3D Audio Perception System for Humanoid Robots

Norbert Schmitz, Carsten Spranger, Karsten Berns
University of Kaiserslautern
Robotics Research Laboratory
Kaiserslautern D-67655
{nschmitz,berns}@informatik.uni-kl.de

Abstract

An audio system is one of the basic components of a humanoid robot designed for natural interaction. For many interaction purposes it is sufficient to use the sound detection and localization as attention system for the vision system. In this paper the audio perception module of the robot ROMAN is presented including the integration into the existing control structure and the localization algorithm using a microphone array with 6 microphones. The reduction of data into so called sector maps is presented and the interaction with the control architecture is shown.

1. Introduction

The goal of autonomous natural interaction is a highly complex task due to the complexity of the communication process. An important aspect is the realization of a hearing sensor which is able to detect and localize sound activity in the environment of the robot. Besides the recognition of the speaker this ability is necessary to detect interesting activities which are not inside the field of vision.

The hearing can be seen as support or guidance for the vision system. Often persons hear a sound like the shutting of a door and turn toward the sound source to see what has happened. The goal of the attention system presented here is the observation of the environment which is not inside the view of the robot. This implies that the precision is not very important since the vision system will search for an object of interest in the surrounding of the detected area. Figure 1 shows a typical interaction situation with the robot head turned toward the interaction partner.

Various approaches for sound localization are presented in literature which can generally be divided into two main groups: Localization with two microphones similar to the human ears and localization using an array of microphones.

[1] uses a microphone pair with the cross-correlation or alternatively the dual-delay line algorithm. The frequency range in this work is limited to the range from 100Hz to 4.000Hz which is suitable for human speech. It is restricted to 2D-localization with the drawback that the localization is not unique due to the number of microphones.

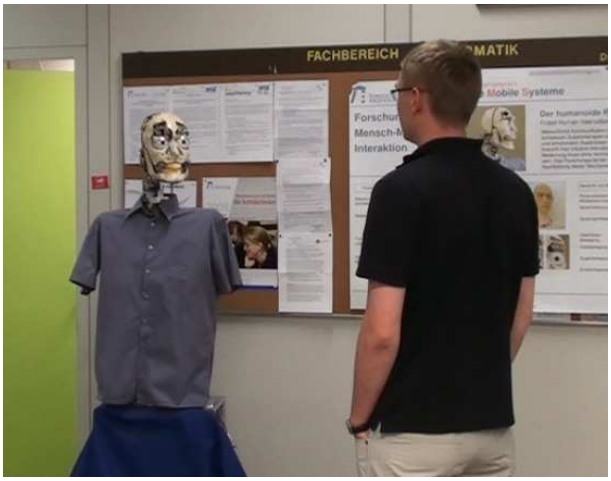

Figure 1. A typical interaction situation in an office environment.

In [4] the time difference of arrival (TDOA) is calculated using the generalized cross correlation in the frequency domain. Experiments with different weighting functions, phase transformations and filters are presented. With the additional analysis of speech it is possible to track different persons with temporal overlapping.

[3] uses the properties of a specially shaped ear to localize sound sources in the vertical and horizontal plane. For the localization the delay of arrival, difference of amplitude and the direction-dependent spectral difference caused by the shape of the ear is used.

Approaches using microphone arrays include the work presented in [7]. Here an array of eight microphones is used to localize a single sound source using the TDOA method. The presented approach is claimed to be frequency indepen-

dent with a resolution of 3° within a maximum distance of 3m.

A very promising approach is described in [6]. Here a beam-forming algorithm is used for localization and additionally a particle filter is applied to realize the tracking of multiple objects.

Humanoid robots designed for natural interaction require a sound localization system to be able to react to sound input stimuli like speech or screams. On the other hand it is not required to exactly follow those sound sources since these objects directly attract the visual attention of the robot. Based on the research results described before the sound localization is realized using the beam-forming algorithm. Besides the localization itself the infrastructure to integrate the audio attention into the control system of the humanoid robot ROMAN is presented. Further details concerning the implementation of the audio system can be found in [5].

The following chapter will briefly introduce the localization algorithm before section 3 describes the implementation of the presented approaches. Experimental results with test scenarios and the humanoid robot ROMAN are shown in section 4.

