A Framework for Spatial Map Generation using Acoustic Echoes for Robotic Platforms

Usama Saqib, Jesper Rindom Jensen

A Framework for Spatial Map Generation using Acoustic Echoes for Robotic Platforms

Usama Saqib, *Student Member, IEEE,* Jesper Rindom Jensen, *Member, IEEE,*

***Abstract*—In this work, we present a framework for constructing a spatial map of an indoor environment using the concept of echolocation. More specifically, we propose a non-linear least squares (NLS) estimator which is combined with a spatial filtering technique, e.g., beamforming, to estimate both the time-of-arrival (TOA) and direction-of-arrival (DOA) of the acoustic echoes. The proposed framework is complemented with an echo detector to classify a spurious estimate and an acoustic reflector, i.e., a wall. Based on these estimators, we then propose two algorithms that can complement existing range sensors and aid a robotic platform in acoustic reflector localization and mapping: a single-channel localization and mapping (ScLAM) and a multi-channel localization and mapping (McLAM). Compared to commonly used sensors, e.g., lidar and cameras, our proposed method can detect transparent surfaces that are typically found in an office environment. To test our algorithms, a proof-of-concept robotic platform was built. According to our evaluation, both qualitative and quantitative experiments reveal that the proposed methods can detect an acoustic reflector up to a distance of 1.5 m at a signal-to-diffuse-noise ratio (SDNR) of 0 dB in a simulated environment and 10 dB in a real environment with an accuracy of 80 %.**

***Index Terms*—robot audition, TOA estimation, DOA estimation, echolocation, localization, mapping**

I. Introduction

Robotic platforms, e.g., drones and unmanned ground vehicles (UGVs), has become an essential part of our society. We use them for tasks that are often monotonous and dangerous for human workers to handle. With the advancement of perception technology, i.e., the ability of a robot to perceive its environment, robots are now able to perform complicated tasks that makes them suitable in different sectors, e.g., agriculture [1], construction [2], supply chain and logistics [3], hospitals [4], etc. Within a warehouse setting, robots are often programmed to follow a predefined trajectory within an environment to transport goods. Over time, robots were equipped with proximity sensors, e.g., lidars, cameras, ultrasonic sensors, etc., for navigation, which also led the robots to plan its own path without human intervention, making these robots more autonomous. According to the IEEE Standard for Robot Map Data Representation for Navigation [5], one way to effectively navigate an indoor environment is to construct a spatial map of an environment, which is normally done using a very popular framework called Simultaneous Localization and Mapping (SLAM) [6]–[8]. The advantage of constructing a spatial map of a surrounding could also aid engineers and building planners to do asset maintenance and survey related work. Additionally, SLAM-based robots also aid rescue workers and surveyors to construct spatial maps of unknown environments, e.g. sewers [9], [10], underground tunnels, etc. Traditionally, lidar and camera-based technologies are used to provide input data to SLAM algorithms to construct a spatial map of an environment [11].

However, lidar and camera-based technologies are susceptible to changing light conditions or foggy environments and are also not suitable for detecting transparent surfaces [12]. This makes these technologies unsuitable to accurately generate a spatial map of a typical office environment [13]. Furthermore, lidar and camera-based technologies has limited field of view (FOV) and thus offers limited coverage when localizing targets around the corner of the room [14]. These issues can be resolved by employing sound [15]. Sound is used by animals (e.g., bats, dolphins, and rats) in nature for orienting themselves within an environment and hunting prey [16] by probing the environment. This process is known as echolocation. An advantage of using echolocation for spatial map generation is that it can enable a robotic platform to navigate an environment under poor lighting conditions. Furthermore, compared to camera and lidar-based technologies, microphones are typically cheaper and may offer omni-directionality. In the past, the concept of echolocation was studied by several researchers to built active SONAR (Sound navigation and ranging) technologies for naval submarines to detect incoming ships and hostile submarines [17].

The use of SONAR in air-borne applications is a challenging and complicated task but an attempt to study this was proposed in [18]. The authors utilizes two ultrasonic transmitters/receivers to effectively localize multiple targets up to a distance of 8 m and classification of the targets were done using template matching. Moreover, the authors in [19]–[21] has also proposed several techniques that utilizes sound to make a distance estimate of an acoustic reflector. However, these works assume that the time-of-arrival (TOA) information of the acoustic echoes are known prior to estimation. TOA measurement of an acoustic echo is usually extracted from the estimated room impulse response (RIR) using a standard peak-picking approach, but this has proven to be a non-trivial and time consuming process [22]. In acoustic signal processing, the RIR is the transfer function between the source and the microphone. It has a distinctive characteristic, i.e., it contains two main components: a direct-path plus early reflections and a stochastic long tail representing late reflections that contributes to the reverberation [23]. The direct-path component corresponds to the shortest distance that a sound travels to reach a receiver while the early reflections correspond to the sound bouncing off an acoustic reflector before reaching the

