Combining Steered Response Power with 3D LIDAR scans for building sound maps

Jani Even, Yoichi Morales, Jonas Furrer, Carlos Toshinori Ishi, Norihiro Hagita

***Abstract*— This paper presents a framework for building 3D map of sounds. The environment is scanned by using a mobile platform equipped with a microphone array and a 3D LIDAR. A steered response power algorithm gives an angular distribution of the sound power at the mobile platform's position. This angular distribution is combined with the distances estimated by the 3D LIDAR in order to generate the spatial distribution of the sound's power. The fusion of the successive measurements obtained while the platform explores the environment results in the creation of the 3D sound map.**

I. INTRODUCTION

In acoustical signal processing, knowing the locations from which sounds are emitted is a very important task referred to as sound source localization (see [5]). Steered response power (SRP) algorithms are among the most effective methods that have been proposed [5], [4]. In particular, the SRP with PHase Transform (SRP-PHAT) [4] is well suited for robotic applications [1].

When the microphone array used for acquiring the audio data fed to the SRP algorithm is mounted on a mobile robot, the operational range of sound localization is extended as the robot can explore the environment. A natural framework for using the robot's mobility for sound source localization is to use a conventional sound source localization algorithm at different locations and combine the results from all these different locations [14], [8], [10], [11], [7].

In this paper, we present a framework for building a 3D map of sound using an autonomous mobile robot equipped with a microphone array and a 3D LIDAR (see Fig.1). The resulting 3D maps, referred to as *sound maps* in the remainder, can be exploited for sound source localization. The proposed method is a multi-modal approach that combines the bearing and power estimates from the SRP algorithm with the range estimates given by the 3D LIDAR. The 3D map is a 3D grid of voxels (3D cubes) that fill the space. Each of the voxel contains the information about the presence of an object at its location but it also contains the probability that this object emits sound. In order to build a precise map by fusing audio and LIDAR data, the platform has to localize itself in the environment. It is possible to proceed to local maxima search on the 3D grid in order to find the locations of the sound sources in the environment.

This research was funded by the Ministry of Internal Affairs and Communications of Japan under the Strategic Information and Communications R&D Promotion Programme (SCOPE).

The authors are with ATR Intelligent Robotics and Communication Laboratories, Kyoto, Japan. even at atr.jp

II. MAP BUILDING

To precisely localize itself in the environment, the mobile robot requires a map describing the environment, referred to as the *geometric map*.

The geometric map is built in advance using the 3D Toolkit library framework [3], [12]. To build the geometric maps, a mobile platform is driven through the environment. During this drive, the wheel encoders provide odometry data and the 3D Lidar provides scan data. Then the scans are aligned by correcting the trajectory of the platform using iterative closest point based simultaneous location and mapping (SLAM) [2]. Rather than using the aligned point cloud, an octree representation of the environment is created [6]. The geometric map refers to the voxels at the lowest level that are occupied (the edge length of the voxels composing the geometric map is 0.05 m). In Fig.2, a view of an indoors environment and the corresponding view in the associated geometric map are juxtaposed. Note that the voxels that compose the octree are clearly visible.

A. ROBOT LOCALIZATION

In this paper, it is assumed that the ground is flat and that the platform's pitch and roll are negligible. Consequently, the pose of the platform is composed of its 2D location $\{x_r(t), y_r(t)\}$ and its orientation $\theta_r(t)$. The altitude is assumed constant $z_r(t) = z_0$ and the pitch and roll null $\{\phi_r(t) = 0, \gamma_r(t) = 0\}$.