2. Sound Localization

The principle of the beam-forming algorithm is to select several points in the environment of the robot and check the probability that a sound is located at that point by delaying the microphone signals according to the distance to the selected point. Using a large amount of those points enables a localization of points with high energy which can be used for a variety of tracking algorithms. For the humanoid robot ROMAN several spheres with equally distributed points on the surface are chosen. All spheres have different radii and identical center points which are located in the selected middle of the microphone array Given these points the beamformer energy which is an indicator for a sound source can be calculated according to the description in [6].

The output $y(n)$ of a delay and sum beam-former at sample index n is defined as

$$y(n) = \sum_{m=0}^{M-1} x_m(n - \tau_m) \tag{1}$$

where M is the number of microphones, m is the index of a microphone, $x_m(n)$ is the signal of the mth microphone and τ_m the delay of sound arriving at the mth microphone from a specified point.

The output energy E is defined as

$$\begin{aligned} E &= \sum_{n=0}^{L-1} [y(n)]^2 \\ &= \sum_{n=0}^{L-1} [x_0(n-\tau_0) + \cdots + x_{M-1}(n-\tau_{M-1})]^2 \end{aligned} \tag{2}$$

where L denotes the length of an audio frame. This equation can be rewritten as

$$\begin{aligned} E &= \sum_{m=0}^{M-1} \sum_{n=0}^{L-1} x_m^2(n-\tau_m) \\ &+ 2 \sum_{m_1=0}^{M-1} \sum_{m_2=0}^{m_1-1} \sum_{n=0}^{L-1} x_{m_1}(n-\tau_{m_1}) x_{m_2}(n-\tau_{m_2}) \end{aligned} \tag{3}$$

Using the cross-correlation $R_{x_{m_1},x_{m_2}}$ and the fact that the first term is constant and can be replaced by K we get

$$E = K + 2 \sum_{m_1=0}^{M-1} \sum_{m_2=0}^{m_1-1} R_{x_{m_1},x_{m_2}}(\tau_{m_1} - \tau_{m_2}) \tag{4}$$

In the frequency domain the cross-correlation can be approximated with

$$R_{ij}(\tau) \approx \sum_{k=0}^{L-1} X_i(k) X_j(k)^* e^{j2\pi k\tau/L} \tag{5}$$

where $X_i(k)$ is the discrete Fourier transform of $x_i[n]$, $X_i(k)X_j(k)^*$ is the cross spectra of $x_i[n]$ and $x_j[n]$ and $(.)^*$ is the complex conjugate.

3. Realization

The control system of ROMAN consists of the basic modules habits, motives and percepts (see [2] for details). The audio group is part of the perception group and therefore has the functionality to provide information about the environment. The audio group is organized in five modules as shown in Fig. 3:

jack_capture The capturing module realizes the communication with the sound card using the jack daemon.

jack_filter The filtering module is intended for filtering operations like noise reduction or band pass filtering.

beam-forming This module realizes the localization algorithm.

particle_filter This module uses the output of the beamformer to generate a reduced information of the environment which is suitable for the control of ROMAN.

visualization This module represents the visual 3d output of the localization process (mainly for debugging reasons).

The whole sound localization system can be configured using a simple syntax as listed in Fig. 2. The configuration includes the sound capturing, the position of the microphones and the refinements of the beam-forming algorithm. The following sections will describe the configuration and the basic modules in detail.

```
jack{
    max_number_of_channel:6
    capture_buffer_size:4096 # samples
    sample_rate:48000 # Hz
}
micro_1{
    posX:-0.038 # m
    posY:-0.186 # m
    posZ: 0.262 # m
}
...
refinement_count{
    refinement_distance_count: 4
}
refinement_distance_1{
    refinement_distance: 0.35 # m
}
...
```

Figure 2. Listing of the configuration file for the audio perception group.

