

Auditory Evidence Grids

Eric Martinson

Naval Research Laboratory
Georgia Institute of Technology
ebeowulf@cc.gatech.edu

Alan Schultz

U.S. Naval Research Lab
schultz@aic.nrl.navy.mil

Abstract – Sound source localization on a mobile robot can be a difficult task due to a variety of problems inherent to a real environment, including robot ego-noise, echoes, and the transient nature of ambient noise. As a result, source localization data are often very noisy and unreliable. In this work, we overcome some of these problems by combining the localization evidence over a variety of robot poses using an evidence grid. The result is a representation that localizes the pertinent objects well over time, can be used to filter poor localization results, and may also be useful for global re-localization from sound localization results.

Index Terms – Sound Source Localization, Evidence Grid, Mobile Robots, Auditory Mapping.

I. INTRODUCTION

Sound source localization algorithms have long used the concept of time-difference on arrival (TDOA) between microphones to try and find the position of the sound source in the environment. Noise, however, remains a problem with these TDOA based solutions. Echoes from the surrounding walls and other hard surfaces in the environment are a good example of this noise. From a stationary position, echoes can be misleading, appearing to come from mirror image sources located behind the walls off which they have been generated. Another common example is environmental noise from fans, electronics, and motors that are not the primary target of sound source localization. The source separation problem is always difficult, but especially so when the environmental sources are located in close proximity to the microphones.

Using a stationary microphone array, these problems of echoes and environmental noise can often be very difficult to overcome, requiring sizeable filters and much customization to



Figure 1. The robot is localizing a human speaker by their speech. The microphone array used in this task can be seen mounted at the highest point on the robot..

the particular environment. A microphone array mounted on a mobile robot, however, is not necessarily as restricted, because there is more information available for filtering out extraneous noise, namely the robot pose. As a robot equipped with a microphone array moves through the environment, only the global position of stationary sound sources will appear to remain in the same place over time. In comparison, echoes from the surrounding walls will appear to come not from a single location but instead from seemingly random locations as the angle of reflection off the walls changes with the changing robot position. The same is true of robot ego-noise (motor, wheel, fans, etc), which can be particularly disruptive to a sound source localization algorithm because of their proximity to the robot. While the robot moves about the environment, ego-noise will appear to travel with the robot instead of appearing to come from a single location, thus allowing much of it to be filtered out.

To use this mobility-inspired advantage for localizing stationary sound sources, we combine the sound localization results with pose information of the robot (Figure 1) acquired via laser-based localization techniques. The results are maps of the global soundscape. The algorithmic technique used to create these auditory maps is that of evidence grids [1, 2].

II. RELATED WORK

The notion of an evidence grid has been around for many years in the context of map building, although not yet applied to the problem of sound source localization. Most typically, a suite of range finding sensors (such as sonar or laser) are used to acquire evidence about walls and other obstacles distributed about an environment. An evidence grid then combines this data from a set of disparate, separated sensors to create a map of an indoor or outdoor environment[3].

While this work will be using evidence grids, there are other existing approaches in robotics to mapping out some form of the acoustic landscape. Noise Mapping[4] uses sound pressure level measurements taken by a mobile robot as it traverses the environment to construct maps of ambient noise levels throughout the environment. Peaks in the noise level map are then indicative of sound sources in the area, although other features, such as excessive robot noise and echoic locations, could also generate such peaks. Auditory evidence grids, by comparison, are less affected by robot ego-noise and echoic environments than noise maps. Furthermore, provided the acoustic localization algorithm can detect the sources, the soundscape can be mapped with many fewer samples taken over a smaller area of the environment.

Besides those approaches that create explicit maps, a number of others also take advantage of movement to improve

sound localization results. Nakadai et al[5] physically rotate their robot so as to separate out robot ego-noise from other sound sources with stationary global positions. Even without physical movement, an array of microphones can sometimes do the same thing by mathematically steering the direction of attention until hitting a peak in correlated noise [6].

