadai, S. Yamamoto, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "A robot referee for rock-paper-scissors sound games," in *Proceedings of IEEE International Conference on Robotics and Automation ICRA*, Pasadena, CA,, May 2008, pp. 3469–3474.

[9] Kazuhiro Nakadai, Hiroshi G. Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino, "An open source software system for robot audition HARK and its evaluation," in *Proceedings of 8th IEEE-RAS International Conference on Humanoid Robots*, Daejeon, Korea (South), Dec. 2008, pp. 561–566.

[10] D. Giuliani, M. Omologo, and P. Svaizer, "Talker localization and speech recognition using a microphone array and a cross-powerspectrum phase analysis," in *Proceedings International Conference on Spoken Language Processing (ICSLP)*, 1994, pp. 1243–1246.

[11] T.B. Hughes, Hong-Seok Kim, J.H. DiBiase, and H.F. Silverman, "Using a real-time, tracking microphone array as input to an HMM speech recognizer," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 249–252 vol.1, May 1998.

[12] I.A. McCowan, C. Marro, and L. Mauuary, "Robust speech recognition using near-field superdirective beamforming with post-filtering," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '00.*, vol. 3, pp. 1723–1726 vol.3, 2000.

[13] J-M. Valin, *Auditory System for a Mobile Robot*, Ph.D. thesis, Department of Electrical & Computer Engineering Université de Sherbrooke, 2005.

[14] D. Halupka, N.J. Mathai, P. Aarabi, and A. Sheikholeslami, "Robust sound localization in 0.18um CMOS," *IEEE Transactions on Systems, Man, and Cybernetics, Part B,*, vol. 53, pp. 2243–2250, 2005.

[15] J.-M. Valin, F. Michaud, J. Rouat, and D. Letourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proceedings International Conference on Intelligent Robots and Systems*, 2003, pp. 1228–1233.

[16] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, Aug 1976.

[17] T. Gustafsson, B. D. Rao, and M. Trivedi, "Analysis of time-delay estimation in reverberant environments," in *Proceedings of the International Conference on Spoken Language Processing*, 2003, vol. 2, pp. 573–576.

[18] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 34, no. 3, pp. 1526–1540, June 2004.

[19] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, 2002.