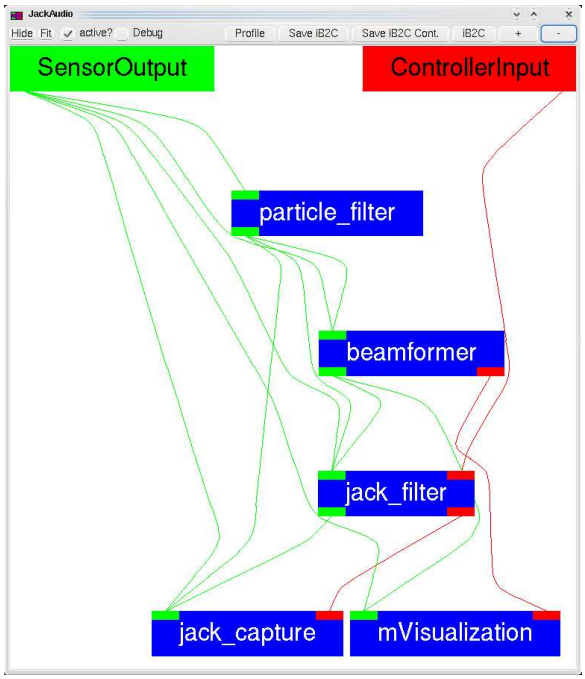

Figure 3. Structural overview of the audio group with the basic modules jack_capture, jack_filter, beam-forming and particle_filter. Additionally the visualization module is displayed.

3.1 Capturing

The capturing module realizes the connection to the sound card using the jack sound server[1] and the alsa sound driver[2]. The advantage of this combination is the flexibility of the jack audio daemon. With this framework it is possible to start the localization running on the real input of the microphones as well as on recorded or simulated sound sources. Additionally it duplicates the input signals which enables the system to use the same audio stream for localization and speech recognition.

The capturing module stores the samples provided by the jack daemon until the desired buffer size is reached and then notifies the filter module that new sample data has arrived.

3.2 Filtering

The filtering module is intended for basic filtering of the input signals like band-pass filtering or noise reduction. Since speech signals are basically located in the range of 100Hz to 10kHz very high and very low frequency signals are not of interest for interaction. The output of the filtering is passed to the beam-forming module.

3.3 Beamforming

The output of the filtering module is read by the beamformer. The energy calculations are realized in this module as described in section 2. The calculation can alternatively be performed in frequency or in time domain with an improved performance in the frequency domain. The output of this module is an energy value for each point in the observed spheres.

3.4 Post-Processing

The immense amount of data – 5 times 2560 energy values per time step – cannot be used directly as input information for the robot control. Two different implementations are realized for tracking a single object and for a generalized overview of the environment.

The tracking of a single object is implemented using a two step average. In a first step the average energy over the

[1] http://jackaudio.org/
[2] http://www.alsa-project.org/

last n cycles for each point in the spheres is calculated. In our experiments we use a value of $n = 20$ which results in an average over the last second since one time step lasts $50ms$. In a second step the l points with maximal average energy are chosen and a weighted average point is calculated. In the experiments l is set to 20. This average calculation reduces the amount of values to 3 – Position X,Y,Z – and is used in the experiments described in section 4.

A second possibility of data reduction uses the sector maps depicted in Figure 4. The sector map is divided into 36 sector with an opening angle of 10 degrees. Each of these sectors is again divided into several chunks depending on the radius of the spheres defined for the beam-forming algorithm. Each sector contains a value in the interval $[0; 1]$. A low value indicates a low sound activity in that sector. Using these sector maps the amount of data stored in a single map is the number of spheres plus one multiplied with the number of sectors – here $5 \cdot 36 = 180$ values.

The value of each sector is calculated by mapping each point on a sphere to the nearest segment. The normalized average value of all points $avg_i \in [0; 1]$ and the update factor $\alpha \in [0; 1]$ are used to calculate the sector value v_i at time step t.

$$v_i^t = v_i^{t-1} \cdot (1 - \alpha) + avg_i \cdot \alpha \quad (6)$$

The 3D information captured by the localization system is used by introducing multiple sector maps at different heights. A reduction to 3 sector maps – one in the center, one $50cm$ above and one $50cm$ below – is a suitable selection concerning precision and amount of data.