III. ALGORITHMIC FOUNDATIONS

The approach taken in this work to map out sound sources in the environment is performed in two separate phases. The first phase estimates the location of the sound source detected in a single sample of measured data. The second phase then converts those data to probabilistic representations and updates the evidence grid.

A. Sound Source Localization

The most common sound source localization algorithms are based on the physics of sound propagation through an environment. If two microphones are located some distance from each other, then the signal received by each microphone due to a single source will be offset by some measurable time. If the value of this time difference between the two received signals can be determined, then the possible positions of the sound source will be constrained to all positions in the room whose geometrical position relative to the array corresponds to a measured time difference.

Predicting what the time difference should be for a particular location (L) in the environment is the simplest part. If the speed of sound (c) is assumed constant (343 m/s), then the time required to travel from a source at L to a microphone is the distance traveled divided by the speed of sound:

$$T(l, m) = d_{l,m} / c \quad (1)$$

To speed up the general computational speed of the algorithm, this value is predetermined for all microphone/location pairs at compile time for a set of 400 evenly spaced grid points in a $6 \times 6 \text{m}^2$ array around the center of the microphone array. The positions could be estimated over the entire map, instead of only for a small subset, but beyond 3 meters the robot is unlikely to distinguish many types of sources from background noise, and the computational time required for the next step limits the total number of grid points that can be checked in real time. The

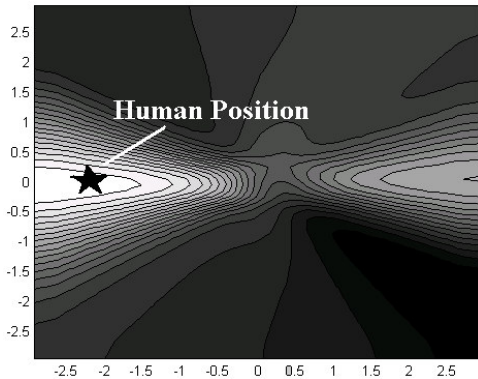


Figure 2. A spatial likelihoods result for detecting human speech. This result demonstrates the common problem of a strong angular performance, but poor distance estimates.

value of 3 meters was determined experimentally using a speech source moving away from the robot.

The algorithm then used for actually estimating sound source positions given these predicted time delays is the spatial likelihood function [7]. Spatial likelihoods are an approach based on maximum likelihood that uses a weighted cross correlation algorithm to estimate the relative energy associated with every possible source location. The general idea is that the resulting cross correlation value, adjusted for the predicted time difference on arrival, will be highest for those position/time differences corresponding most closely with the true value. As this work is using an array of microphones, the cross correlation value is actually determined separately for each microphone pair, and then summed across all microphone pairs for every position:

$$F_l = \sum_{a=1}^N \sum_{b=1}^N \int_{\omega} W(\omega) M_a(\omega) \overline{M_b(\omega)} e^{-j\omega(T(l,a)-T(l,b))} d\omega \quad (2)$$

where (M_a) is the Fourier transform of the of the signal received by microphone (a), $\overline{M_b}$ is the complex conjugate of (M_b), (ω) is the frequency in [rad/s], and (W) is a frequency dependant weighting function:

$$W(\omega) = \frac{1}{|M_a(\omega)| |M_b(\omega)|} \quad (3)$$

Called the “phase transform” (PHAT)[7], this weighting scheme does not use an existing noise measurement to bias the cross correlation, but instead only depends on the current magnitude at each frequency. Other weighting schemes designed to include knowledge of ambient noise in the estimates were also tried, but gave similar performance to the PHAT scheme.

The position (1) that corresponds to the highest cross correlation value (F) is then the most likely position to contain the sound source. In theory, given enough microphones in an array, it should be possible to exactly localize upon the source generating the noise. In practice, however, given the small distances between microphones in an on-robot array, as well as the levels of ambient noise and echoes from the environment, we have observed high amounts of error in the localization from one location (Figure 2). That error tends to be concentrated mostly along the axis stretching from the center of the array out through the sound source location, meaning that the cross correlation results are generally better at estimating angle to the sound source rather than distance.

