Acoustic/Lidar Sensor Fusion for Car Tracking in City Traffic Scenarios ⋆

Hamma Tadjine * **Daniel Goehring** **

* *IAV GmbH, Carnotstraße 1, Berlin, 10587, Germany (e-mail: Dr.Hadj.Hamma.Tadjine@iav.de).*
** *Freie Universität Berlin, Arnimallee 7, 14195 Berlin (e-mail: daniel.goehring@fu-berlin.de)*

Abstract: In this paper we describe a sound source localization approach which, in combination with data from lidar sensors, can be used for an improved object tracking in the setting of an autonomous car. After explaining the chosen sensor setup we will show how acoustic data from two Kinect cameras, i.e., multiple microphones, which were mounted on top of a car, can be combined to derive an object's direction and distance. Part of this work will focus on a method to handle non-synchronized sensory data between the multiple acoustic sensors. We will describe how the sound localization approach was evaluated using data from lidar sensors.

Keywords: Sound source localization, Autonomous driving, Sensor data fusion.

1. INTRODUCTION

The ability to quickly detect and classify objects, especially other vehicles within the surrounding is crucial for an autonomous car. However, cluttered environments, occlusions and real-time constraints under which autonomous vehicles have to operate let this task remain a key-challenge problem. In recent years, tremendous progress has been made in the field of self-localization, world modeling and object tracking, mainly thanks to lidar, radar, and camera based sensors but also because of algorithmic advances, e.g., how to model uncertainties [Thrun (2005)] and how to apply these methods to sensor fusion [Schnuermacher (2013)], or how to train object classifiers using machine learning techniques [Mitchell (1997)]. In the past, acoustic sensors have played a minor part in robotics, especially in autonomous driving or for outdoor robotics in general only. One reason for this might be the omnipresent noise in most city road traffic and outdoor scenarios and the domination of other sensors like lidar, camera, or radar. In this paper we want to present how an autonomous vehicle can localize other vehicles in a real-world road-traffic environment. For this task we wanted to use low-cost off-the-shelve microphone arrays like the ones provided in a Microsoft Kinect camera. Since it is usually hard to determine the euclidic distance to an object with acoustic data, we will to focus on angular direction approximation. This data can still be very helpful, especially when combined with data from other sensors, e.g., lidar data from laser scanners. One possible scenario, even though not pursued in this work, would be to localize the direction at which an emergency vehicle was detected and then to assign this direction to a tracked object using lidar data. Another challenge in our scenario are the moving sound sources and comparably high velocities of other vehicles, in addition to temporarily occluded, emerging and disappearing vehicles. The presented solution was implementated on a real autonomous car using the OROCOS realtime robotics framework. For evaluation of the algorithm the acoustic data were synchronized and evaluated with lidar objects from Ibeo Lux sensors.

⋆ Part of this work has been funded by DAAD and DFG; in addition the authors would like to thank Prof. Dr. Raúl Rojas (Freie Universität Berlin, Germany) and Dr. Gerald Friedland (International Computer Science Institute, Berkeley, CA, USA) for their support.

2. RELATED WORK

A lot of progress for sound source localization has been achieved in the speech and language processing community, as in [Benesty (2007)] on beam-forming methods, or for dialog management [Frechette (2012)].

In the robotics community and especially for indoor robots there are a variety of publications on sound source local-

Fig. 1. Test Car MadeInGermany from Freie Universität Berlin, the Kinect devices were placed on top of the roof, in front of the Velodyne HDL 64 lidar sensor.

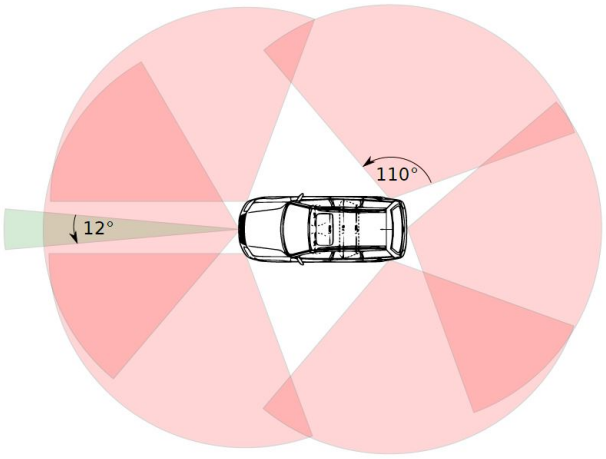

Fig. 2. The six fields of view for the six lidar sensors Lux from Ibeo and one of the radar sensors are shown, facing to the front.

