Acoustic SLAM for unprepared environments with Gaussian Processes

B. Sc. Oliver Neumann
Dipl. Phys. Jana Mayer
Dr. Ing. Benjamin Noack

Karlsruhe Institute of Technologie (KIT)
Institute for Anthropomatics and Robotics (IAR)
Intelligent Sensor-Actuator-Systems (ISAS)
Adenauerring 2 76131 Karlsruhe Germany

Abstract. Simultaneous localization and mapping (SLAM) became one of the most important research fields in robotics. Modern approaches of SLAM are based on discrete landmarks that are used to estimate the position of the actor. But there are environments where no landmarks could be detected or the landmarks are not suitable for accurate localization. Instead of discrete landmarks, researchers used physical phenomena as continuous functions for localization purposes. Roughly, the continuous function is usually approximated with a Gaussian process as the map. Measurements then will be fused with an estimated position (e.g. from an inertial measurement unit) to increase accuracy. Those approaches have promising performance and there are a lot of use cases. From indoor robot cleaning to autonomous submarines for undersea pipeline investigation up to autonomous drones for weather observation. In this project we want to investigate if a speaker and microphone together with an inertial measurement unit (IMU) mounted on a robot, is capable of localizing itself in an unprepared environment. Therefore a specific signal will be sent via the speaker and from the microphone input, a continuous function will be derived based on the concept of room impulse response. This function will be approximated with Gaussian processes and should be used as input for SLAM.

1   Motivation

Robots often have to act in an unknown environment. For that, a robot has to recognize the environment and locate itself in it. For example, an actor wants to find the dishwasher to clean up the kitchen. This problem is also known as simultaneous localization and mapping (SLAM) in the context of robotics. Common approaches using discrete landmarks recognized by the robot to set up a landmark map of the environment (see figure 1). With this map of landmarks, it's possible to estimate the position of the robot. A typical example of landmarks are features from images, which are taken by a camera from the robot.

For SLAM there are a variety of use cases but there are also scenarios where it's not possible to obtain landmarks that are suitable for SLAM approaches. For

example, at the sea, water is usually muddy and cameras would only see over a short distance. There it wouldn't be possible to extract image features suitable for SLAM approaches. That's where scientists switched from discrete landmarks to continuous functions given by physical phenomena. A common approach is to use the magnetic field for indoor localization. This is also a good example to show also what performance such approaches could have.

But it's not only a workaround to use SLAM in situations, where discrete landmarks are not possible to obtain. Continuous approaches could be as good as discrete ones and in combination, they could be even better. Also, when it comes to costs if the aim of a robot is to measure physical phenomena, why don't use this data for SLAM instead of buying a camera and more expensive hardware for real-time feature extraction of the images.

There is a huge potential of SLAM with continuous functions and the science community is only at the beginning. This project wants to examine if an actor could locate itself in an unprepared unknown environment only by using a microphone, speaker and inertial measurement unit (IMU). For that, the speaker sends out a signal which is received by the microphone. But the signal will be changed mainly because of interference of the signal with the room.

One Task is now to derive a continuous function from the microphone signal. For that, the room will be treated as an electrical system. The signal from the speaker will be the input and the microphone signal the output of the system. So an impulse response could be calculated. This concept is already known as the room impulse response (RIR). Based on that, various continuous functions have to be validated which one is the best for indoor SLAM approaches. For example, using only the ten highest peaks of the RIR. But before that, data has to be recorded for analysis, test and validation purposes. At the end of this project, the performance of this indoor SLAM approach should be compared to other acoustic SLAM approaches and also to other indoor SLAM approaches with other physical phenomena like the magnetic field.
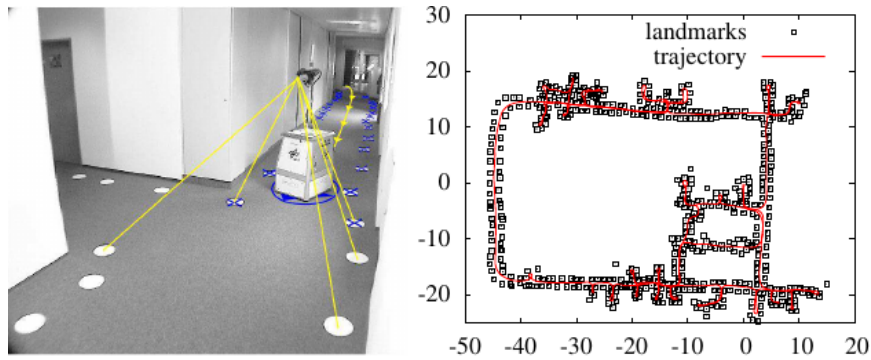

Fig. 1: An illustration for SLAM with a prepared environment for the robot. Landmarks are mounted on the floor which the robot can recognize (left image). The generated map is shown on the right [1].

