Combining Sound Localization and Laser-based Object Recognition

Laurent Calmes[†‡], **Hermann Wagner**[†]
[†] Institute for Biology II
Department of Zoology and Animal Physiology
RWTH Aachen University
52056 Aachen, Germany
calmes@pool.informatik.rwth-aachen.de
wagner@bio2.rwth-aachen.de

Stefan Schiffer[‡], **Gerhard Lakemeyer**[‡]
[‡] Knowledge-based Systems Group
Department of Computer Science 5
RWTH Aachen University
52056 Aachen, Germany
{schiffer,gerhard}@cs.rwth-aachen.de

Abstract

Mobile robots, in general, and service robots in human environments, in particular, need to have versatile abilities to perceive and interact with their environment. Biologically inspired sound source localization is an interesting ability for such a robot. When combined with other sensory input both the sound localization and the general interaction abilities can be improved. In particular, spatial filtering can be used to improve the signal-to-noise ratio of speech signals emanating from a given direction in order to enhance speech recognition abilities. In this paper we investigate and discuss the combination of sound source localization and laser-based object recognition on a mobile robot.

Introduction

Speech recognition is a crucial ability for communication with mobile service robots in a human environment. Although modern speech recognition systems can achieve very high recognition rates, they still have one major drawback: in order for speech recognition to perform reliably, the input signals need to have a very high signal-to-noise ratio (SNR). This is usually achieved by placing the microphone very close to the speaker's mouth, for example, with the help of a headset. However, this is a requirement which in general cannot be met on mobile robots, where the microphone can be at a considerable distance from the sound source, thus corrupting the speech signal with environmental noise. In order to improve SNR, it is very useful to know the direction to a sound source. With the help of this information, the sound source can be approached and/or spatial filtering can be used to enhance a signal from a specific direction.

In order to obtain reliable directional information, at least two microphones have to be used. Although the task would be easier with more microphones, we deliberately chose to restrict ourselves to two because the processing of only two signals is computationally less expensive and standard, off-the-shelf hardware can be used. Furthermore, two microphones are easier to fit on a mobile robotic platform than a larger array.

We investigated the combination of our existing sound localization system (Calmes, Lakemeyer, & Wagner 2007) with the robot's knowledge about its environment, especially the knowledge about dynamic objects in this paper. By combining several sensor modalities, sound sources can be matched to objects, thus enhancing the accuracy and reliability of sound localization.

The paper is organized as follows. First, we describe our approach to sound localization. Then we present how our laser-based object recognition works. Finally, we report on experiments we conducted to show how combining these two kinds of information improves the detection of sound sources followed by a brief discussion of the results and future work.

Sound Localization

We use a biologically inspired approach to sound localization. The major cue for determining the horizontal angle (azimuth) to a sound source in humans as well as in animals is the so-called interaural time difference (ITD). The ITD is caused by the different running times of the sound wave from the source to each ear.

L.A. Jeffress proposed a model in 1948 which tried to explain how ITDs could be evaluated on a neuronal level (Jeffress 1948). This model has two major features: axonal delay lines and neuronal coincidence detectors. Each coincidence detector neuron receives inputs from delay lines from the left and the right ear and fires maximally if excited from both sides simultaneously. As action potentials are transmitted by axons at finite speeds, different delay values are implemented by varying length of the axonal delay lines. Each coincidence detector is tuned to a best delay by the combination of the delay values from both input sides.

By this arrangement, the axonal delay lines compensate the ITD present in the ear input signals and only neurons with a best delay corresponding to the external delay will fire. Thus the timing information is transformed into a place code in a neuronal structure.

Strong physiological evidence for the Jeffress model was found in birds (Parks & Rubel 1975; Sullivan & Konishi 1986; Carr & Konishi 1988; 1990). In the case of mammals, it is currently debated whether these animals have delay lines at all (McAlpine & Grothe 2003).