The input used for the localization task is 250-ms of synchronized audio data from every microphone in the array. As the sources of interest were largely speech and music sources, the sample frequency was limited to 8192 samples per channel per second to minimize computational requirements while still working with the dominant frequency ranges present in those sources.

B. Building the Evidence Grid

The evidence grid representation uses Bayesian updating to estimate the probability of a sound source being located in a set of predetermined locations (i.e. a grid cell center). Initially, it is assumed that every grid cell has a 50%

probability of containing a sound source. Then as each new sensor measurement is added to the evidence grid those probabilities for each grid cell are adjusted. For the simplicity of adding measurements together, we use the log odds notation to update the evidence grid. Equation 4 demonstrates this additive process for each new measurement:

$$\log\left(\frac{p(SS_{x,y} | z^t, s^t)}{1 - p(SS_{x,y} | z^t, s^t)}\right) = \log\left(\frac{p(SS_{x,y} | z_t, s_t)}{1 - p(SS_{x,y} | z_t, s_t)}\right) + \log\left(\frac{p(SS_{x,y} | z^{t-1}, s^{t-1})}{1 - p(SS_{x,y} | z^{t-1}, s^{t-1})}\right) \quad (4)$$

In these equations, $p(SS_{x,y}|z^t,s^t)$ is the probability of occupancy given all evidence (sensor measurements z , and robot pose s) available at time (t), and $p(SS_{x,y}|z_t,s_t)$ is the *inverse sensor model*, or probability that a single grid cell contains the sound source based on a single measurement.

The *inverse sensor model* used in this work is simply the scaled result of the cross correlation measurements. All results were scaled between two chosen probabilities [P_{low} and P_{high}] so that the lowest cross correlation value resulted in a probability of P_{low} and the highest in P_{high} . The resulting $p(SS_{x,y}|z_t,s_t)$ can be determined by using equation 5:

$$\begin{aligned} K_1 &= (F_{min}(t) - F_{max}(t)P_{low} / P_{high}) / (1 - P_{low} / P_{high}) \\ K_2 &= (F_{max}(t) + K_1(P_{high} - 1)) / (P_{high} - K_1) \\ p(SS_t | z_t) &= (F_t(t) - K_1) / K_2 \end{aligned} \quad (5)$$

Where $F_{min}(t)$ and $F_{max}(t)$ are the lowest and highest F_l values calculated for the measurement taken at time (t). To then extract the resulting $p(SS_{x,y}|z_t,s_t)$ from $p(SS_t|z_t)$ the robot pose (s_t) is used to convert from local coordinates (l) to global coordinates (x,y).

In this work, the spatial likelihood results were scaled between [0.2, 0.95], but this could be varied when tracking different types of sources. These scaling numbers were chosen empirically based on spatial likelihood reliability. As the spatial likelihoods would generally only point at one source at a time, areas not indicated with a high cross correlation result were not necessarily devoid of sources so setting the probability at 0 would unfairly penalize the quiet source. Similarly, spatial likelihoods could also make a mistake in the direction they pointed, and so 100% confidence was inappropriate in scaling the results.

IV. HARDWARE SETUP

The robot hardware used in this work was a B21R research robot manufactured by iRobot. The robot is equipped with a SICK laser measurement system (LMS) mounted in the robot base, and two onboard computers for processing. Robot pose information is then provided by the continuous localization[8] algorithm, which uses a spatial evidence grid representation (different from auditory evidence grids) constructed from LMS range data and robot base odometry. The robot also has an additional SICK LMS mounted above

the robot base, as well as a full sonar ring that was which were not used in these experiments.