2   State of the Art

2.1   Modern SLAM

SLAM describes the field of problems where an agent has to locate itself in an unknown environment. Therefore a map of the environment has to be created where the actor can locate itself in there [1]. The map could consist of discrete landmarks, measurements of continuous function, but also planes as an approximation of the environment. There are several SLAM approaches. All have in common that all measurements, as well as agent positions, are Gaussian distributed. That's why the map is only an estimation and all SLAM approaches itself are only probabilistic approaches [1].

SLAM could be subdivided into two common classes [1]. First, approaches where only the most probable current position of the agent within the map is wanted, are called filtering or on-line SLAM approaches. In contrast to that, when the current most probable full trajectory of the robot within the map is calculated, those approaches are called smoothing or full SLAM. Both approaches acting in real-time but only smoothing approaches calculating a full trajectory of the robot which could change over time.

Regardless of filtering or smoothing approaches, they often use Kalman or particle filter for current state estimation [1]. Kalman filter is a mathematical concept consists of two steps [2]. The prediction and update steps. In the prediction, the next state is predicted by using all data given until this timestep. In the update phase, the predicted state is improved by combining it with the new measured state. By changing the parameters it can be controlled if the Kalman filter should trust more the predictions or measurements. It is suggestive to trust more the predictions if the measurements are normally distributed with a large deviation. If the measurements have nearly no error, it would be better to trust the measurements. Kalman filter is a famous concept in the science community.
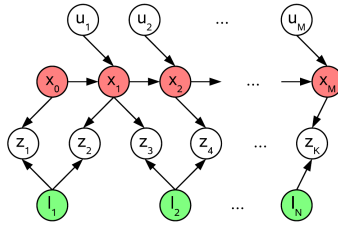

Fig. 2: SLAM graph with x_i as the robot position at time step i. l_j is the jth landmark. z_k is the kth by the robot measured landmark. u_i is the control input at time step i which is for example odometry data [3]. For understanding: At time step 1 the robot position was x_1 and it measured landmark l_1 and l_2 which can be seen by the measurements z_2 and z_3. The input from the control unit was u_1.

One large disadvantage is that Kalman filters are only following one hypothesis. For example, when tracking a car at a crossing it could be better to follow the hypothesis, that the car is going straight, but also the hypothesis that the car will turn right or left. When tracking using the Kalman filter, the car could be lost after the crossing. That's where the concept of particle filters was introduced [4]. In contrast to Kalman filters, particle filters can follow multiple hypotheses. To be exact, each particle can follow one hypothesis.

State-of-the-art SLAM techniques using graph-based approaches. The idea of those approaches is easy to understand. While the robot is moving and making measurements a graph is build up as illustrated in figure 2. With modern graph algorithms the position $x_0, ..., x_n$ can now be estimated. In figure 3 the performance of those approaches is illustrated.
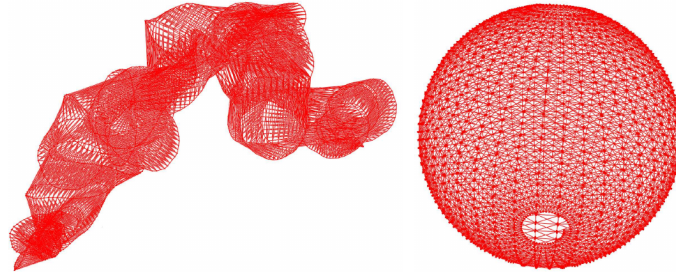

Fig. 3: A showcase of a current state-of-the-art graph-based SLAM approach. Initial position estimation on the left and next to it graph-based SLAM position estimation. The robot was moved in a simulation on the surface of a sphere [1].