The simplest computational implementation of the Jeffress model consists of a cross-correlation of the input signals. Our algorithm is a modification of the one proposed in (Liu *et al.* 2000). All processing takes place in the frequency domain after Fourier transformation. Delay line values are computed so that the azimuthal space is parti-

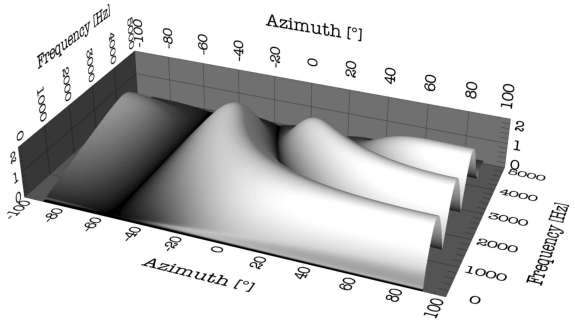

Figure 1: 3D coincidence map generated using two unit impulses. Sampling frequency was 16 kHz. The left channel was leading the right by 5 samples, resulting in an ITD of 312.5 µs. This corresponds to an azimuth of $-55°$. The z-axis denotes dissimilarity, i.e. low values correspond to high coincidence.

tioned into sectors of equal angular width, with each coincidence detector element corresponding to a specific azimuth. For each frequency bin, delaying is implemented by a phase adjustment in the left and right channels at each coincidence detector corresponding to the precomputed delay values. Coincidence detection is performed by computing the magnitude of the difference of the delayed left and right signals for each frequency and each coincidence detector element. Plotting these magnitudes against coincidence location and frequency results in a three-dimensional coincidence map. Figure 1 shows an example of such a map. It was computed by synthetically generating two unit impulses, with the left one leading the right one by 5 samples. At a sampling frequency of 16 kHz, this corresponds to an ITD of 312.5 µs. The frequency independent minimum corresponds to the simulated sound source azimuth of $-55°$. Low values in the map correspond to high coincidence for a given frequency and coincidence detector. The final localization function is computed by summing up the 3D coincidence map over frequency. Minima in the resulting function specify the location of the detectors at which highest coincidence was achieved. As each detector corresponds to a specific azimuth, the angle to the sound source can easily be determined from positions of the minima.

From the localization function, a quality criterion is derived (roughly corresponding to the cross-correlation of the input signals) by normalizing to the range of the absolute maximum and minimum. The coincidence location corresponding to the normalized minimum with the value 0 will be assigned a so-called peak height of 100%, other minima will be assigned a correspondingly lower value. Furthermore, coincidence locations with a peak height less than 50% will be discarded.

Figure 2 shows the localization accuracy of our algorithm with three different stimuli measured in an office room. The sound source was at a distance of approximately 1 m from the microphones. Sound source azimuth was varied in $10°$ steps from $-70°$ to $+70°$. Each individual data point shows the average of 400 measurements. Error bars indicate 99% confidence interval (Calmes, Lakemeyer, & Wagner 2007). As can be seen, the algorithm performs very good for broad-

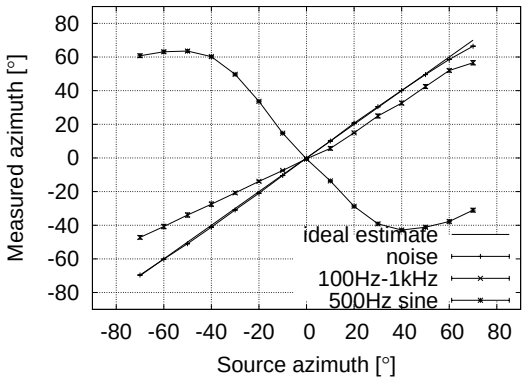

Figure 2: Accuracy of the sound localization algorithm. Averages of 400 measurements for each source position ($-70° \ldots +70°$ in $10°$ steps) are shown. Broadband noise, bandpass noise (100 Hz–1 kHz) and a 500 Hz sine were used as stimuli. Error bars indicate 99% confidence interval.

band noise and quite well for bandpass noise. The complete mislocalization of the 500 Hz sine is caused by interference with reverberations.

Under favorable acoustic conditions (high signal to noise ratio and broadband signals), the precision of the algorithm matches the accuracy of biological systems. As an example, the barn owl, a nocturnal predator, is able to hunt in total darkness by localizing the sound its prey generates. It can achieve an angular resolution of some $3°$ in azimuth as well as elevation (Knudsen, Blasdel, & Konishi 1979; Bala, Spitzer, & Takahashi 2003). Humans achieve a precision of about $1°$ in azimuth (an overview on human sound localization can be found in (Blauert 1997)).