The equipment used for gathering the acoustic data was an array of (4) Audio-Technica AT831b lavalier microphones mounted at the top of the robot. These microphones were each connected to battery powered preamps mounted inside the robot body and then to an 8-Channel PCMCIA data acquisition board. The equipped robot is seen in Figure 1.

V. RESULTS

To test the algorithm, we ran the robot in 20 trials, varying two parameters: (1) the set of sources active in the environment, and (2) whether or not the robot was moving while gathering data. A total of 10 different configurations of sources were tested, where a source configuration is defined as a unique set of active sources in the environment. For the following trials, 9 sources were mapped by the robot: 2 human speakers (male and female), 1 tape recording of human speech, 2 radios playing different types of music, and 4 air vents in the laboratory. Figure 3 shows the relative positioning of each of the sources, along with the grid used for localizing the robot in the 12x12m² laboratory. In general, the robot was not always exploring the entire but was instead restricted to a subsection in the vicinity of the sound sources of interest.

Robot movement during these tests was varied according to whether or not it was stationary while sampling audio data. In both cases, the robot was teleoperated in a large circle in the vicinity of the sound sources. In the first case, however, the robot would stop 6-7 times to gather samples of the auditory scene using its microphone array. In the second case, the robot would continue to gather microphone array measurements while it was moving. The reason for the different data collection techniques was to evaluate the effect of increased ego-noise on the robot, as movement increased the volume of wheel and motor noise generated present.

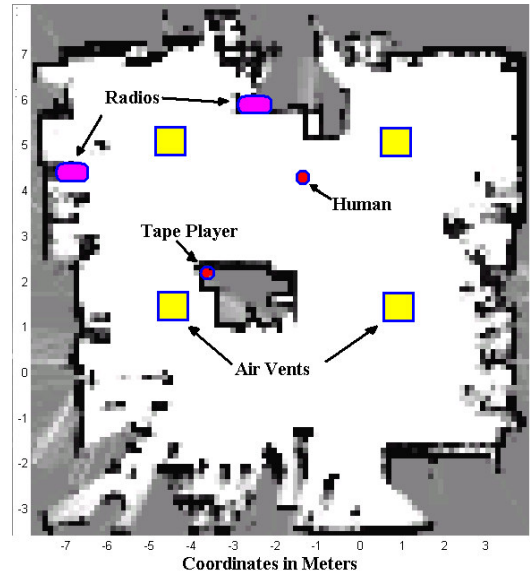


Figure 3. Spatial evidence grid used by the robot for localization, with source positions shown relative to the obstacle positions in the room. Note that not all sources were active in every test.

The results of all mapping experiments are shown in Table 1, where a successful test is defined as correctly mapping all active sources. The number of sources listed in this table does not include air vents. As the vents could not be fully disabled, they were on during all trials, but were too quiet to detect except when all other sources were disabled.

# of Sources	# of Source Config.	Successes: Pausing while Collecting	Successes: Moving while Collecting
1	5	5	5 (4)
2	4	4	3
>2	2	0	0

Table 1. For two or fewer sources, the robot correctly mapped all active sources in 17 out of 18 trials. One trial using a moving robot and a single human speech source, however, resulted in an additional illusory peak in the evidence grid. Trials with three or more sources, however, always missed at least one active source.

In general, as demonstrated by the table results, the auditory evidence grid algorithm worked very well for mapping one or two sources. In only one test with two sources, did the robot fail to correctly map all of the active sources. There was an additional test using one active source, in which a phantom, or illusory, peak appeared strong enough in the evidence grid to be mistaken for a real source, but the active source was still correctly mapped. Note that in both of these trials, the robot was moving while collecting data instead of stopping, so movement obviously did introduce some additional error, but the algorithm still succeeded in most cases to successfully map 1-2 sources (Figure 4).