ization available. The interaural time difference method (IDT) has been widely applied, as in [Liua (2010)]. In [Liu (2010)] and in [Li (2012)] the generalized cross-correlation function (GCC) is extended to localize different sound sources using an array of four different microphones. A further approach unsing a four-microphone-array in a room and time-delay estimates is provided by [Pineda (2010)], with focus on a geometric analysis and under optimization criteria. In [Valin (2003)] and in [Valin (2004)], a robot with 8 microphones was used to localize moving sound sources. The work of [Markowitz (2014)] gives a broader perspective on how people can interact with robots by using speech.

This paper is structured as follows: Section 3 will introduce the accoustic sensor setup and setup of lidar sensors, which will be used to evaluate the presented approach. Section 4 will describe the applied and implemented algorithms with an emphasis towards the sensor fusion method in this approach. In Section 5 we will perform experiments and present the results. Section 6 will summarize the approach and will give an outlook for future work.

3. SENSOR SETUP

As a test platform, we used the autonomous car named "MadeInGermany" from Freie Universität Berlin, cf. Fig. 1. The car is fully equipped with a combined lidar system from Ibeo, including 6 laser scanners, as shown in Fig. 2, a second 64 ray lidar sensor from Velodyne, in addition 7 radar sensors for long and short distance perception, at least 5 different cameras for lane marking and traffic light detection, including a stereo camera system for visual 3D algorithms, and a highly precise GPS unit. The car can be operated via a CAN-bus interface, thus, no further actuators are necessary to operate the throttle or brake pedals.

Different configurations were tried for the Kinect camera devices. To be independent from rain or snow and also to avoid wind noise while driving, we would have preferred to put the acoustic sensors inside the car. Unfortunately, the disadvantage of this configuration would have been the weaker signal strengths as well as signal reflections inside the vehicle. Therefore, we decided to mount both Kinect devices outside on the roof of the test platform, see Fig. 3.

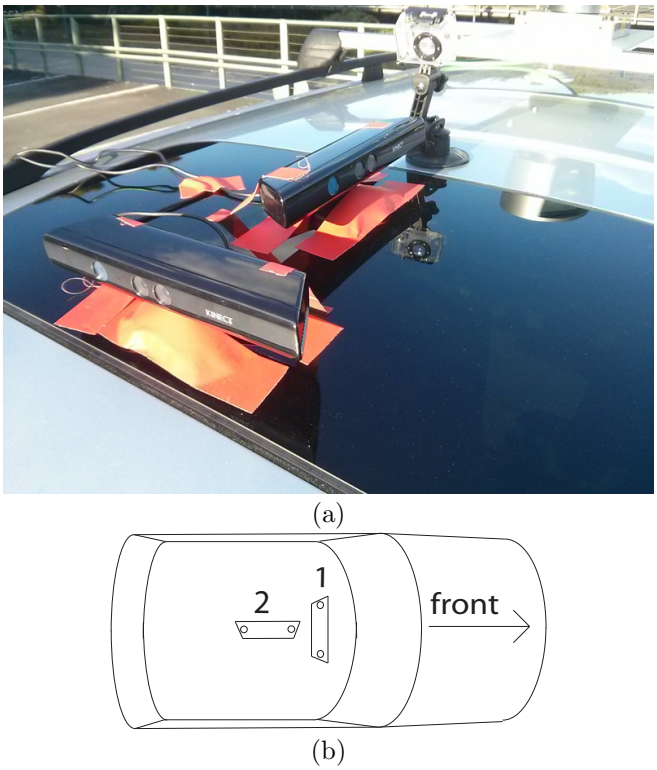

Fig. 3. Kinect sensor setup. (a) the device in the lower left corner is facing to the front of the car, the other one to the left. (b) View from above.