2.2   Gaussian Processes

When switching from discrete landmarks to continuous functions, these functions need to be approximated within the area of the environment. So it is possible to have values between the measurements which is necessary to improve the position estimation of the actor. When knowing the input space (in this case the environment) but don't want to make any assumptions to the underlying function, Gaussian processes (GP) are the solution [5].

Basically, GP are defined by a mean function $m(t)$ and a covariance function $k(x_i, x_j)$ also called the kernel. If $m(t) = 0$ the GP are called centered. For the kernel radial basis functions can be used. Such GP are called radial. For a radial covariance function follows $k(x_i, x_j) = k(||x_i - x_j||)$. One very common radial kernel is the squared exponential kernel [5]

$$k(x_i, x_j) = \sigma^2 \exp\left(-\frac{(x_i - x_j)^2}{2l^2}\right).$$

As mentioned before, for SLAM it should be possible to model normally distributed measurements. With GP that is possible. Assume the noise is normally distributed with variance σ^2. Then measurements can be described like $y_i = f(x_i) + \mathcal{N}(0, \sigma^2)$. Combining the covariance function with the model of noisy measurements, that will lead to the slightly changed kernel

$$k(x_i, x_j) = \sigma^2 \exp\left(-\frac{(x_i - x_j)^2}{2l^2}\right) + \delta_{ij}\sigma_{noise}^2 \ .$$

In this kernel δ_{ij} is the Kronecker delta. This function is only 1 when $i = j$ otherwise it is 0. So, σ_{noise}^2 is only added to the measurements. Figure 4 gives an example of GP with and without noisy measurements.

Assume now there are n measurements $y_1, ..., y_n$ taken at $x_1, ..., x_n$. Then the covariance matrix would be

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \dots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \dots & k(x_n, x_n) \end{bmatrix} .$$

Estimating y_* at x_* and assuming the kernel is radial, would lead to a covariance matrix

$$\begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix} \text{ with}$$

$$K_* = [k(x_*, x_1), \dots, k(x_*, x_n)],$$

$$K_{**} = k(x_*, x_*).$$

Because of GP modeling data from a multivariate normal distribution, y and y_* can be modeled as

$$\begin{bmatrix} y \\ y_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix}\right).$$

Using the mathematics of multivariate Gaussian distributions, the most likely value of y_* can be calculated by

$$y_*|y \sim \mathcal{N}(K_* K^{-1} y, K_{**} - K_* K^{-1} K_*^T)$$

$$\Rightarrow \bar{y}_* = K_* K^{-1} y \ .$$

The derivation of the formula is well explained by Ebden [5]. Notice that also the variance for y_* is given by

$$var(y_*) = K_{**} - K_* K^{-1} K_*^T \ .$$

All that, uncertainty in measurements, prediction of y_* and probability of the prediction can be seen in Figure 4. Note that for linear increasing number of measurements the size of the covariance matrix rises quadratically and for y_* the inverse of that matrix has to be calculated. So it is computational expensive for large-scale data. There are approaches which facing that problem. Another remark is that the standard GP is not possible to model uncertainty in the input data. So it is not possible to model uncertainty in the position of the robot with GP.
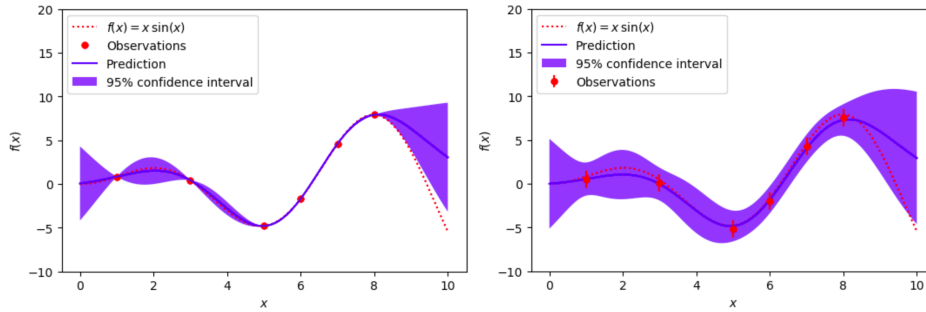
Fig. 4: A plot of two GP regressions with the same measurement points x_i but the one at the right has noisy measurements with variance $\sigma^2_{noise} = 1$. Also noticeable the probability of the predictions represented through the confidence interval.