3.5 Visualization

The visualization module shows the output of the beamforming algorithm and the information reduction in the sector maps. Depending on the calculated beam energy a color is assigned to every point on the localization spheres where blue indicates a low and red a high energy. The energy is normalized to the maximum energy of all measurement points in one time frame. Additionally to the points the sector maps are visualized. Each sector is color coded corresponding to the coloring of the measurement points. Figure 5 shows a screen shot of the scene during runtime. Each measurement point of the beamforming algorithm is visible so that the spheres can be seen. Additionally the three sector maps are sketched.

4 Experiments

The evaluation of the sound localization system has been performed in three steps. In each step the recursion depth of the spheres is 4 which leads to 2560 points for each sphere. The experiments are performed with 5 spheres located at distances of 8m, 2m, 0.95m, 0.65m and 0.35m.

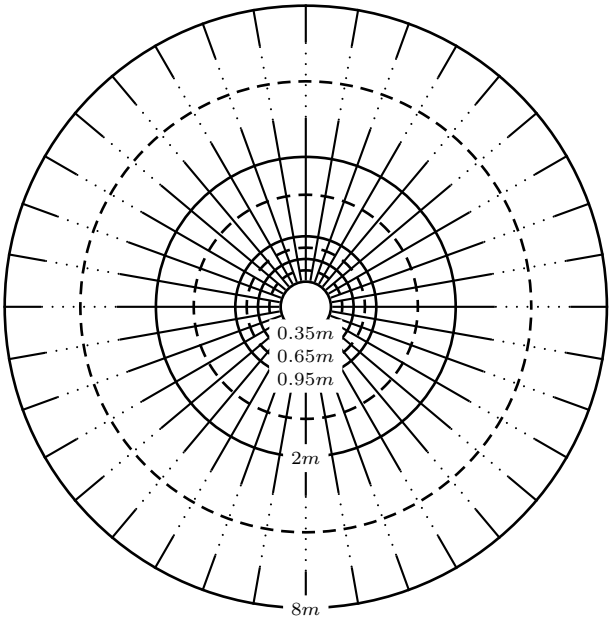

Figure 4. Top view of a sector map with the beam-former spheres shown as solid lines. The sector map has 36 sectors and each sector is divided into 5 distance fields (borders are drawn with a dashed circle) defined as the center distance between two spheres.

In the first phase the beam-forming and tracking is verified using a playback of an previously recorded sound. An arbitrary point in 3D space is selected and the delay of arrival for each microphone is calculated. The channels corresponding to the microphones are delayed to simulate a sound from that specific source. This experiment proves the principal functionality of the localization without reflections or any other real environment effects.

The second test is realized on a microphone array consisting of four microphones arranged in a regular tetrahedral structure. As sound source a small transistor radio has been selected which plays some recorded news. The experimental results have shown that a regular arrangement of microphones has drawbacks since the delay between to microphone pairs is often equal.

The last experiment – which is described here in detail – is realized using the microphone array mounted on the humanoid robot ROMAN. The robot is placed in a cluttered office environment and specific points in the direct neighborhood of the robot are marked and the distance to the robot is measured. Figure 6 shows the experimental setup with the marked positions on the floor and the robot in the background. During the experiment the previously mentioned transistor radio is moved every 2.5sec to one of the thirteen positions and located on a pole to include changes

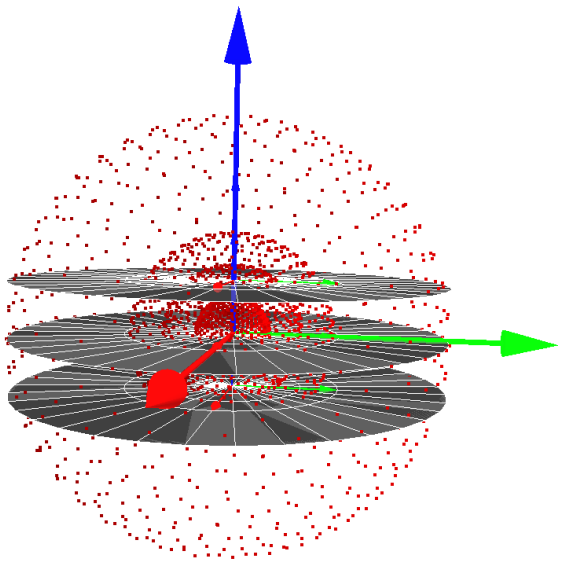

Figure 5. Example screen shot of the beam-former visualization. This visualization displays the normalized energy of each point on one of the available spheres. The color indicates the energy in that specific point. The sector maps used by the robot are also plotted.

of height. The maximum energy output of the beam-former as well as the filtered output has been recorded.