Since the localization is reduced to a 2D problem, laser range finders (LRFs) scanning in the horizontal plane at a height h_{LRF} are used to localize the mobile platform in a 2D map. The 2D map is created by taking an horizontal slice of the geometric map at the height $\{h_{\text{LRF}} - \epsilon, h_{\text{LRF}} + \epsilon\}$ and flattening it. Then the referential in the 2D map coincide perfectly with the one in the geometric map. Fig.3 gives the naming conventions for the pose $\{x_r(t), y_r(t), \theta_r(t)\}$ in the referential of the 2D map. The green arrows shows the

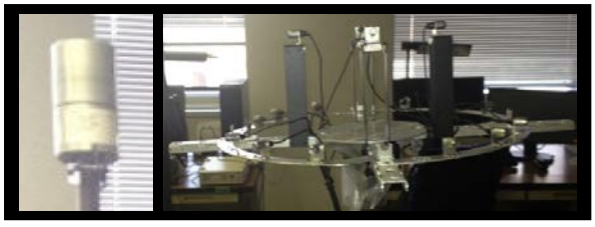

Fig. 1. 3D LIDAR (left) and microphone array (right).

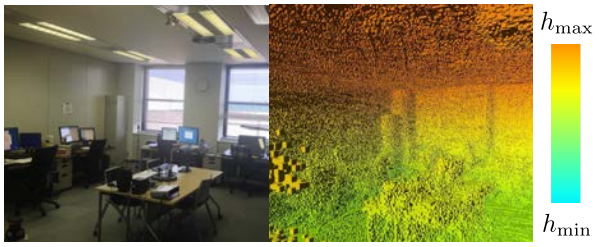

Fig. 2. Photo of the indoor environment and the corresponding view in the geometric map.

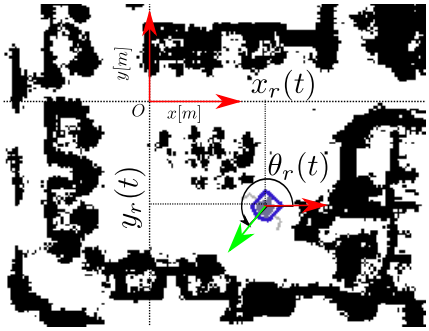

Fig. 3. Mobile platform localization in the $2D$ map created from the geometric map.

orientation of the mobile platform. This map correspond to the environment depicted in Fig.(2).

The localization algorithm is a particle filter (see [13] and references herein). In the prediction step, the particles are propagated accordingly to the odometry data. In the correction step, the likelihood of the particle is computed by using the ray tracing approach to match the LRFs scan to the 2D map. Resampling is performed when the number of effective particles is too low.

The number of particles is 200 and correction is performed when the platform moved by 0.1 m or rotated by 5 degrees.

III. STEERED RESPONSE POWER

In this paper, we use an SRP-PHAT algorithm to process the signals from the microphone array mounted on the mobile platform. The audio processing is done in the frequency domain. The frequency domain signals are denoted by $X_n(f,t)$ where n is the microphone index, f the frequency bin index and t the frame index. They are obtained by applying a short time Fourier transform (STFT) to the audio signals. The analysis window is W points long and the shift of the window is $W/2$.

First the PHAT transform is applied to the frequency components

$$V_n(f,t) \quad = \quad \frac{X_n(f,t)}{|X_n(f,t)|}. \qquad (1)$$

Then the power of the received sound is estimated for a set of candidate directions $\{\theta_i,\phi_i\}_{i\in[1,I]}$. The green dots in Fig.4 represent a set of candidate directions.

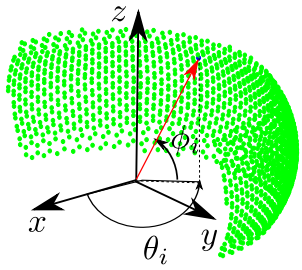

Fig. 4. Set of candidate directions (green dots) and conventions for the angles $\{\theta,\phi\}$ in the array referential.