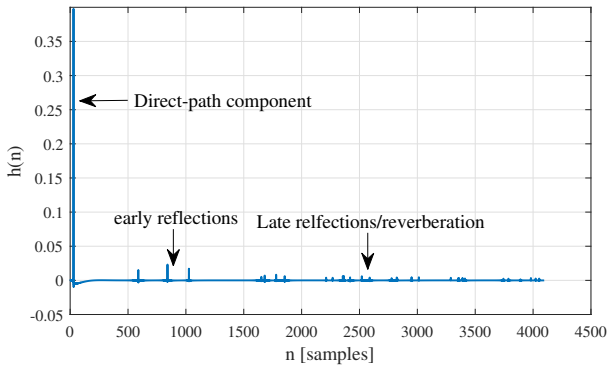


Fig. 1: An example illustrating synthetic Room Impulse Response (RIR)

receiver as shown in Fig. 1. Within the context of robotic platforms, the individual RIRs have to be estimated as the robot moves within an environment for TOA estimation of the early reflections as it corresponds to an information about the geometry of the room [24]. Furthermore, [25] proposed an algorithm called BatSLAM that utilizes a transmitter/receiver in the ultrasonic frequency range to generate a spatial map of an indoor environment. In [16], the researchers built a robotic platform that navigates an outdoor environment in order to construct a spatial map as well as classify flora using an artificial neural network. However, both these approach utilizes specialized sensors that operates in a specific frequency range of the sound, i.e., the ultrasonic range. Many robotic platforms that exist in the market are, on the other hand, intended for human-robot interaction (HRI), e.g., the NAO robot. These are often equipped with standard microphones and loudspeakers that operates in the audible frequency range. The estimation of TOAs and DOAs from an observed signal in the audible frequency range has been addressed previously in [26]–[29]. However, the presence of the direct-path component within the recording makes it difficult to detect and estimate early reflections as it contains the highest energy as seen in Fig. 1. Therefore, in [26], [30], the direct-path component was assumed to be absent (e.g., removed by preprocessing) from the synthetic data recordings before TOA/DOA estimation, but, in reality, this is a non-trivial problem on its own and could thus have a detrimental effect on the estimation. Additionally, the estimators proposed in [26], [30] always gives an estimate regardless of whether there actually is an acoustic reflector in the vicinity of the robot, which may lead to a noisy map generation. To mitigate this, we propose introducing a statistical echo detector within our framework [31]. The detector proposed in this paper is a binary classifier, where the statistics of the background noise is used to optimally define a threshold value against which the spurious estimates are differentiated from those corresponding to actual acoustic reflectors. Furthermore, in order to construct a spatial map of an indoor environment, the direction of arrival (DOA) of the acoustic echo is also required. This helps a robot estimate both the distance and the orientation of the acoustic reflectors, i.e., walls. Several techniques on DOA estimation exist in the literature [32]–[35] to estimate the required information, e.g., TOAs and DOAs, but without addressing the influence of the strong direct-path component of the sound source that could have a detrimental impact in TOA and DOA estimation. To address the presence of the direct-path component, we consider a setup consisting of a microphone array, e.g., a uniform circular array (UCA), and a loudspeaker situated at the center of the array as in Fig. 2. This setup could be placed on different kinds of robotic platforms, e.g., drones, UGVs, etc. Based on these conditions, we propose a framework that extends our existing approach [12], [30] to multi-channel. The framework consist of four main blocks: 1) a spatial filtering block, that utilizes an adaptive beamformer to filter the observed signals 2) a non-linear least squares (NLS) estimator proposed in [30] to estimate TOA from the filtered signal 3) an echo detector block, that takes the statistics of the background noise into account to optimally decide a threshold for deciding if the said estimates belong to actual acoustic reflectors and 4) a mapping block to utilize the DOA and TOA to generate a spatial map, as shown in Fig. 3. The process involves probing the environment with a known sound which is recorded by a UCA. The recorded audio data is then processed to estimate the DOAs and TOAs of the acoustic reflections. The advantage of the approach proposed in this paper is that it reduces the influence of the direct-path component resulting from the sound source, e.g., loudspeaker, which is achieved by first processing the acoustic echo with a spatial filter and later using the filtered signal for TOA estimation, and, then, classify whether an estimate belongs to an acoustic reflector or empty space. Finally, the estimates classified as belonging to acoustic reflectors, are used for generating a spatial map. Furthermore, the proposed methods are derived in the frequency-domain which provides a decrease in computational load. The proposed method estimate the parameter of interest directly from the observed signals and does not rely on estimating RIRs as the robot moves. Moreover, to test our proposed framework, we built a proof-of-concept robotic platform to generate real data for a multi-channel scenario similar to our earlier work where we introduced a single-channel estimator for acoustic reflector localization [12]. The dataset is made public[1].