Figure 8 shows the filtered output as 3d plot as well as single plots for each distance in X,Y and Z direction. The true position of the sound source is plotted to measure the localization error. Although the maximum error is up to 1.5m in distance (X direction from time step 350 to 450) the average distance error is below 0.5m. Transferred to polar coordinates the angular error is within $4.5°$.

5. Summary and Outlook

In this paper the audio perception module of the humanoid robot ROMAN has been presented. Based on the flexible capturing module using the jack audio server various simulation and real scenarios can be realized. The localization module is implemented using the beam-forming algorithm which assigns a sound probability to each virtual measurement point in the environment of the robot. The filtering module finally uses the output of the beam-former to generate sector maps which reduce the large amount of data to the essential information of the localization process. These maps are then transferred to the control architecture

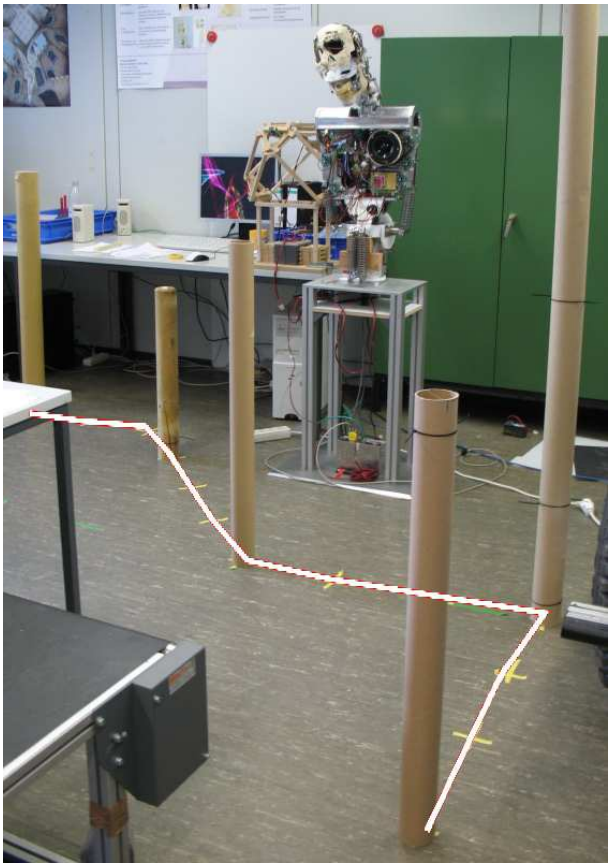

Figure 6. Experimental setup in the laboratory. In the background the humanoid robot ROMAN and the microphone array are visible. The white line marks the path of the sound source described in the final experiment.

of the robot.

Future extensions of the audio perception module should extract further information of the sound source. Characteristic properties of speech can be detected and the probability of speech can be added to the sector map representation.

6. Acknowledgments

This publication is supported by the DFG's International Research Training Group (IRTG) Visualization of Large and Unstructured Data Sets Applications in Geospatial Planning, Modeling, and Engineering

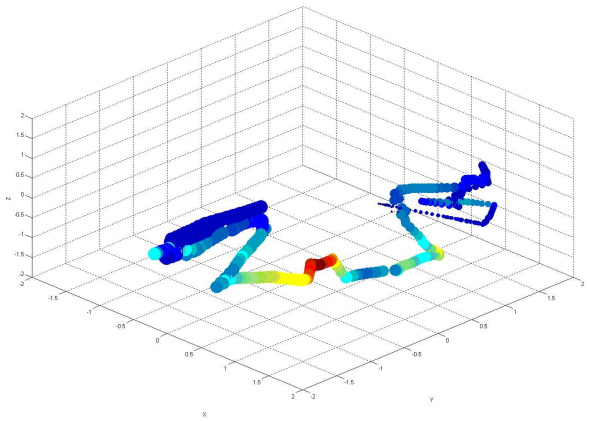

Figure 7. 3D plot of the estimated position of the sound source. The size of the circles are an indicator of time (small circles are estimated prior to big circles) and the color indicates the absolute energy (increasing energy from blue to red). The axis display the distance in the range $[-2, 2]$m.