But in contrast to the technical implementation, biological systems can maintain high accuracy in acoustically more challenging environments, with e.g. high noise and reverberation levels as well as in the presence of multiple sound sources.

The major advantage of using interaural time differences over other binaural sound localization cues which rely on the particular anatomy of the head, is their relative independence on the microphone (ear) mounting. Basically, the only parameter affecting ITDs is the distance between the microphones.

This comes with the drawback that with ITDs only the azimuth to a sound source can be determined in a range of $-90°$ to $+90°$, resulting in ambiguities whether a source is above, below, in front or behind the "head". In mobile robotics applications related to speech recognition, the relevant information is the azimuth to a source, so localization can be restricted to the horizontal plane. This assumption eliminates the above/below ambiguities, leaving the front/back confusions which can only be resolved by incorporating additional environmental knowledge.

Laser-based Object Recognition

In order to acquire information on dynamic objects in the robot's vicinity it needs to know the structure of the envi-

ronment (i.e. a map) as well as where it is located within this environment. With both information the robot can distinguish between features that belong to the environment and dynamic objects. Thus, our approach requires a (global) localization capability. The primary sensor our robot uses for localization and navigation is a 360° laser range finder. In the following we briefly describe how we do localization and object recognition with this sensor.

Localization

Our self-localization uses a Monte Carlo approach to localization (Dellaert *et al.* 1999). It works by approximating the position estimation by a set of weighted samples:

$$\mathbf{P}(l_t) \sim \{(l_{1,t}, w_{1,t}), \ldots, (l_{N,t}, w_{N,t})\} = \mathbf{S}_t$$

Each sample represents one hypothesis for the pose of the robot. Roughly, the Monte Carlo Localization algorithm now chooses the most likely hypothesis given the previous estimate, the actual sensor input, the current motor commands, and a map of the environment. In the beginning of a global localization process the robot has no clue about its position and therefore it has many hypotheses which are uniformly distributed. After driving around and taking new sensor updates the robot's belief about its position condenses to some few main hypotheses. Finally, when the algorithm converges, there is one main hypothesis representing the robot's strongest belief on its position.

Novelty Filter For localization we use an occupancy grid map (Moravec & Elfes 1985) of the environment. This allows us to additionally apply a Novelty filter as described in (Fox *et al.* 1998) in the localization process. It filters readings which, related to the map and the current believed position, are too short and can thus be classified to hit dynamic obstacles.

Suppose we have a map and believe we are at position l_b in this map. Then we can compute the expected distance d_e our laser range finder should measure in any direction. In conjunction with the statistical model of the laser range finder we can compute the probability that a reading d_i is too short.

Localization Accuracy With the above approach we are able to localize with high accuracy in almost any indoor environment. Depending on the cell size of the occupancy grid the average error usually is around $15\ \mathrm{cm}$ in position and 4° in orientation even in large environments. The method is presented in detail in (Strack, Ferrein, & Lakemeyer 2005).

Object Recognition

Based on the laser readings that where classified to be dynamic we perform object recognition. In a first step, groups of dynamic readings are clustered. This is done based on the fact that readings belonging to one particular object cannot be farther away from each other then the diameter of the object's convex hull. To be able to distinguish between different dynamic objects, we use the laser signature of the objects for classification by size and form on the clustered groups afterwards. The dynamic objects are classified each time new laser readings arrive. Thus, they can of course change both in number and position. To stabilize the robot's perception we make use of the Hungarian method (Kuhn 1955) to track objects from one cycle to the next.

The object recognition was originally developed for robotic soccer. In the soccer setting we are able to distinguish between our own robots and opponents, and even humans can be told apart. Though, the most important information there is whether the object is a teammate or an opponent obstacle. Our heuristic for classification is still rough at the moment. Nevertheless, the object recognition output is accurate enough to perform an association between sound sources and dynamic objects.

Turning Delay The localization module consists of several components that run with different frequencies. The classification routine that our object recognition bases upon is called with a frequency four times lower than new laser readings arrive. Thus, there is a certain delay within the detection of dynamic objects which has to be taken into account in our evaluation. This delay becomes especially obvious when the robot is turning.