2.3   Scalable Magnetic Field SLAM

One impressing approach from Manon Kok and Arno Solin showing full SLAM in 3D using the magnetic field and GP [6]. The magnetic field was approximated with GP and because of performance issues the environment is divided in 3D hexagons as illustrated in figure 5. So there is a GP for each hexagonal region. This reduces the size of the covariance matrix for each GP and increases performance.

To approximate the magnetic field, they use a combination of two kernels, the linear and squared exponential kernel and a mean centered around zero [6], that leads to the following Gaussian process

$$GP(0, k_{lin}(x_i, x_j) + k_{se}(x_i, x_j)) \text{ with}$$

$$k_{lin}(x_i, x_j) = \sigma^2_{lin} x_i^T x_j,$$

$$k_{se} = \sigma^2_{se} \exp\left(-\frac{||x_i - x_j||^2}{2l^2}\right).$$

For position estimation, Kok and Solin don't use any graph-based approach. Instead, they use a combination of Kalman and particle filter. To update measurements a Kalman filter is used. For predictions in time particle filters are used. A detailed description can be found in their paper [6]. Figure 5 demonstrates the performance of their approach. This paper shows vivid what SLAM approaches with continuous functions are capable of.
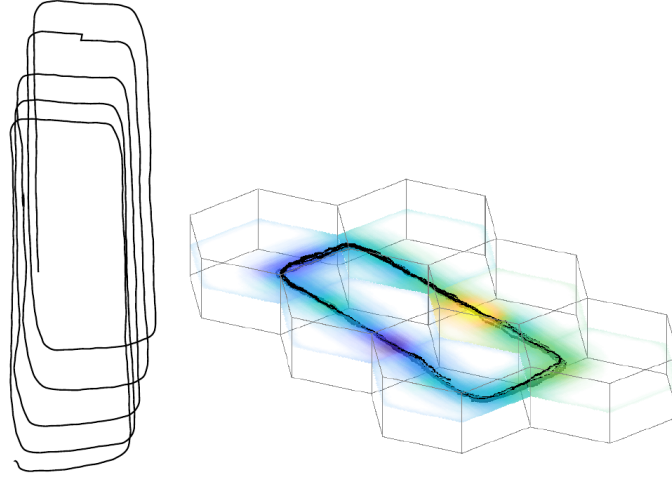
Fig. 5: Performance demonstration of the SLAM algorithm by
Kok and Solin [6]. The trajectory on the left shows the odometry
data collected by an iPhone 6s and next to it the resulting
trajectory with the approximated magnetic field.

2.4   Terrain Field SLAM

Another interesting approach from Hyeonwoo Yu and Beomhee Lee shows which
different continuous functions could be used for SLAM [7]. In their paper, they
described a SLAM algorithm where the continuous function is the vibration
caused by the interaction between the robot and the terrain. Their experimental
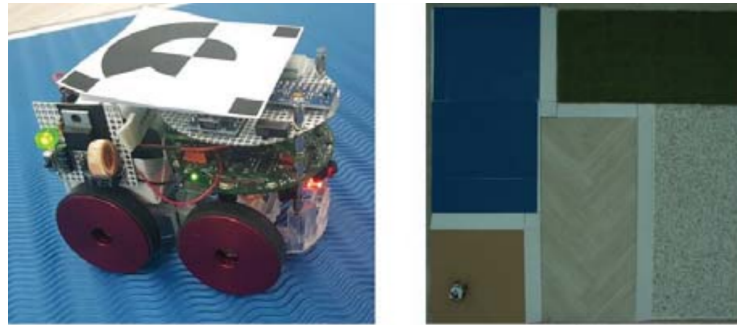setup and the robot can be seen in figure 6.

Fig. 6: Illustration of the setup by Yu and Lee. The mobile robot
at the left and on the right the environment for the robot to move
in. The environment containing five different subsurfaces [7].