For each of the candidate directions, the frequency domain processing is decomposed in 3 stages. First the response is steered in the candidate direction $\{\theta_i,\phi_i\}$ by applying a delay and sum spatial filter

$$Y(\theta_i,\phi_i,f,t) \quad = \quad H(\theta_i,\phi_i,f)\begin{bmatrix} V_1(f,t) \\ \vdots \\ V_N(f,t) \end{bmatrix}, \qquad (2)$$

with

$$H(\theta_i,\phi_i,f) = \frac{1}{N}\left[e^{-j\tau_1(\theta_i,\phi_i,f)},\cdots,e^{-j\tau_N(\theta_i,\phi_i,f)}\right], \quad (3)$$

where $\tau_n(\theta_i,\phi_i,f)$ is the phase delay at the microphone n in the frequency bin f for a signal coming from the direction $\{\theta_i,\phi_i\}$. Assuming that the sound sources are in the far field, the filter is entirely characterized by the angles $\{\theta_i,\phi_i\}$ and the microphone positions.

Then the power of the beamformer output is estimated by a K frame averaging

$$S(\theta_i,\phi_i,f,k) \quad = \quad \frac{1}{K}\sum_{t=0}^{K-1}|Y(\theta_i,\phi_i,f,k-t)|^2. \quad (4)$$

Note the introduction of the index k to show that the power has a different rate (the period is $KW/2$ samples).

Finally, the steered response power in the direction $\{\theta_i,\phi_i\}$ is obtained by selecting a limited band of frequencies

$$S(\theta_i,\phi_i,k) \quad = \quad \sum_{f=f_{\min}}^{f_{\max}} S(\theta_i,\phi_i,f,k). \qquad (5)$$

In the remainder, the term *audio scan* refers to the set of candidate directions $\{\theta_i,\phi_i\}_{i\in[1,I]}$ and their associated power $S(\theta_i,\phi_i,k)$ computed at a given frame k. The kth audio scan is denoted by $\mathcal{S}(k) = \{S(\theta_1,\phi_1,k),\cdots,S(\theta_I,\phi_I,k)\}$.

An audio scan is represented as a colored portion of a sphere in Fig.5 (left). The color is function of the power for each of the candidate directions. This audio scan clearly exhibits an area of higher power on the top left side indicating the presence of a sound source.

IV. AUDIO LIKELIHOOD

In order to fuse the sound source localization results of different audio scans together, the power $S(\theta_i,\phi_i,k)$ is

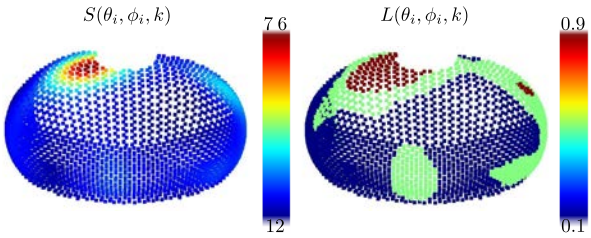

Fig. 5. Transformation by the thresholding function of the power (left) into a likelihood (right).

first transformed in a likelihood $L(\theta_i, \phi_i, k)$. This likelihood $L(\theta_i, \phi_i, k)$ expresses the belief of having a sound source in the candidate direction $\{\theta_i, \phi_i\}$.

From the sound source localization literature and the idea behind the SRP approach, it is expected that a large power should correspond to a strong belief. For example in [9], [7] a scale version of the power was used as likelihood. In audio source tracking, creating a likelihood by scaling the power is also common [15].

In this paper, rather than using a scaled power, a nonlinear function is applied in order to create the likelihood. The selected nonlinear function is a double thresholding function

$$F(x) = \begin{cases} p_{\text{min}}, & \text{if } x < T_1 \\ p_{\text{max}}, & \text{if } x > T_2 \\ p_{\text{med}}, & \text{else} \end{cases} \quad (6)$$

Fig.5 shows the transformation of an audio scan with the nonlinear function (the parameters are set to $T_1 = 20$, $T_2 = 30$, $p_{\text{min}} = 0.1$, $p_{\text{med}} = 0.5$, $p_{\text{max}} = 0.9$).