Experiments

Based on the combination of both the sound sources detected and the objects recognized we investigated how to steer the robot's attention towards a direction of particular interest.

Matching Sound Sources and Objects

Our framework features a multi-threaded architecture. Several modules are running in parallel each with its own cycle time. The sound localizer component is able to produce azimuth estimates at a rate of about $32\ \mathrm{Hz}$. A signal detector, calibrated to the background noise level, ensures that only signal sections containing more energy than the background noise are used for localization. If new sound sources are detected they are written to a blackboard where any other module can retrieve them from. The information is organized in a list which contains the azimuth of the sound sources detected along with the corresponding peak heights. It is sorted by descending peak height. Based on the information provided by the localization module, the object recognition module clusters the laser readings that have been classified as dynamic and computes the positions of dynamic obstacles thereupon. Those objects are also written to the blackboard.

Our attention module which determines which action to take runs with a frequency of $10\ \mathrm{Hz}$, i.e. a new cycle starts every $100\ \mathrm{ms}$. In the first step, we check whether there is new data from the sound localizer. If not, we are done already and skip this cycle. If there are sound sources available, we retrieve the corresponding list of angles and proceed.

For now, we only work on one sound source, that is the one with 100% peak height. However, with some minor modifications we could also process all sources detected. We retrieve the relative angle to this source. Then we iterate over all dynamic objects and search for the one object that is in the direction of the sound source. Due to front/back confusions, we have to check for both directions. If we find

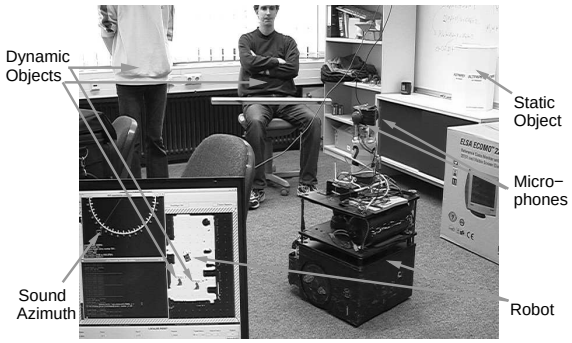

Figure 3: Initial test setup

an appropriate object to match the sound with, we schedule a command to the motor to turn towards this object (and not to the sound source itself). An object is considered appropriate if the relative angle from the robot to this object does not differ more than a given tolerance value from the relative angle to the sound source.

Figure 3 shows our initial test setup. The robot just detected a sound in the direction of the sitting person and has matched it to a corresponding dynamic obstacle. It is about to turn towards this object. In the upper right corner of the picture one can see a box which was used to generate noises that do not have any corresponding dynamic object. We generated the noise by simply hitting the box with a stick.

Initial Tests

A first series of tests showed that in the vast majority of cases the robot was able to correctly discriminate sounds emanating from dynamic objects (i.e. persons) from noises emitted by the static object. The correct turning behavior could be observed as long as a dynamic object was not too close to the static object. In that case, the robot would react to the noise emitted by the static object, but would nevertheless turn towards the dynamic object.

A noteworthy observation is that the matching of sound sources to dynamic objects sometimes helped in resolving the front/back confusions immanent in our sound localization method. If there is no object in front of the robot corresponding to the sound's azimuth but there is one behind it, the robot would turn to the one behind it. Unfortunately, in symmetric situations ambiguities remain. There are cases in which there were objects in front of the robot as well as behind it which both could match the estimated sound source azimuth.

As the tolerance between the angle to the sound source and the angle to the dynamic object was arbitrarily chosen to be rather large ($30°$), these front/back confusions could certainly have been reduced by choosing a smaller value. This would also keep the robot from reacting to noise from static objects if there was a dynamic object in the vicinity.

Evaluation Setup

After the initial test described above we prepared and conducted a more extensive series of tests for evaluation purposes. The quantitative evaluation took place in the seminar room of the Department of Computer Science 5. The

Speaker #	x	y	object
1	$-1.00\,\mathrm{m}$	$1.00\,\mathrm{m}$	yes
2	$-0.15\,\mathrm{m}$	$1.75\,\mathrm{m}$	no
3	$-1.50\,\mathrm{m}$	$1.25\,\mathrm{m}$	no
4	$-1.75\,\mathrm{m}$	$-0.15\,\mathrm{m}$	yes
5	$2.25\,\mathrm{m}$	$-0.75\,\mathrm{m}$	no
6	$2.00\,\mathrm{m}$	$0.75\,\mathrm{m}$	yes

Table 1: List of positions of the loudspeakers and whether or not there was a dynamic object associated.