For more than 2 sources, the auditory evidence grid representation was not as successful. The algorithm was tested on mapping air vents (only 2 vents successfully mapped), and mapping three speech/music sources (tape player, and 2 radios), both when moving, and pausing to listen. While watching the evidence grids form in real-time, all of the sources would appear for a time while the robot was in close proximity, but then, as the robot moved away from the source, one or more sources would be suppressed by new measurements. The reason for this suppression is due to the underlying spatial likelihood estimation process. When localizing sources using the spatial likelihood method for acoustic localization, only one source at a time is reliably detected. Each measurement that points at one source, however, will effectively suppress the evidence grid in other areas not being pointed at, including areas containing another source. Therefore, if too few measurements point at a source because it is too quiet or too far away, then the cumulative effect of the suppression may end up being greater than the cumulative positive effect.

This suppression problem, however, can be overcome with some post-processing of the data to correctly map the missing sources. As discussed earlier, each spatial likelihood measurement is most strongly associated with a single angle, so for each measurement we can calculate the most likely angle to the source by compressing the log-likelihoods along the angular axis at some increment δ . Let F_θ be the log-likelihood of the detected source being located along angle θ , and $F_{\phi,r}$ be the log-likelihood of the sound source being

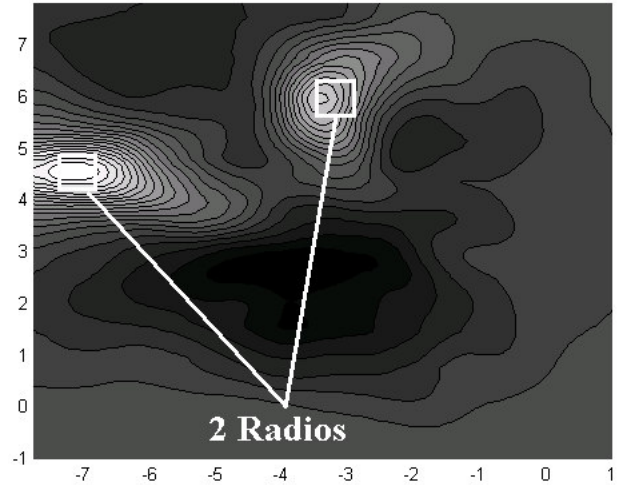


Figure 4. Two Radios are mapped with a moving robot using four microphones. This map was created from 190 spatial likelihoods and smoothed using a square sliding window (0.6-m in width).

located at cylindrical coordinates (ϕ, r) . Then the most likely angle towards the detected source is the angle (θ) with the highest log-likelihood.

$$F_\theta = \sum_{r=0}^3 \sum_{\phi=\theta-\delta/2}^{\theta+\delta/2} \frac{F_{\phi,r}}{1 - F_{\phi,r}} \quad (6)$$

Now, using this notion of most likely source angle, we can determine which spatial likelihood measurements actually point at sources found:

$$|\theta_{best} + \theta_{robot} - \alpha_{source}| \leq threshold \quad (7)$$

where θ_{best} is the most likely angle as predicted by the spatial likelihood function in local coordinates, θ_{robot} is the orientation of the robot in global coordinates, and α_{source} is the angle from the robot to a detected source in global coordinates. If the difference between the angle to the source location (as predicted by the evidence grid) and the most likely angle (as predicted by a single spatial likelihood measurement) is less than some threshold, then that measurement is estimated to be pointing at the source.

The final post-processing step is to create a second evidence grid out of those measurements not estimated to be pointing at a known loud source, thus removing its suppressive effect on weaker sources. The results of this post-processing can be seen in Figure 5, where the missing source from the 3-source trial is successfully found by excluding measurements pointing at the loudest source. The same approach resolves the missing source problem for both the three-source trials (moving and not) and the one two-source trial with a moving robot. Note that this technique for reducing the suppressive effect of the loudest source does not always fully remove the source from the resulting evidence grid, as some few spatial likely measurements may still indicate multiple sources. Altogether, extracting larger number of sources only takes a few seconds to run on an average processor. It is currently only run in post-processing because of the need for the suppressive source position. If, however, that knowledge was known a priori, then this step could be run in real time instead.