3.1 Kinect Sensor

Each Kinect sensor is equipped with four different, non-equally spaced microphones which are aligned in line, cf. Fig. 4. As a result of this configuration, only pairs of microphones are linearly independent. To achive the highest precision for an angle estimation, we decided to use the two microphones with the largest distance to each other, i.e., the two outer microphones on the left and right side of the Kinect device, depicted in Fig. 4. Another advantage is that the signal strength for those microphones is almost equal. This is not necessarily true for the inner two microphones which are located more inside the kinect case.

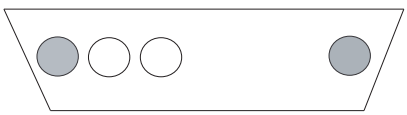

Fig. 4. Kinect microphone configuration, 4 mics are aligned in a line, we used to two outer mics (gray).

In the next section we want to describe how the sound source estimation and sensor fusion of the two Kinect devices was implemented.

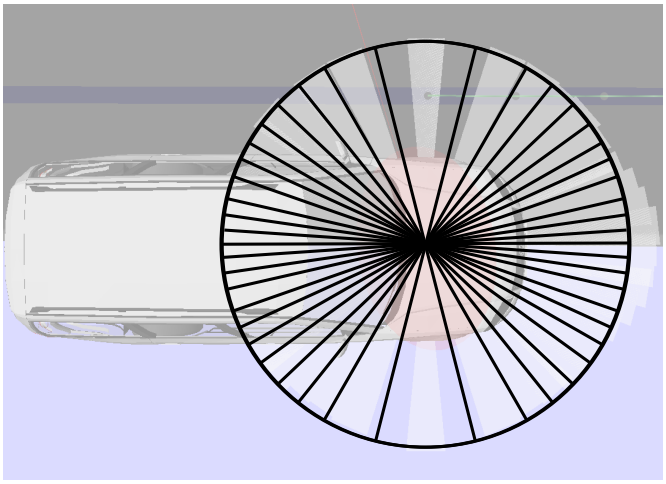

Fig. 5. Each shift between the two microphone signals corresponds to a segment (an interval) of possible angles, given that shifts can take only integer values.

4. SOUND SOURCE LOCALIZATION

In this section we are going to show how to localize an object using two microphones only. Furthermore we will focus on the direction accuracy given all possible directions. In the second part we will show how the resulting angular probabilistic distribution functions of two Kinect devices can be combined. One advantage of this method will be to constrain the set of possible solutions.

4.1 Calculation for one Kinect with two microphones

Estimation of the sound source using two microphones was designed straightforward using a cross-correlation function over the two microphone signals. Given the signal of the left microphone f and the right one g, for a continuous signal the cross-correlation function $f \star g$ with respect to the shift τ can be calculated as:

$$(f \star g)(\tau) = \int_{-\infty}^{\infty} f(t) \cdot g(t+\tau) \quad dt \tag{1}$$

Since we handle digital signals, for discrete functions the cross-correlation is calculated similarly with respect to a given shift n between the two signals f and g:

$$(f \star g)[n] = \sum_{-\infty}^{\infty} f[m] \cdot g[m+n] \tag{2}$$

Now we want to take a look at the real Kinect audio signals. Both Kinect microphones were sampled with 16800 Hz. For every calculation step we compared 216 data points from the two signals with a shift n ranging from -20 to +20. These 216 data points (provided by a module including the open source libFreenect library) showed to be sufficient for the cross-correlation calculation and allowed us to estimate the sound direction with more than 70 Hz. Each shift between the two signals would result in a certain direction. Regarding the number of possible shifts between the two signals, the two outer microphones of the Kinect are about 22 cm apart, we therefore assumed a base distance of $b = 0.22m$. With the speed of sound at $v_s = 340\frac{m}{s}$ at sea level and with a sampling rate for each microphone of $f_k = 16800Hz$, there is a maximum and a minimum value for possible shifts. These two boundaries correspond to the sound source being perfectly on the left or on the right side of the device. The maximum and minimum shift can be calculated as:

$$\begin{aligned} n_{max} &= b \cdot f_k \cdot v_s^{-1} && (3)\\ &= \frac{0.22m \cdot 16.8kHz}{340ms^{-1}} && (4)\\ &\approx \quad 11 && (5)\\ n_{min} &= -b \cdot f_k \cdot v_s^{-1} && (6)\\ &\approx -11 && (7) \end{aligned}$$