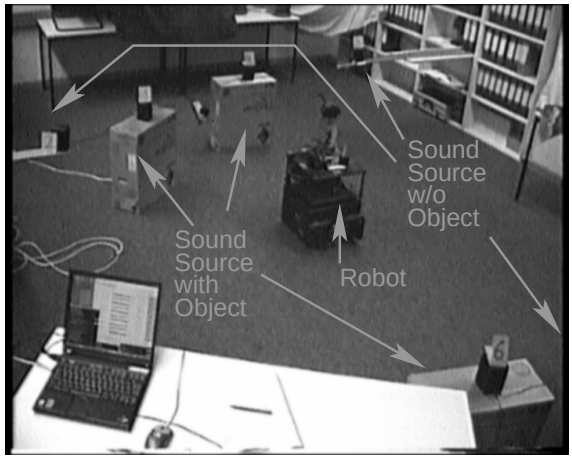

Figure 4: Extensive evaluation setup

room has a size of about $5\,\mathrm{m} \times 10\,\mathrm{m}$. The robot was placed at the center of this room at coordinates $(0, 0)$. We placed six sound sources (loudspeakers) around the robot, three of which had a (dynamic) object associated to them. The coordinates of the sound sources are shown in Table 1. Loudspeakers 1, 4 and 6 were placed on cardboard boxes so that the robot's laser scanner could detect an object corresponding to these sources. Loudspeakers 2, 3 and 5 were mounted in such a way that no object could be detected. The evaluation setup is shown in Figure 4. Within this setup we conducted 4 evaluation experiments. An experiment consisted of 100 trials, where each trial consisted of randomly selecting a loudspeaker for noise playback. The task of the robot was to turn towards an object if the source was associated with this object. We conducted two experiments (200 trials) with a fixed angular tolerance of $23°$ (cf. Section Initial tests) and two experiments (200 trials) with a varying tolerance value described in the following.

Adaptive Tolerance Control Because the accuracy of the sound localizer decreases with more lateral source azimuths, the two latter experiments were conducted with a variable angular tolerance. With this adaptive tolerance control (ATC), angular tolerance was varied linearly between $5°$ (for a source azimuth of $0°$) and $30°$ (for a source azimuth of $\pm 90°$) computed by $tol_{atc} = 25° \cdot \left| \frac{azimuth_{src}}{90°} \right| + 5°$

Data Analysis

Table 2 shows the results of the experiments. There were three cases in which a trial was considered as being correct:

1. No object was associated with the source emitting a sound and the robot did not select any target.

	# of trials	%correct	%symmetric (of correct trials)
ATC	200	63.00	2.38
No ATC	200	57.50	8.70

Table 2: Performance evaluation for ATC and fixed angular resolution (%symmetric indicates the percentage of correct trials caused by front/back confusions due to the sound localizer)

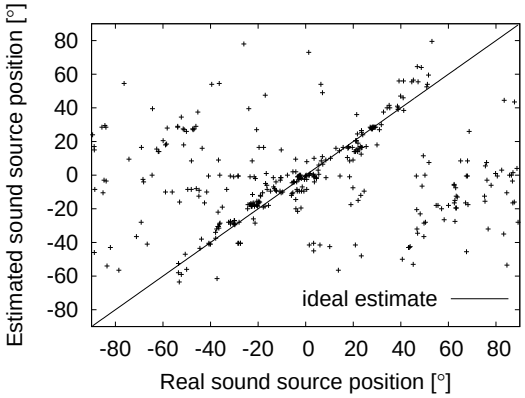

Figure 5: Real vs. estimated sound source azimuth of all trials (non-ATC and ATC combined)

2. There was an object associated with the source emitting the sound cue and the robot selected that object (with the given angular tolerance) as its target.
3. Either there was an object associated with the source or there was no object associated but one on the opposing side of the robot. Then the robot selected an object symmetric to the source (front/back confusions) as target.