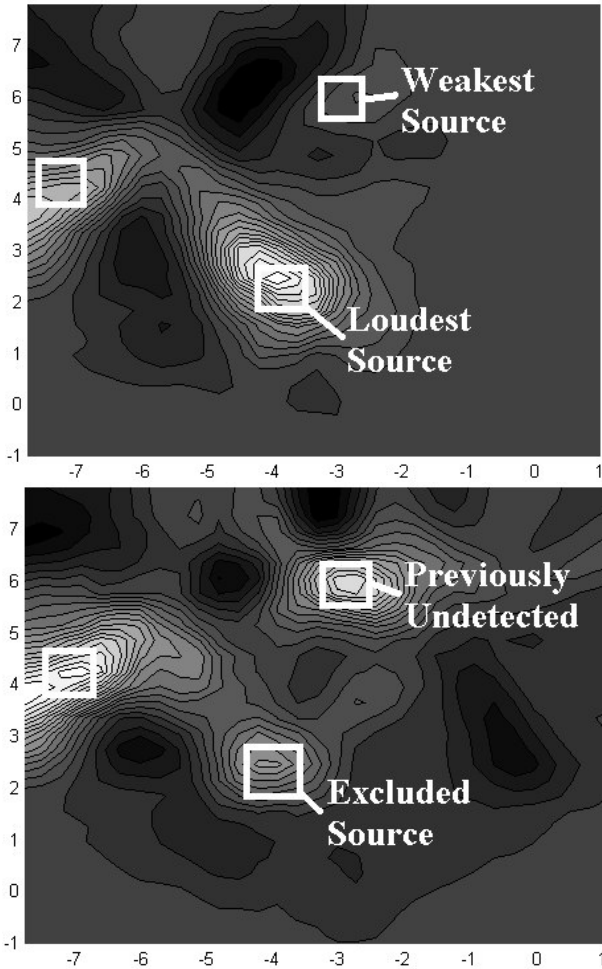


Figure 5. Two radios and a speech source are active while the robot collects data. In the original evidence grid [Top], only the loudest sources are evident. By creating a second evidence grid without using those spatial likelihoods directly pointing at the loudest source [Bottom], we can still correctly localize the weaker sound sources.

VI. DISCUSSION

As discussed earlier, the goal of the auditory evidence grid approach is to reduce the effect of echoic environments and robot ego-noise on sound source localization, and in general, what the testing revealed, was that the evidence grid representation was highly appropriate for accomplishing that goal. The resulting maps were not strongly influenced by the echoic effects of an indoor environment, and could easily be used to find 1-2 sources despite the presence of robot ego-noise, whether the robot was moving or not. Furthermore, we could extend the algorithm to an even larger number of sources by using some additional post-processing. This is not to say, however, that the evidence grid representation is perfect. What follows here is a discussion of how different design decisions may affect the accuracy of the resulting sound source map:

- *Moving when gathering auditory data* - As was seen earlier, the evidence grid representation still works when the robot is moving (instead of moving, pausing to collect, and then moving again), but more problems occurred in trials

where the robot was moving than when not. There are two reasons for this decreased accuracy in evidence grid. The first reason is that, when moving, the robot pose estimation algorithm introduces more relative error into the representation. As the robot pose estimates are used to align overlapping spatial likelihood measurements, this results in wider, lower peaks in the resulting evidence grid. The second problem when moving comes from the louder robot ego-noise generated by the robots wheels and motors. If the robot is generating more noise when moving, than there will be more noise present in the environment that can partially or totally mask the active sound sources being investigated. Algorithmically, this results in degraded spatial likelihood results, and less certainty on the origin of the loudest sound. The effect of this on the resulting evidence grid is twofold: (1) poorer spatial likelihood accuracy results in more phantom peaks, making it harder to distinguish actual sources; and (2) rougher object contours will be evident in the final map.