Yu and Lee also used GP to infer the terrain field. A performance demonstration can be seen in figure 7. Even if they don't face the problem of doing SLAM and Gaussian regression on-line and there is a little deviation over time, the results are considerable.
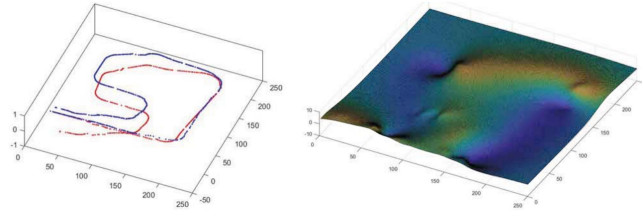

Fig. 7: Performance illustration of the terrain based SLAM approach by Yu and Lee [7]. Robot trajectory (red) on the left and ground truth (blue). Righ shows inferred terrain field.

2.5   Acoustic SLAM

When thinking about other potential phenomena for SLAM approaches, acoustic would be a good option. Speakers are cheap and normally a lot of electronic devices have one. Also, acoustic approaches could be used underwater in contrast to camera-based SLAM or the approaches described in section 2.3 and 2.4.

What nearly all acoustic-based approaches have in common, that they have a prepared environment. Evers et al. in their paper about A-SLAM [8] is a good example. He and his team use speakers placed in a room emitting a specific signal and an actor equipped with a microphone array to calculate the incoming directions of the speaker signals. Each speaker acts as a landmark which position has to be estimated so it shouldn't move over time. Although it is a landmark-based approach, it shows what performance an acoustic-based SLAM could have.
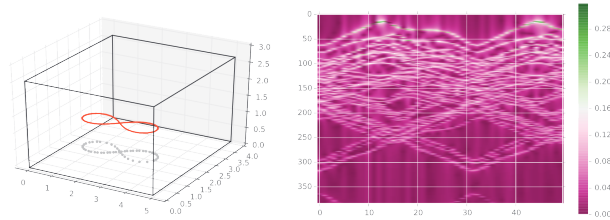

Fig. 8: RIR illustration of Dokmanic et al. [9]. Left Image shows the trajectory in red of the robot. At the right a plot of the RIR where x-axis is the time and y-axis the RIR.

There are also acoustic SLAM approaches based on continuous functions. One is from Dokmanic et al. who use the concept of room impulse response (RIR) from room reconstruction for a SLAM approach [9]. Therefore he placed a speaker in the room which is sending a specific signal. An actor can now move around and calculate the RIR based on the current position. This RIR is used for SLAM. One thing to point out that in this paper the RIR is a smooth continuous function when the robot is moving and when the robot passes the same point the RIR will be similar (see figure 8).

3   Pre-Studies

Based on the work of Dokmanic et al. [9] the question is now if the RIR will also be suitable for SLAM when the speaker is also mounted on the robot. In the case of Dokmanic et al. [9] described in section 2.5, the RIR is a continuous function that won't change until the speaker moves or the room will change. In this project, the speaker will move along with the robot. So the RIR as a 2D function over the room will change every time. In the pre-studies, it will be investigated if the RIR is suitable for SLAM when measuring the RIR at the position of a moving speaker.

But first a short recap of the room impulse response (RIR). The principle of sending a signal in a room and recording it with a microphone could be treated as an electrical system where the sent signal is the input and the recorded signal the output of the electrical system. In the concept of electrical systems, the output $y(t)$ depends on the input $x(t)$ and a function $g(t)$ which is convolved with the input in time space. So it follows

$$y(t) = (g * x)(t).$$

When now stimulating the system with an infinite small function also called impulse, the function $g(t)$ is called the impulse response. Normally the impulse response isn't calculated in time space because of the convolution operator. To calculate the impulse response the input and output is transformed in frequency space with Fourier transformation. In frequency space, the convolution becomes a multiplication and the impulse response can be calculated by dividing the output $y(\omega)$ by the input $x(\omega)$. Which is the same like

$$y(\omega) = g(\omega)x(\omega) \Rightarrow g(\omega) = \frac{y(\omega)}{x(\omega)}.$$

Obviously, the input signal can't be infinitely small. Also, because of the time-frequency uncertainty relation, an impulse has to have a certain length if a wide frequency band wants to be stimulated. For the pre-studies, we decided to use a chirp signal which is a sinusoidal function with a rise in frequency. Before calculating the impulse response, we decided to window the output with a Tukey window. The windows are overlapping slightly so that the area under the overlap is one. Each window is then Fourier transformed and the mean in

frequency space is calculated. Figure 9 shows an example of a chirp signal, the recorded equivalent and the impulse response in frequency space.