References

[1] L. Calmes. A binauraul sound source localization system for a mobile robot. Master's thesis, RWTH Aachen, 2002.

[2] J. Hirth, N. Schmitz, and K. Berns. Emotional architecture for the humanoid robot head roman. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2150–2155, Rome, Italy, April 11-13 2007.

[3] J. Hoernstein, M. Lopes, and J. Santos-Victor. Sound localization for humanoid robots - buildind audio-motor maps based an the hrtf. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1170–1176, 2006.

[4] G. Klaassen. Speech-based localization of multiple persons for an interface robot. Diplomarbeit, University of Amsterdam, 2005.

[5] C. Spranger. Räumliche geräuschquellen-lokalisierung für interaktive humanoide roboter. Master's thesis, University of Kaiserslautern, 2008.

[6] J.-M. Valin, F. Michaud, and J. Rouat. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. In *Robotics and Autonomous Systems Journal (Elsevier)*, volume 55, pages 216 – 228, 2007.

[7] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau. Robust sound source localization using a microphone array on a mobile robot. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1228–1233, 2003.

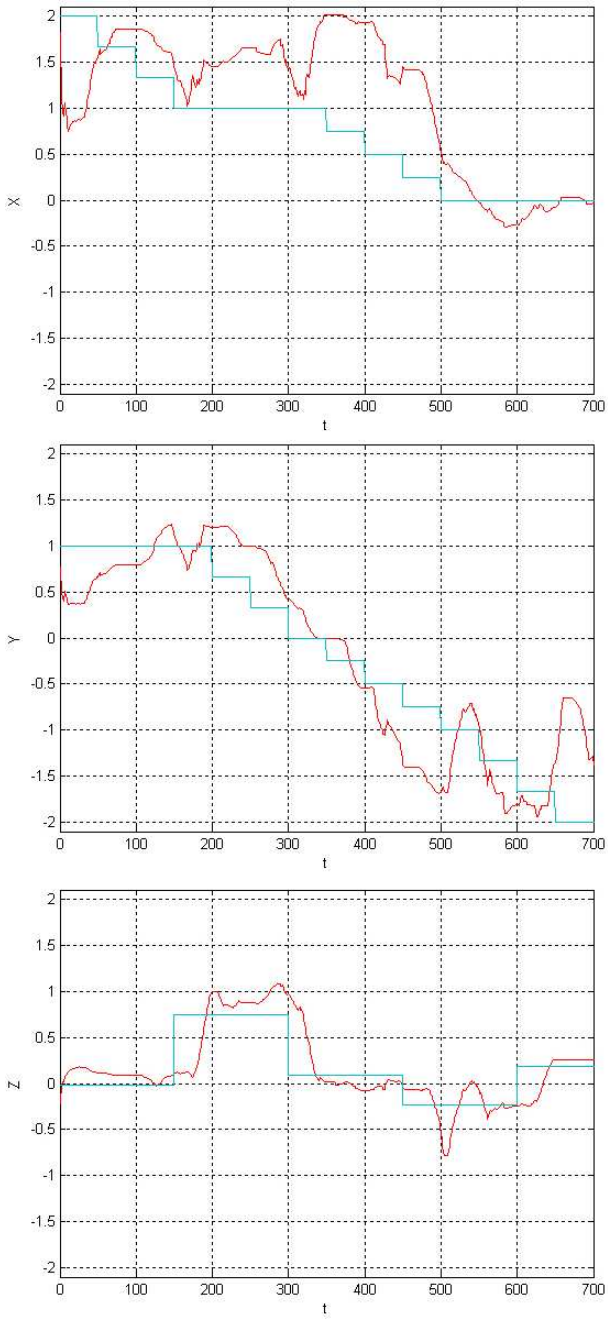

Figure 8. Recordings of the experiments with the humanoid robot (6 microphones). The plots –from top to bottom – show the calculated versus the true position of the sound source during the test run. The distance in meter is plotted versus the time step where each time step corresponds to 50ms.