, resulting in approx. 22 possible values for shifts, making it sufficient to check these 22 possible shifts. As we will see later, on a planar surface with two microphones there are ususally two solutions for each signal shift (except for $n_{min} = -11$ and $n_{max} = 11$). Thus, we can map each shift n to two angular segments (angular intervals) which are symmetrically located with respect to the connecting line between the two microphones. The angular segments (or intervals) are depicted in Fig. 5.

The calculation of the corresponding angle for a given signal shift is straightforward, too. Given the speed of sound v_s we can translate each shift n into a distance $n \cdot v_s$. Now we have a triangle with a base length of $b = 0.22m$ and a known difference of the two other sides of $n \cdot v_s$ towards each other. Since the real distance to the sound source is unknown, we have to make an assumption, e.g., 25 m (the result of the calculation converges for higher distances) and can solve the angle to the object for each microphone using the Law of Cosines. A geometric sketch of the triangle is shown in Fig. 6.

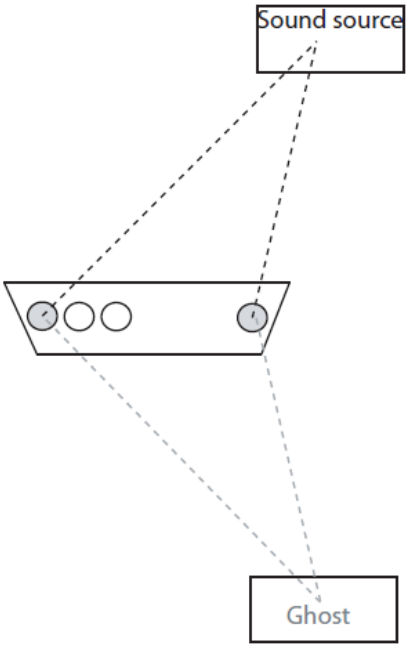

Fig. 6. Given the base distance of the triangle, the difference of the two sides and an assumed far distance (for drawing reasons the distance here is very close) to the object, the angles of the object to each microphone can be calculated - and should converge with increasing distance. Two solutions remain.

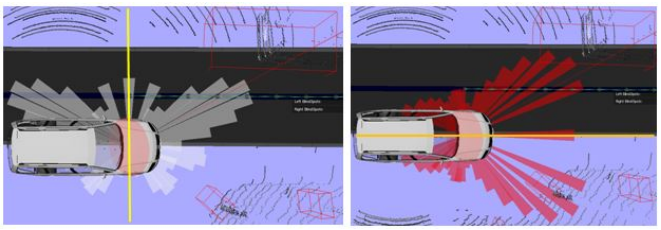

Fig. 7. Symmetry of angular distribution for front facing Kinect (left) and sideways facing Kinect (right). Symmetry axis depicted in yellow.

4.2 Sensor fusion of two Kinect devices

Since the two Kinect devices are not synchronized, we cannot just combine the data of the four outer microphones for triangulation. Moreover, we decided to combine the resulting probability distributions, cf. Fig. 5 of the Kinect devices with each other. As mentioned earlier, the probability of each segment containing the angle to the sound source is calculated from the cross-correlation function. Since both Kinect devices are rotated to each other by 90 degrees, the segment sizes do not match and thus cannot be combined directly. To overcome this problem, we subsample the angular segments for each Kinect with 64 equally-spaced angular segments. In a next step, after we generate the two equally spaced angular segment sets, we can combine them by pairwise multiplication of the probabilities of two corresponding segments, i.e., segments that contain the same angles. As a result of this combination via multiplication, we get a final segment set which represents the resulting probability distribution for both Kinect sensors (belief distribution). While each Kinect device alone cannot distinguish between objects in front and objects behind (see symmetry depictions in Fig. 7), after combination with the second sensor, those symmetries vanish. We show the calculation schematically in Fig. 8 and a step by step calculation with experimental data in a real traffic scenario in Fig. 9.