When looking at different phase plots of the RIR at the same position, it can be seen that the phase changes randomly. Because with the platform that is used to play the sound back, it is not possible to guarantee that the recording and sound playing starting at the same time. So there is always a random shift in time which depends on the CPU. If later needed, it may possible to compensate this shift by using cross-correlation to find the starting point of the chirp signal. But the error has to be significantly lower than 250 µs when using frequencies up to 4000 Hz which could be difficult because of the noisy recordings.
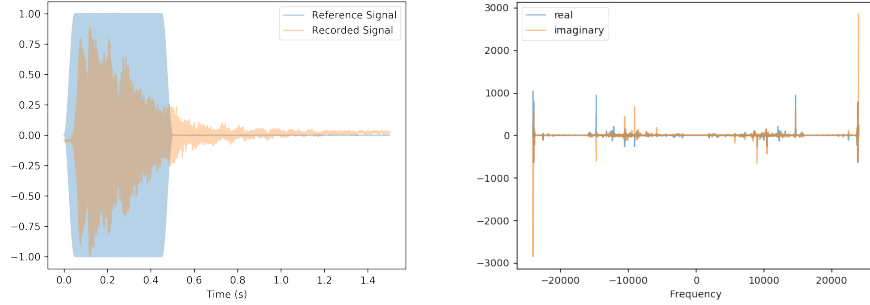

Fig. 9: On the left in blue, the chirp signal with a frequency starting at 200 Hz and going up to 4000 Hz linearly. The chirp signal is 0.5 seconds long and was multiplied with a Tukey window to prevent clicking sounds. In orange the recorded signal and on the right the RIR.

But it seems that the phase isn't necessary for this use case. When looking at the peaks of the magnitude, RIR at the same position will look similar as illustrated in figure 10. But that's expectable when looking at the work of Dokmanic [9]. Also expectable is that the peaks in magnitude are at a different position when changing the room which can be seen in figure 11. More interesting is the RIR of the same room but different positions. It turns out that the RIR is recognizably different in the same room but in different positions which can be seen in figure 12. Also interesting, peaks in the same room but different positions seem to be at similar positions when moving around (like in figure 8). But when changing the room the peak characteristic will change. That probably leads to a continuous function which is usable for SLAM purposes.
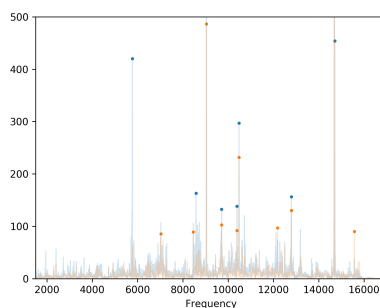
Fig. 10: Two RIR of the position but different times. Notice, most of the peaks are at the same position.
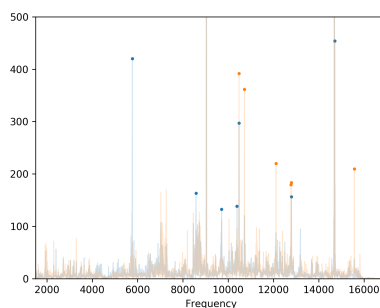

Fig. 11: Two RIR of different rooms. Notice, the peak characteristic has changed.
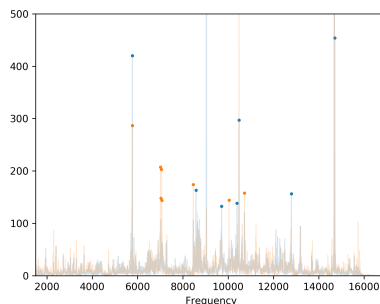

Fig. 12: Two RIR of the same room but different positions. Notice, the change in peaks is recognizable but some peaks are only shifted slightly.

4   Project Aims

The overall project goal is to implement a SLAM approach based on the RIR with a mobile platform equipped with an IMU, speaker, and microphone. To reach that, intermediate targets have to be defined. At first, the aim is to show that the RIR is a continuous function also for moving speakers not only for stationary ones which is already shown by Dokmanic [9] as described in section 2.5. For that, data of a moving agent has to be recorded which also can be a human holding the equipment and moving around for first tests.