- *Number of Microphones* - Many robots are now being equipped with a binaural microphone array (i.e. two microphones) to mimic human hearing, and there is no reason why spatial likelihoods cannot be computed using only 2 microphones. However, with a binaural setup, the accuracy of calculated spatial likelihoods decreases in both distance and angle. So to test the effect of a binaural setup on auditory evidence grid, we reused the data from the same trials discussed earlier, and only used two microphone streams instead of all four to generate the spatial likelihoods. The resulting effects on the evidence grid from this binaural approach is actually very similar to those seen when moving while gathering audio data: (1) more phantom noise sources, or peaks in the evidence grid are generated, and (2) the object peaks have rougher contours. However, as demonstrated in Figure 6, the same sources were generally still evident for both 2 and 4 microphone configurations in most trials.

- *Map Resolution* - To detect sources in real-time the evidence grid and spatial likelihood grid cell size was a minimum of 0.3m. This is a relatively coarse resolution that may have affected the resulting accuracy. To determine exactly how the resulting map was affected, we recreated the trial maps at a higher resolution (0.1m) using the data collected earlier. The result of increasing resolution was that it shifted the center of the peak in the evidence grid towards a more accurate center. However, that center would have otherwise been included in a larger grid cell at a lower resolution, so it was not unexpected. Unfortunately, though, changing resolution did not appear to affect the creation of phantom peaks or rougher contours.

- *Filtering the Sample Set* - If a priori knowledge exists about the set of sound sources being mapped, then another method for removing error from the map is to filter the sample set. One such filter tested was an rms threshold, like that employed in Linear Predictive Coding [9] for detecting speech over the telephone. The resulting maps for speech sources had smoother contours and better defined peaks. There is a tradeoff, however, in that fewer samples were used to create the maps in general, and that some source types (non-speech) were removed by this filter entirely.

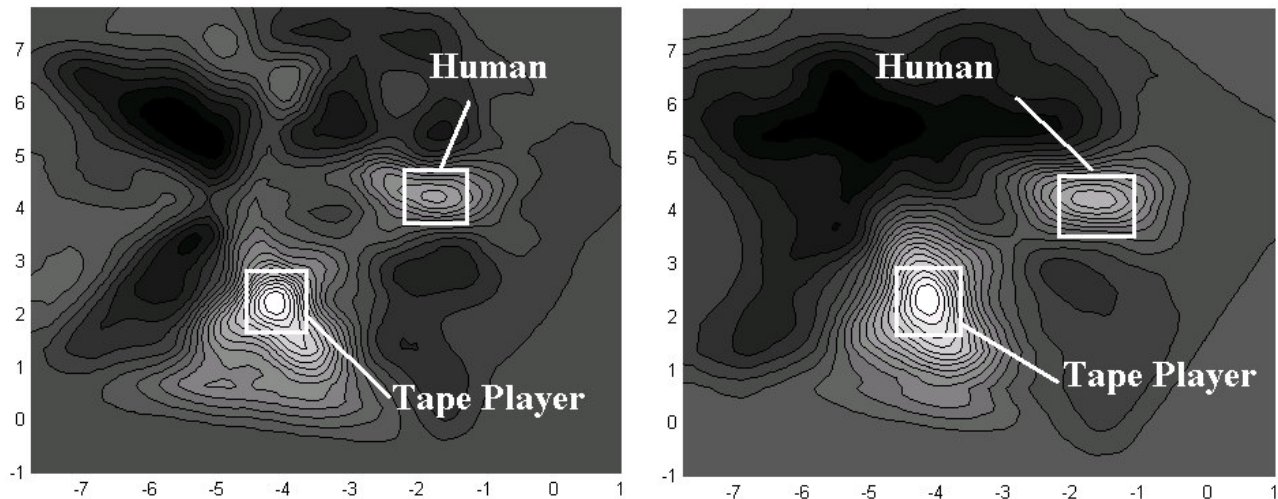


Figure 6. Auditory evidence grids localizing two speech sources (a stationary human speaker and a tape player) from 463 data points collected at 6 positions. Both grids are created from the same 463 recorded samples (not all are speech) using either a 2 [Left] or 4 [Right] microphone array. Note the rougher contours and phantom peaks found in the grid created by only 2 microphones.