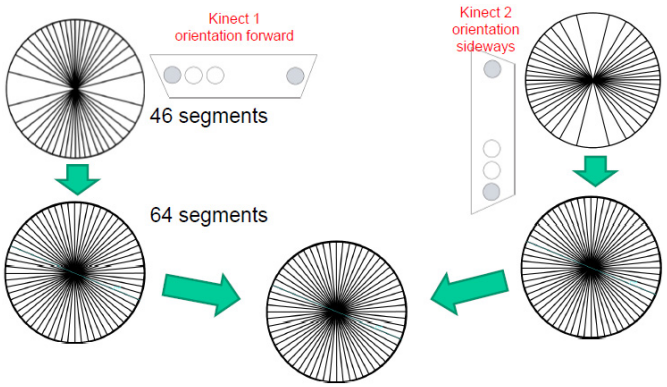

Fig. 8. Schemtatic calculation. The upper two segment sets result from the two different Kinect sensors. Since the segment sizes of the two sets are not equally aligned with respect to each other, we need to subsample them seperately into two segment sets with 64 equally sized segments. In a next step, they can be combined via pair-wise multiplication into a final segment set.

After sensor fusion, the resulting segment set corresponds to a probability distribution (belief distribution) of possible directions, i.e., where the sound source is located. To calculate a discrete value for the most likely direction, we selected the segement with the highest probability value assigned and took the mean value of that particular segment as the resulting angle. There would have been more sophisticated methods, e.g., integrating over different segments; also we thought about how to calculate directions to multiple sound sources but left this open to future research work.

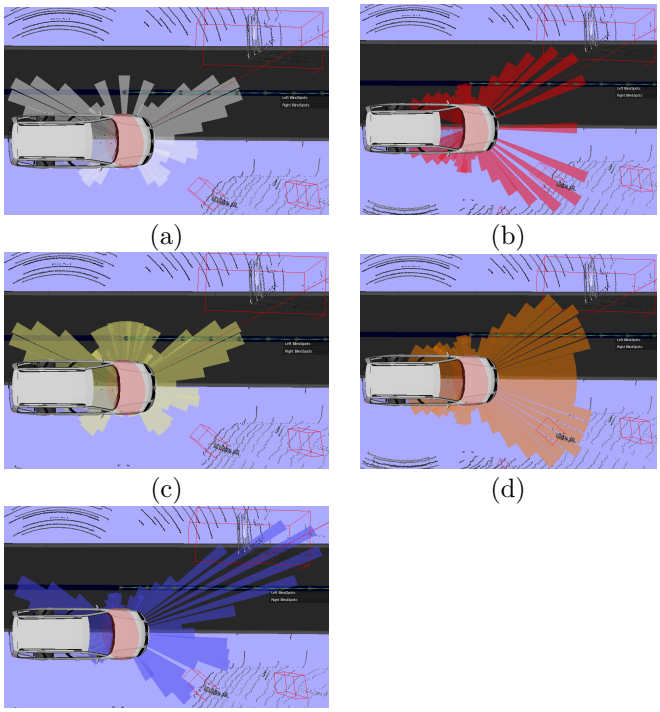

Fig. 9. Illustration of the approach, sound source vehicle in the upper right corner. Segment lengths correspond to cross-correlation amounts of the underlying signal shift and can be interpreted as a probability for the sound source lying in that angular interval. (a) Non-uniform segments for front facing Kinect and (b) left facing Kinect; (c) uniform (equally spaced) segments for front facing Kinect after subsampling, (d) uniform segments for left facing kinect; (e) uniform segments after combining (c) and (d), the resulting probability distribution (belief) for the sound source direction.

5. EXPERIMENTAL EVALUATION

As mentioned above, the proposed algorithm was implemented for our autonomous vehicle and tested in a real traffic scenario. The algorithms were tested within a modular robotics framework, the Open Robot Control Software Project Orocos (2011) under an Ubuntu 12.4. 64bit operating system. The data from both Kinect sensors were integrated into our AutoNOMOS software project and time stamped to compare them with our lidar sensory data. The six lidar Lux sensors from Ibeo run with a frequency of 12.5 Hz, the Kinect sensors ran with 70 Hz.