Consequently while the mobile platform is navigating the environment for each audio scan $\mathcal{S}(k) = \{S(\theta_1, \phi_1, k), \cdots, S(\theta_I, \phi_I, k)\}$ a *likelihood scan* $L(k) = \{L(\theta_1, \phi_1, k), \cdots, L(\theta_I, \phi_I, k)\}$ is create by applying the nonlinear function. Each of these likelihood scans contains the likelihood of having a sound source for one of the candidate directions $\{\theta_i, \phi_i\}$.

V. AUDIO MAP BUILDING

To understand the creation of the sound map, let us first discuss about the structure used to store the audio information. The sound map is an octree representation [6]. At the finest level of decomposition, the edge length of the voxels is 0.05 m and the voxels centered at the position $\{x, y, z\}$ is denoted by c_{xyz}. The voxels of the sound map have some fields to store the audio information:

- $\mathcal{L}(c_{xyz})$ denotes the log-odds of having a sound source within the voxel c_{xyz},
- $\mathcal{M}(c_{xyz})$ counts the number of times the voxel c_{xyz} was updated during sound map creation,
- $\mathcal{U}(c_{xyz})$ contains the last time the voxel c_{xyz} was updated.

Initially, all the voxels at the lowest level in the sound map are considered not occupied.

The candidate directions $\{\theta_i, \phi_i\}$ are defined in the referential centered at the microphone array depicted in Fig.4. This referential is rigidly attached to the mobile platform. The axis directions coincide with the platform's ones but the array origin is at the position $O_a = (x_a, y_a, z_a)_r$ (the subscript r denotes coordinate in the robot's frame).

A scan of the 3D LIDAR is composed of Q points $M_j = (x_j, y_j, z_j)_r$ in the robot's frame. The range in the direction $\{\theta_i, \phi_i\}$ is obtained by finding the point M_j the closest to that direction. For this purpose, let us define the audio direction

$$\vec{v}_a(i) = \begin{bmatrix} \cos(\theta_i)\cos(\phi_i) \\ \sin(\theta_i)\cos(\phi_i) \\ \sin(\phi_i) \end{bmatrix},$$

and the LIDAR direction

$$\vec{v}_L(j) = \frac{\overrightarrow{M_j O_a}}{|\overrightarrow{M_j O_a}|}.$$

Then the index j_i of the closest point M_j to the direction i is selected by finding

$$j_i = \text{argmin}_j \; 1 - \vec{v}_L(j) \cdot \vec{v}_a(i).$$

The point is considered valid if $1 - \vec{v}_L(j) \cdot \vec{v}_a(i) < \epsilon$, where ϵ is a small threshold, and the range associated to $\{\theta_i, \phi_i\}$ is $\rho_i = |\overrightarrow{M_j O_a}|$.

Then to relate the likelihood $L(\theta_i, \phi_i, k)$ to a geometric structure in the environment, the candidate direction has to be combined with the estimated pose of the platform. For the likelihood scan $L(k)$, the pose $\{x_r(t), y_r(t), \theta_r(t)\}$ of the platform with t the closest to k is considered.

This combination is illustrated in Fig.6. For simplicity, a top view is presented and the elevation angle ϕ_i is omitted. The circles represents the points M_j of the LIDAR scan.

For each of the candidate direction, a ray is casted from the array origin O_a in the referential of the sound map. Namely a ray of length ρ_i is casted from the point $(x_a(t), y_a(t), z_a(t))_w$ in the direction $\{\theta_r(t) + \theta_i, \phi_i\}$. Note that the coordinate of the array origin is a function of t in the world frame (denoted by subscript w) as the robot moves. Thus the end point falls in a voxel c_{xyz} of the sound map. Then the likelihood $L(\theta_i, \phi_i, k)$ is used to update the log-odds of having a sound source in this voxel.