After the knowledge that the RIR is continuous for the moving case, a function suitable for approximation with Gaussian Processes (GP) and SLAM has to be derived. The reason why not using the RIR itself is that, depending on the window size, GP would have a bad performance. As shown in section 2.5 the peaks are most interesting. That leads to the attempt using the position of the top ten peaks which would increase the performance a lot. But also other approaches have to be examined. For example, using the top lowest magnitudes when thinking of standing and extinguishing waves.

When a function suitable for real-time SLAM is found the SLAM approach has to be implemented, tested, and validated. For implementation, other approaches for continuous SLAM described in chapter 1 can be used for orientation. For testing purposes, the performance of the implemented SLAM approach should be examined. Where the performance could be the error or real-time requirements. When it comes to valuation, the approach could be compared to other approaches that are described in chapter 1. Especially to the one from Dokmanic [9] where the project is largely based on.

5   Work plan

To reach the goals of this project, the project is divided into six phases. Each phase leads to a milestone at the end of the phase and has also a time buffer of one week. An overview of the work plan can be seen as a Gantt chart in figure 13. Each phase is now explained in detail in the following sections.

5.1   Data Acquisition

For the first phase, a workload of six weeks is scheduled. It is also divided into three subphases. Firstly, a phase to understand and analyze the equipment for data recording. In the pre-studies, the microphone and speaker were already used but for the SLAM approach we need sound recordings in combination with position data. For that, several options are given by the Institut which have to be compared and understood. For example the HTC Vive for measuring position data in combination with the robot platform.

During this phase, the next phase can start where data is acquired for later analysis. This data could also have position data but that's not necessary. This data is used to compare different approaches for continuous functions based on

the RIR. It also should show later if the RIR with moving speaker will behave similarly to the approach from Dokmanic [9] as described in section 2.5.

After the analysis data is collected and all equipment is understood, the test and validation data can be gathered. This data should be used later to test and validate the implemented SLAM approach. So it should be a combination of sent output, recorded input, and position data for every timestep t.

The milestone of this phase will be to gather analysis, test, and validation data. If this phase will unexpectedly take much longer so the whole project will be in danger, at least the RIR approach explained in chapter 3 can be applied to the data to show if the RIR for moving speakers behave continuously.

5.2   Data Analysis

After the data acquisition phase, the data has to be analyzed in this phase to derive a continuous function suitable for SLAM and approximable with Gaussian Processes in real-time. For this, a workload of seven weeks is scheduled. This phase is also divided into two phases. The first one to analyze the data to find suitable continuous functions and the second phase to compare several approaches for continuous functions.

The milestone in this phase will be to find out the most promising continuous functions derived from the acoustic signal usable for SLAM approaches. If the project needed too much time until this phase, several approaches for continuous functions suitable for SLAM could be published which shows that acoustic SLAM is probably suitable for completely unprepared environments.

5.3   Gaussian Process Optimization

This phase will be a short phase with a workload of one week. In this phase, the Gaussian processes will be optimized for the different continuous functions given by the phase before. So the milestone will be a set of optimal Gaussian processes for the given functions. If no time after this phase is left for further work, a paper of the functions suitable for SLAM with some plots of the continuous function in space could be published.

5.4   Implementing Test Environment and SLAM Approach

A workload of five weeks is scheduled for this phase. This phase is also divided into three phases where one phase is implementing a test environment to simulate a virtual robot with the data given by the phase described in 5.2. This test environment is also used for later validation purposes. The next phase is to plot a map of the continuous functions to finally see if it's suitable for SLAM and if there are any errors or other issues within the data or with the derived functions. In the end, a SLAM approach will be implemented which is also the milestone of this phase. As in the phase before, if no time is left the derived functions and plots showing the continuous functions suitable for SLAM can be published.

5.5   Test and Validation

The test and validation phase is scheduled with a workload of three weeks and is divided into three subphases. First testing the SLAM approach given by the phase before in the test environment. After that, the SLAM approach should be tested on real data with a robot platform. At the end, the results should be compared to other approaches especially to the one from Dokmanic [9] described in section 2.5. The milestone of this phase will be the validation of this approach and the comparison to others. If the time is getting short, the test on real data could be canceled and the SLAM approach will only be tested in the virtual environment on the data given by the first phase described in 5.2.