5.1 Test scenario

We tested our approach in a Berlin traffic scenario, close to the campus of the Freie Universität Berlin. Because driving the car was causing a lot of wind noise, we decided to park the car on the road side of the Englerallee, a medium-sized traffic road with trees, parked cars and houses on the side. Vehicles on the street were maintaining a velocity of 50-60 km/h (approx. 35 mph). Since there are trees on the middle strip seperating the two road lanes, cars of the more distant lane were partially occluded by trees while passing. We were interested in the angular accuracy of our approach in comparison to object angles from the lidar sensor. Therefore, data from the lidar sensor (a point cloud) was clustered into 3d-objects and tracked over time, resulting in a position and velocity vector for all clustered lidar objects. Since we were interested in moving (non-static) objects only, we compared the calculated angle from audio data to the closest angle of a moving object (from lidar).

5.2 Experimental results

In Fig. 10 we evaluated the angular error over time. We therefore took more than 5000 measurements, the resulting error-over-time function is depicted in Fig. 10.

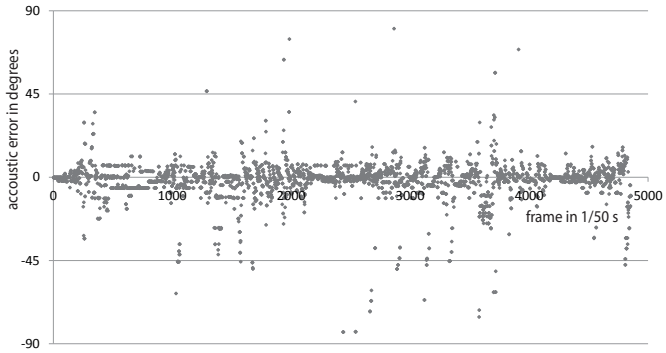

Fig. 10. Experimental evaluation: Difference between the angle from Kinect data to lidar data over time, 5000 data points were recorded. The standard deviation for the difference is $\sigma = 10.3$ degrees.

We also plotted the angular errors over all distances to the objects, as can be seen in Fig. 11. What is interesting, the highest angular errors occurred not for the farest objects but for objects within medium distances. One explanation could be that objects very far away would occupy a very small angular segment in the laser scanner, while objects closer occupy larger angular segments. Since the laser scanner always takes the center point of the detected object as a reference, and since the Kinect sensor will receive the loudest noise from the closest part of the vehicle, which is usually not the center of a car but the approaching front or leaving rear, this might be one reason for an increased detection error. Another reason could be increased reflection of noise on houses or trees for certain distances, which need further analysis.

In Fig. 12 we plotted the standard deviation of the angular error for different distance intervals, which showed the same result in terms that medium distances generated the highest error rates. The calculation time of the algorithm

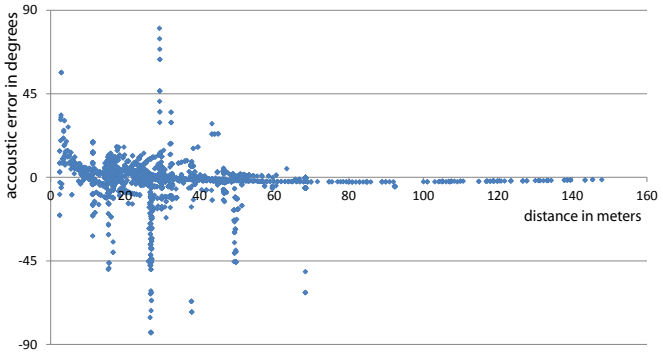

Fig. 11. The angular error over different object distances (measured by lidar). Higher error rates occured for medium distanced objects.

was negligible so that all experiments were performed under realtime constraints

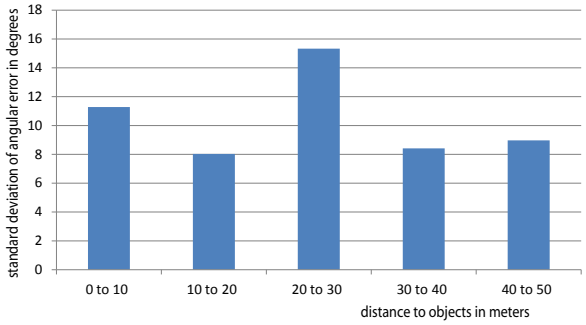

Fig. 12. Angular detection error for different distance intervals. While for high distances the angular error standard deviation was about 9 degrees, for medium distances it was approx. 15 degrees.