VII. CONCLUSIONS

In conclusion, this work has demonstrated the successful application of evidence grids to the sound source localization problem. Using a mobile robot equipped with a microphone array and a laser range finder for robot localization, we were able to detect a wide variety of sound sources, including people, air vents, and radios, and place them on an easy to understand map of the local environment. These maps could be created on the fly for small numbers of sources, or for larger number of sources with some small amount of post-processing. Most importantly, however, the auditory evidence grid algorithm appeared resistant to both echoic locations and robot ego-noise. While reducing ego-noise still leads to more precisely localized sources, a moving robot does not have to stop or otherwise change its movement pattern to successfully detect and localize sources. Therefore a robot could perform other tasks while gathering information about the auditory soundscape in which it is working.

In addition to robustness in the presence of noise, it has also been demonstrated in this work that a large number of microphones is not necessary to create effective auditory evidence grids. While more microphones continues to mean a better, more accurate result, as long as there are at least two synchronized microphones, sound sources can still be mapped. This then opens up a wide range of applications that may have access to a standard PC soundcard equipped with stereo microphone input, but not to higher end data acquisition equipment. So this is also a practical approach to sound source localization on a mobile robot.

Ultimately, the goal of the auditory evidence grid work is to provide a map to an autonomous robot of all the sound source locations in the environment. Then, with such a map, there are at least two more very interesting problems. First, the robot can possibly use this map of the sound sources in the

environment improve its own localization within an environment. Second, like a map of obstacles, a map of ambient sound sources could be used as a spatial guide to position a robot intelligently with respect to the surrounding soundscape and therefore generally improve overall performance at tasks requiring auditory sensing.

VIII. ACKNOWLEDGEMENTS

This research has been funded by the ONR Intelligent Systems Program (Work Order #N0001405WX30022).

IX. REFERENCES

- [1] A. Elfes, "Multi-source spatial data fusion using Bayesian reasoning," in *Data Fusion in Robotics and Machine Intelligence*, M. A. Abidi and R. C. Gonzales, Eds. New York: Academic Press, 1992.
- [2] H. Moravec, "Sensor Fusion in Evidence Grids for Mobile Robots," in *A.I. Magazine*, 1998, pp. 61-74.
- [3] S. Stillman and I. Essa, "Towards reliable multimodal sensing in aware environments," presented at 2001 Workshop on Perceptive user interfaces, Orlando, FL, 2001.
- [4] E. Martinson and R. C. Arkin, "Noise Maps for Acoustically Sensitive Navigation," *Proceedings of SPIE*, vol. 5609, 2004.
- [5] K. Nakadai, K. Hidai, H.G. Okuno, and H. Kitano., "Epipolar Geometry Based Sound Localization and Extraction for Humanoid Audition," presented at IEEE/RISJ International Conference on Intelligent Robots and Systems (IROS), Maui, Hawaii, 2001.
- [6] M. Brandstein and H. F. Silverman, "A Practical Methodology for Speech Source Localization with Microphone Arrays," *Computer, Speech, and Language*, vol. 11, pp. 91--126, 1997.
- [7] B. Mungamuru and P. Aarabi, "Enhanced Sound Localization," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 34, 2004.
- [8] A. Schultz and W. Adams, "Continuous localization using evidence grids," presented at IEEE International Conf. on Robotics and Automation, Leuven, Belgium, 1998.
- [9] T. E. Tremain, "The Government Standard Linear Predictive Coding Algorithm: LPC-10," in *Speech Technology Magazine*, 1982, pp. p. 40-49.