5.6   Paper Writing

In the end, all knowledge of the project has to be gathered and written down into a paper. For that, a workload of three weeks is scheduled. It seems a little bit short but every phase should produce a short document with the milestone. Also, each phase has a buffer of one week. So if every phase before will work as planned, there will be six weeks buffer left for this phase. The milestone of this phase will be the paper.
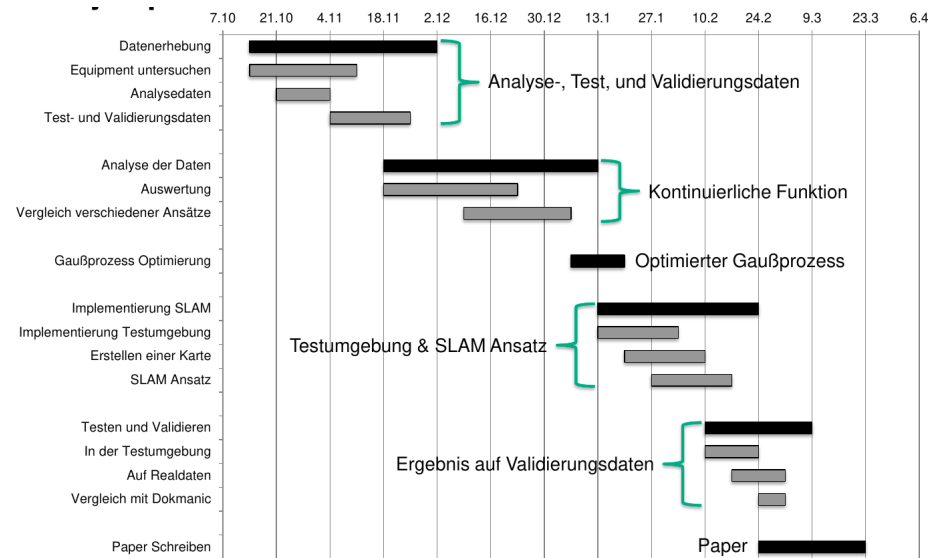

Fig. 13: This is the Gantt chart to illustrate the project plan.
TODO CHANGE TO ENGLISH

References

1. G. Grisetti, R. Kummerle, C. Stachniss, and W. Burgard, "A tutorial on graph-based SLAM," vol. 2, no. 4, pp. 31–43.
2. R. E. Kalman and Others, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.
3. M. Kaess, A. Ranganathan, and F. Dellaert, "iSAM: Incremental smoothing and mapping," vol. 24, no. 6, pp. 1365–1378.
4. Li Liang-qun, Ji Hong-bing, and Luo Jun-hui, "The iterated extended kalman particle filter," in *IEEE International Symposium on Communications and Information Technology, 2005. ISCIT 2005.*, vol. 2, pp. 1213–1216, Oct 2005.
5. M. Ebden, "Gaussian processes for regression: A quick introduction," p. 11, 2015.
6. M. Kok and A. Solin, "Scalable magnetic field SLAM in 3d using gaussian process maps,"
7. H. Yu and B. Lee, "Terrain field SLAM and uncertainty mapping using gaussian process," p. 4, 2018.
8. C. Evers, A. H. Moore, and P. A. Naylor, "Acoustic simultaneous localization and mapping (a-slam) of a moving microphone array and its surrounding speakers," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6–10, March 2016.
9. I. Dokmanić, L. Daudet, and M. Vetterli, "From acoustic room reconstruction to slam," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6345–6349, March 2016.
10. I. Wieser, A. V. Ruiz, M. Frassl, M. Angermann, J. Mueller, and M. Lichtenstern, "Autonomous robotic slam-based indoor navigation for high resolution sampling with complete coverage," in *2014 IEEE/ION Position, Location and Navigation Symposium - PLANS 2014*, pp. 945–951, May 2014.
11. D. Hanley, A. B. Faustino, S. D. Zelman, D. A. Degenhardt, and T. Bretl, "Mag-PIE: A dataset for indoor positioning with magnetic anomalies," in *2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pp. 1–8, IEEE.
12. M. F. Huber, "Recursive gaussian process: On-line regression and learning," vol. 45, pp. 85–91.
13. M. F. Huber, "Recursive gaussian process regression," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3362–3366, IEEE.
14. A. C. Damianou, M. K. Titsias, and N. D. Lawrence, "Variational inference for uncertainty on the inputs of gaussian process models,"
15. C. Evers and P. A. Naylor, "Acoustic slam," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1484–1498, Sep. 2018.