6. CONCLUSION

We presented, implemented and tested an approach which allows a road vechicle, equipped with to off-the-shelve Kinect cameras to localize objects in a distance of up to 50 meters and with a velocity of 50-60 km/h. We showed how to combine probabilistic density functions from two Kinect microphone devices using equally spaced angular interval segment sets, which helped to disambiguate possible angular locations while keeping the whole belief distribution. The algorithm can easily perform under real-time constraints with a frequency of 70 Hz. We also showed how the acoustically derived angle to the sound source could be assigned to moving objects from lidar sensors.

6.1 Future work

Future work needs to focus on localization and tracking of multiple objects, since in real traffic scenarios there are usually multiple vehicles in close proximity. Handling wind noise will be a crucial and challenging task for sound localization while in motion. Noise reflections on trees,

buildings and cars provide another challenge. Distance estimation, at least to some extend could support the data fusion problem with objects from other sensors. Band pass filters, e.g., application of Fast Fourier Transformation (FFT) shall be considered in future works. FFT can help to select specific signals, e.g. emergency vehicles with certain signal horn frequencies and signal patterns. Here the detection of alternating sound frequencies, as for emergency horns, would be helpful, too. Another research path worth following could be acoustic object tracking and velocity estimation, taking advantage of the doppler effect, i.e., the change of a frequency spectrum for an approaching or leaving vehicle.

REFERENCES

J. Benesty, J. Chen, Y. Huang, J. Dmochowski: On microphone-array beamforming from a mimo acoustic signal processing perspective. *In: IEEE Transactions on Audio, Speech and Language Processing*, 2007.

M. Frechette, D. Letourneau, J.-M. Valin, F. Michaud: Integration of Sound Source Localization and Separation to Improve Dialog Management on a Robot. *In: Proceedings of the IEEE/RSJ International Conference of Intelligent Robots and Systems (IROS)*, 2012.

Xiaofei Li, Maio Shen, Wenmin Wang, Hong Liu: Real-time Sound Source Localization for a Mobile Robot Based on the Guided Spectral-Temporal Position Method, *In: International Journal of Advanced Robotic Systems*, 2012.

Hong Liu: Continuous sound source localization based on microphone array for mobile robots, *In: Proceedings of the IEEE/RSJ International Conference of Intelligent Robots and Systems (IROS)*, 2010.

Rong Liua and Yongxuan Wanga: Azimuthal source localization using interaural coherence in a robotic dog, *In: modeling and application, Robotica / Volume 28 / Issue 07, Cambridge University Press*, December 2010.

Markowitz: Robots that Talk and Listen, *Technology and Social Impact*, 2014.

Tom Mitchell: Machine Learning, *McGraw Hill*, 1997.

J. Murray, S. Wermter, H. Erwin: Auditory robotic tracking of sound sources using hybrid cross-correlation and recurrent networks, *In: Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, 2005.

Open Robot Control Software Project *http://www.orocos.org/*, 2011.

Xavier Alameda Pineda, Radu Horaud: A Geometric Approach to Sound Source Localization from Time-Delay Estimates, *In: Robotica*, 12/2010.

M. Schnrmacher, D. Ghring, M. Wang, T. Ganjineh: High level sensor data fusion of radar and lidar for car-following on highways, *In: Recent Advances in Robotics and Automation*, 2013.

S. Thrun, W. Burgard, D. Fox: Probabilistic Robotics, *MIT Press*, 2005.

J.-M. Valin, F. Michaud, J. Rouat, D. Letourneau: Robust Sound Source Localization Using a Microphone Array on a Mobile Robot, *In: Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, 2003.

J.-M. Valin, F. Michaud, B. Hadjou, J. Rouat: Localization of Simultaneous Moving Sound Sources, *In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2004.