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 2003) 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 informations can be helpful and we discuss results obtained so far.

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 a sound wave needs to reach both ears.

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 (Carr & Konishi 1988; 1990; Parks & Rubel 1975; Sullivan & Konishi 1986). 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 partitioned 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. 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.

The major advantage of using interaural time differences over other 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 $+90^\circ$, 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 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 be resolved in most cases by incorporating the environmental knowledge of the robot.

Laser-based Object Recognition

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.

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. 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. With the above approach we are able to localize with high accuracy in almost any indoor environment. The method is presented in detail in (Strack, Ferrein, & Lakemeyer 2005).

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.

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 only important information there is whether the object is a teammate or an opponent obstacle. Therefore, our heuristic for classification is still rough at the moment.

Experimental Evaluation

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 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

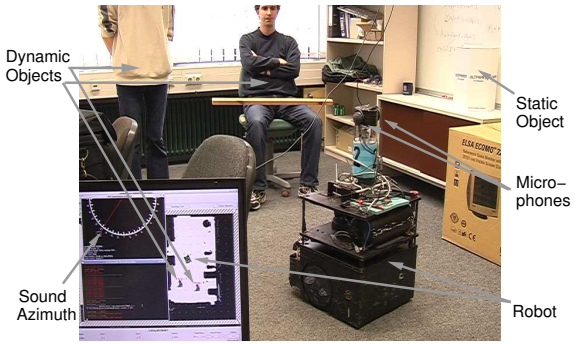

Figure 1: Evaluation setup

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 Hz, i.e. a new cycle starts every 100 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 these directions. If we find 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 30° from the relative angle to the sound source.

Figure 1 shows our evaluation 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.

Preliminary Results

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.

The matching of sound sources to dynamic objects helped in resolving front/back confusions. 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 remained. There were cases in which there were objects in front of the robot as well as behind it which could both 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.

References

Calmes, L.; Lakemeyer, G.; and Wagner, H. 2003. A sound-localization algorithm for a mobile robot. In *Abstractband der 96. Jahresversammlung der Deutschen Zoologischen Gesellschaft*.

Carr, C. E., and Konishi, M. 1988. Axonal delay lines for time measurement in the owls brain stem. *Proc Natl Acad Sci 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. *J Neurosci* 10:3227–3246.

Dellaert, F.; Fox, D.; Burgard, W.; and Thrun, S. 1999. Monte Carlo localization for mobile robots. In *Proc. of the Int. Conf. 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: Proceedings of the Fifteenth National/Tenth Conference 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. *J. Comp. Physiol. Psychol.* 41(1):35–39.

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

Liu, C.; Wheeler, B. C.; W. D. O'Brien, J.; Bilger, R. C.; Lansing, C. R.; and Feng, A. S. 2000. Localization of multiple sound sources with two microphones. *J. Acoust. Soc. Am.* 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 IEEE Int. Conf. on Robotics and Automation (ICRA)*, 116–121.

Parks, T. N., and Rubel, E. W. 1975. Organization of projections from n. magnocellularis to n. laminaris. *J. Comp. Neurol.* 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 Natl Acad Sci USA* 83:8400–8404.