The rationale behind the use of ray casting is to trace back the sound until its sources as in [7]. Contrary to [7], in this paper, the range of the sound source is given by the 3D LIDAR and not estimated from the position in the geometric map.

In practice, the ray casting is limited to a maximum range R_{max} as sound intensity decreases rapidly with the distance.

The audio related fields of the voxel are updated as follows

$$\begin{aligned} \mathcal{L}(c_{xyz}) &= \mathcal{L}(c_{xyz}) + \log \frac{L(\theta_i, \phi_i, k)}{1 - L(\theta_i, \phi_i, k)} \\ \mathcal{M}(c_{xyz}) &= \mathcal{M}(c_{xyz}) + 1 \\ \mathcal{U}(c_{xyz}) &= t_k, \end{aligned}$$

where t_k is the time corresponding to the frame k. At initialization $\mathcal{L}(c_{xyz}) = 0$, $\mathcal{M}(c_{xyz}) = 0$ and $\mathcal{U}(c_{xyz})$ is undetermined. The choice $\mathcal{L}(c_{xyz}) = 0$ means that a voxel has equal chance to emit or not sound.

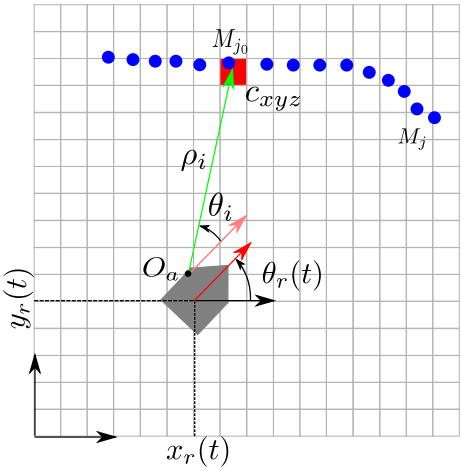

Fig. 6. Ray casting from the mobile platform pose $\{x_r(t), y_r(t), \theta_r(t)\}$ in the direction $\{\theta_i, \phi_i\}$ with a range ρ_i that falls in the voxel c_{xyz} (in red).

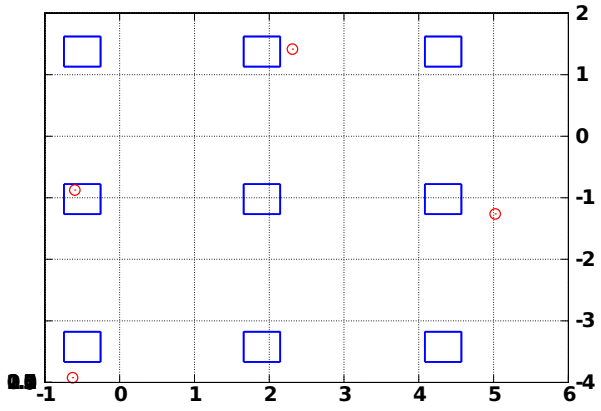

Fig. 7. Top view of the grills placement in blue and the estimated positions in red.

The log-odds $\mathcal{L}(c_{xyz})$ is no longer updated when it goes out of the interval $[\mathcal{L}_{\min}, \mathcal{L}_{\max}]$. Meaning that the odds of having a sound source at the voxel c_{xyz} is considered high or low enough to stop updating it.

The voxels having a count $\mathcal{M}(c_{xyz}) > \epsilon_C$ are considered occupied.

VI. EXPERIMENTAL RESULTS

This part reports the results of the experiments conducted to detect sound sources by using the sound map framework.

The experimental setting corresponds to the indoors environment depicted in Fig.2. At the time of the experiments, the sound sources in this environment are the grills of the air conditioning system. The grills are in the ceiling of the room and have a square shape (0.5 m edge). Fig.7 shows a top view of the room with the grills in blue.