We logged all relevant state data from the sensors, the generated noises and the motion commands issued to the robot. As can be seen in Table 2, the overall accuracy of the system is not very high, although the system managed to produce a correct response to the given stimulus in more than $50\ \%$ of the trials. A slight improvement could be achieved with the adaptive tolerance control algorithm in comparison to the fixed tolerance value. In the following sections, we will analyze the system's performance in more detail.

Sound Localization Performance In order to assess the sound localization performance within our evaluation, real source positions (with respect to the microphone assembly) were plotted against the azimuths returned by the localization system for all trials (non-ATC and ATC combined). These data are shown in Figure 5. From this, it becomes evident that the sound localization system did not perform very well, especially when one compares Fig. 5 to Fig. 2. Because of the differing conditions (larger room, larger distance to sound sources), we did not expect as high a precision as for the broadband noise in Fig. 2 (although we used broadband noise signals). Still, we were surprised by the low performance. We will address this issue again in the discussion at the end of this paper.

For almost all absolute sound source azimuths above $55°$ the detection error was greater then $25°$. We already mentioned that the sound localizer's accuracy decreases with increasing laterality of the source azimuth. However, this cannot be the only reason for the rather weak performance of the sound localizer in our evaluation setup, as there are also significantly high errors in the detection for source azimuths less than $45°$.

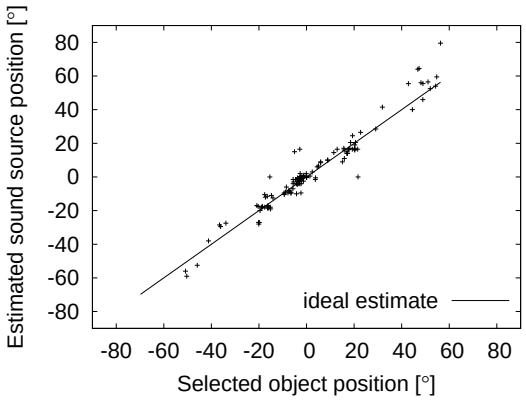

Figure 6: Real positions of selected targets vs. estimated source positions (correct trials, non-ATC and ATC combined)

We will now show that the additional information about dynamic obstacles can, at least partly, make up for the sound localizer's performance.

Object Association Performance Figure 6 shows the positions of the selected objects plotted against the sound localization estimates for all correct trials (non-ATC and ATC combined). In this case, estimated sound source positions correspond well with the target objects. Deviations from the correct azimuths are consistently within the limits of the respective angular tolerance applied for each trial.

As one can see we were able to identify and make up for the low reliability of sound localization estimates with a fairly simple association algorithm. By only allowing object associations within a certain angular tolerance, output from the sound localizer with a large error could be eliminated successfully. For one, this is a cheap way to determine whether the sound source localization works correctly. For another, in some cases symmetric confusion could be resolved by combining the sound sources with dynamic objects. However, there have also been erroneous associations with alleged symmetric objects.

Discussion

Our experiments show that, in order to use sound localization effectively in realistic environments for mobile robotics applications, the acoustic information has to be combined with data from other sensor modalities. In this sense, the unreliable behavior of the sound localization algorithm in this case might well have been a blessing in disguise. With a sound localization system in good working order, the experiments would not have yielded such interesting results. As it is, only the combination of object recognition and sound localization makes it possible for the robot to detect and eliminate errors in estimated sound source positions.

The question remains why the sound localization system did not perform well during our experiments. The initial

evaluation of the algorithm (Calmes, Lakemeyer, & Wagner 2007) showed that, although the algorithm can be very accurate, it is sensitive to reverberations. The room in which the experiments took place is larger than any in which the system has been tested before and relatively empty. This leads to perceivable reverberations which could account for (some of) the error.

Furthermore, previous experiments had all been conducted with no obstruction between the microphones. On the robot the two microphones were mounted on opposite sides of a plastic tube with a diameter of approximately $13\ \mathrm{cm}$. This might have altered ITDs in a frequency-dependent way, as from a critical frequency upwards, the sound wave would have to bend around the tube to reach the second microphone. Measuring the head-related transfer functions (HRTFs; frequency- and sound source position dependent variations of the binaural cues) of the robot's microphone mount might show if these could affect accuracy negatively. In that case, taking into account the HRTFs during localization could alleviate the problem.