To build the sound map, the mobile platform was driven three time around the table in the center of the room in a clockwise manner (see the 2D map in Fig.3). The parameters of the methods are set to $W = 400$, $K = 10$, $f_{\min} = 1000$ Hz, $f_{\min} = 3000$ Hz, $\mathcal{L}_{\min} = -40$ and $\mathcal{L}_{\max} = 40$. The

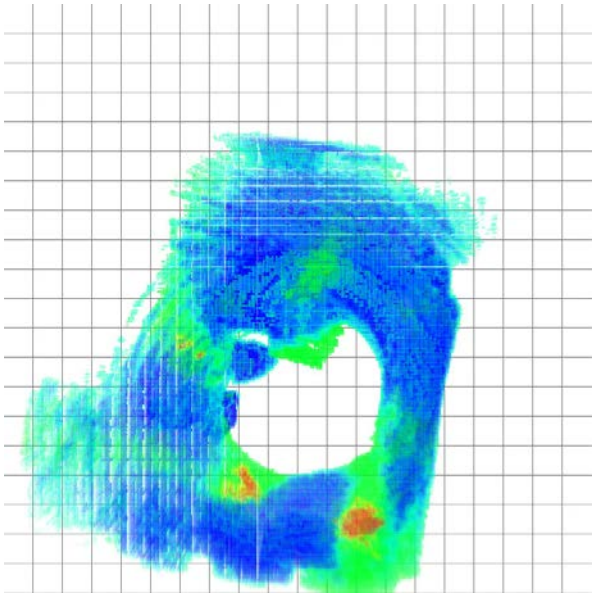

Fig. 8. Top view of the 3D sound map, the color represents the probability of sound source presence (ϵ_C=0).

parameters of the nonlinear function are set to $T_1 = 20$, $T_2 = 30$, $p_{\min} = 0.1$, $p_{\text{med}} = 0.5$, $p_{\max} = 0.999$. The angles of the candidate locations for the SRP algorithm are limited to $\theta \in [-45, 15]$ and $\phi \in [075]$. The maximum range is set to $R_{\max} = 6$ m.

Fig.**??** shows part of the sound map creation while the robot is moving. Fig.8 shows the top view of the sound map generated. Areas of high log-odds are visible around the location of the grills. The localization of the sound sources is estimated by clustering the voxels with positive log-odds (probability of having a sound source larger than 0.5). The clustering method is a kmeans method seeded with the positions of the local maxima of the log-odds. Each obtained cluster is assigned to the closest grill, then that grill is marked as detected and the distance to this grill is computed. In fig.7, the red circles indicate the positions of the detected sound sources (note that only a few sources are detected).

The sound source detection is evaluated in term of localization error E for these detected sources. The average error is 0.49 m with a standard deviation of 0.22 m. The errors are relative to the centers of the grills that have a 0.5 m edge. As a comparison, the maps presented in [8], [7] exhibit localization errors in the 0.2∼0.3 m range for the 2D case.

The undetected grills are the ones that were not for a long time in the aperture of the SRP while the platform made three loops around the table in the center of the room. The threshold ϵ_C for the count of the number of time a voxel was updated affects the number of occupied voxels. By plotting only the voxels c_{xyz} of the sound map such that $\mathcal{M}(c_{xyz}) > \epsilon_C$, it is possible to refine the map as illustrated in Fig. 10. The delimitation of the map and the areas of higher likelihood appear more clearly when the voxels with

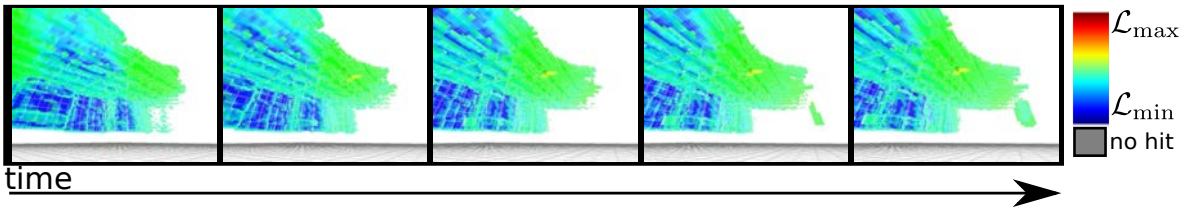

Fig. 9. Sound map creation.

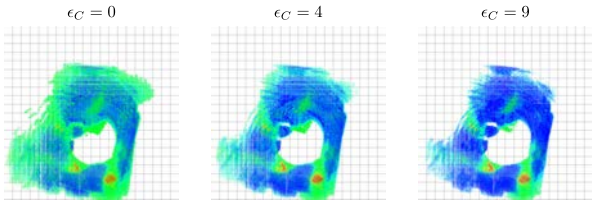

Fig. 10. Top view of the probabilistic 3D sound map for different count threshold ϵ_C.

few update are filtered out.

VII. CONCLUSIONS

This paper introduced a framework for creating a 3D description of the environment that contains the probability that a structure to emit sound. It is a multi-modal approach that combines 3D SRP with 3D LIDAR scans. Experimental results in an indoors environment showed that using the proposed approach it is possible to detect air conditioning grills and associate them with geometric features in the environment. The future work is to experiment in more diverse environments in order to determine the best parameter settings for different situations.

REFERENCES

[1] A. Badali, J.-M. Valin, F. Michaud, and P. Aarabi, "Evaluating real-time audio localization algorithms for artificial audition on mobile robots," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*, 2009, pp. 2033–2038.

[2] P. Besl and H. McKay, "A method for registration of 3-d shapes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[3] D. Borrmann, J. Elseberg, K. Lingemann, A. Nüchter, and J. Hertzberg, "The Efficient Extension of Globally Consistent Scan Matching to 6 DoF," in *Proceedings of the 4th International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT '08)*, Atlanta, USA, June 2008, pp. 29–36.

[4] M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *IEEE Conference on Acoustics, Speech, and Signal Processing, ICASSP 1997*, 1997, pp. 375–378.

[5] H. DiBiase, J. nad Silverman and M. Brandstein, *Microphone arrays : Signal Processing Techniques and Applications*. Springer-Verlag, 2007.

[6] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An efficient probabilistic 3D mapping framework based on octrees," *Autonomous Robots*, 2013, software available at http://octomap.github.com. [Online]. Available: http://octomap.github.com

[7] N. Kallakuri, J. Even, Y. Morales, C. Ishi, and N. Hagita, "Probabilistic approach for building auditory maps with a mobile microphone array," in *Proceedings of 2013 IEEE International Conference on Robotics and Automation, ICRA 2013*, 2013, pp. –.

[8] E. Martinson and A. C. Schultz, "Auditory evidence grids." in *IROS*. IEEE, 2006, pp. 1139–1144.

[9] ——, "Robotic discovery of the auditory scene," in *ICRA*, 2007, pp. 435–440.

[10] K. Nakadai, H. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "An open source software system for robot audition hark and its evalation," in *IEEE-RAS International Conference on Humanoid Robots*, 2008, pp. 561–566.

[11] Y. Sasaki, S. Thompson, M. Kaneyoshi, and S. Kagami, "Map-generation and identification of multiple sound sources from robot in motion," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2010*, 2010, pp. 437–443.

[12] slam6d, "Slam6d - simultaneous localization and mapping with 6 dof," Retrieved December May, 20 2011 from http://www.openslam.org/slam6d.html, 2011.

[13] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005.

[14] J.-M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, vol. 3, sept.-2 oct. 2004, pp. 2123 – 2128 vol.3.

[15] D. B. Ward, E. A. Lehmann, and R. C. Williamsin, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech and Audio Processing*, vol. 11, pp. 826–836, Nov. 2003.