Finally, during the experiments we only took into account the best azimuth provided by the sound localizer. It could be that, when multiple azimuths were detected, the correct source position was among them, but not considered the best by the sound localization algorithm. Considering all source position estimates instead might also help in increasing the accuracy of the system.

Once this question is solved, we plan to replace the simple object association method by a more sophisticated algorithm based on Bayesian inference. This would make it possible to track multiple hypotheses of sound sources based on the auditory information, the map of the environment and the knowledge about dynamic objects in a more robust manner.

Obvious applications for our system lie in general attention control for mobile robots by detecting and tracking humans (dynamic objects emitting sound) and as a frontend for a speech recognition system. Realistic scenarios will impose noisy conditions not unlike those we experienced in our evaluation setup. Thus, directing attention towards a specific person will enable the robot to move closer to that person and employ directional filtering methods to enhance the speech signal from that particular direction.

Another extension for future work could be to integrate qualitative spatial descriptions to allow for an even more natural integration of sound information in human-robot interaction.

Additional Information

You can download a subtitled video of one of our evaluation runs at `http://robocup.rwth-aachen.de/soulabor`.

Acknowledgements

This work was supported by the German National Science Foundation (DFG) in the HeRBiE project and the Graduate School 643 *Software for Mobile Communication Systems*. Further, we thank the reviewers for their comments.

References

Bala, A.; Spitzer, M.; and Takahashi, T. 2003. Prediction of auditory spatial acuity from neural images on the owl's auditory space map. *Nature* 424:771–774.

Blauert, J. 1997. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press.

Calmes, L.; Lakemeyer, G.; and Wagner, H. 2007. Azimuthal sound localization using coincidence of timing across frequency on a robotic platform. *Journal of the Acoustical Society of America*. (accepted for publication).

Carr, C. E., and Konishi, M. 1988. Axonal delay lines for time measurement in the owls brain stem. *Proc. of the National Academy of Sciences USA* 85:8311–8315.

Carr, C. E., and Konishi, M. 1990. A circuit for detection of interaural time differences in the brainstem of the barn owl. *Journal of Neuroscience* 10:3227–3246.

Dellaert, F.; Fox, D.; Burgard, W.; and Thrun, S. 1999. Monte Carlo localization for mobile robots. In *Proc. of the International Conference on Robotics and Automation (ICRA)*.

Fox, D.; Burgard, W.; Thrun, S.; and Cremers, A. B. 1998. Position estimation for mobile robots in dynamic environments. In *AAAI '98/IAAI '98: Proc. of the 15th National/10th Conf. on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, 983–988. Menlo Park, CA, USA: American Association for Artificial Intelligence.

Jeffress, L. 1948. A place theory of sound localization. *Journal of Comparative Physiology and Psychology* 41(1):35–39.

Knudsen, E. I.; Blasdel, G. G.; and Konishi, M. 1979. Sound localization by the barn owl (tyto alba) measured with the search coil technique. *Journal of Comparative Physiology* 133(1-11).

Kuhn, H. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* 2:83–97.

Liu, C.; Wheeler, B. C.; O'Brien, Jr., W. D.; Bilger, R. C.; Lansing, C. R.; and Feng, A. S. 2000. Localization of multiple sound sources with two microphones. *Journal of the Acoustical Society of America* 108(4):1888–1905.

McAlpine, D., and Grothe, B. 2003. Sound localization and delay lines – do mammals fit the model? *Trends in Neurosciences* 26(7):347–350.

Moravec, H., and Elfes, A. 1985. High resolution maps from wide angular sensors. In *Proc. of the International Conference on Robotics and Automation (ICRA)*, 116–121.

Parks, T. N., and Rubel, E. W. 1975. Organization of projections from n. magnocellularis to n. laminaris. *Journal of Comparative Neurology* 164:435–448.

Strack, A.; Ferrein, A.; and Lakemeyer, G. 2005. Laser-based Localization with Sparse Landmarks. In *Proc. RoboCup 2005 Symposium*.

Sullivan, W. E., and Konishi, M. 1986. Neural map of interaural phase difference in the owl's brain stem. *Proc. of the National Academy of Sciences USA* 83:8400–8404.