The remainder of the paper is then structured as follows: Section II describe the signal model and problem formulation, Section III describes the non-linear least squares (NLS) estimator, Section IV describe the first multi-channel localization and mapping (McLAM) algorithm while Section V describes single-channel localization and mapping (ScLAM) algorithm. Additionally, the robotic architecture and components description are detailed in Section VI before proceeding to the Section VII. In Section VIII we test the performance of the proposed method on a robotic platform and then we discuss our findings and conclude the paper in Section IX and Section X. Moreover, in Section X, we propose different ways on how this research could be extended.

[1]The dataset used for simulation and evaluation can be found at http://homes.create.aau.dk//ussa/journal/index.php

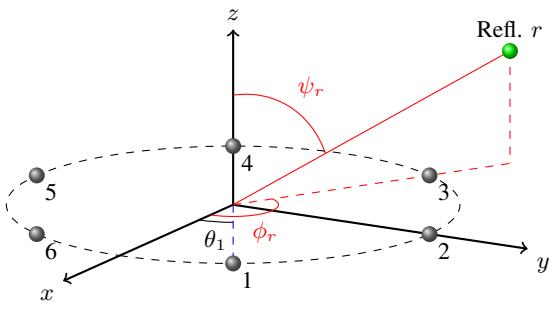


Fig. 2: Example of a uniform circular array with six microphones.

II. Signal Model and Problem Statement

A. Time-domain model

Consider an array with M microphones recording a sound emitted from a loudspeaker, including its acoustic reflections from walls, etc. The microphones and loudspeaker are collocated and the loudspeaker is assumed to be a point source. We can then formulate a general model for the recorded signal at microphone m, for $m = 1, \ldots, M$ at the kth robot position, as

$$\begin{aligned} y_{m,k}(n) &= (h_{m,k} * s)(n) + v_{m,k}(n), \\ &= x_{m,k}(n) + v_{m,k}(n) \end{aligned} \tag{1}$$

where $h_{m,k}(n)$ is the acoustic impulse response of the room measured from the loudspeaker to microphone m at robot position, w_k, for $k = 1 \ldots K$. Moreover, $v_{m,k}(n)$ is the additive background noise, including interfering sources plus the ego-noise of the robot at position, w_k. The operator $*$ represents the convolution operator, and $x_{m,k}(n) = (h_{m,k} * s)(n)$. In what follows, the background noise is assumed to be white Gaussian noise, but prewhitening techniques could be employed in cases where such assumptions are not met [36], [37]. By decomposing (1) as a sum of its direct-path component and its reflections and expressing the transfer function between the loudspeaker and a microphone in terms of its gain and delay, the signal model in (1) can be written as:

$$\begin{aligned} y_{m,k}(n) = &\sum_{r=1}^{\infty} g_{m,r,k} s(n - \tau_{ref,r,k} - \eta_{m,r,k}) \\ &+ v_{m,k}(n) \end{aligned} \tag{2}$$

where $g_{m,r,k}$ is the attenuation of the rth reflection from the loudspeaker to the microphone m at position, w_k, while $\tau_{ref,r,k}$ is the TOA of the reflected sound received at the reference point of the UCA at robot position, w_k, while $\eta_{m,r,k}$ is the time-difference-of-arrival (TDOA) between the reference point and the microphone m. In our definition in (2), the direct-path component corresponds to $r = 1$. The acoustic impulse response has a certain structure and is distinctively described in two parts: the direct-path plus early reflections and late reflections often described as a stochastic and dense tail. This means that we could rewrite (1) as the sum of the first R reflections to facilitate TOA and DOA estimation as shown

$$\begin{aligned} y_{m,k}(n) = &\sum_{r=1}^{R} g_{m,r,k} s(n - \tau_{ref,r,k} - \eta_{m,r,k}) \\ &+ d_{m,k}(n) + v_{m,k}(n), \end{aligned} \tag{3}$$

$$= x_{m,k}(n) + v'_{m,k}(n) \tag{4}$$

where $d_{m,k}(n)$ is the stochastic and dense tail of the late reflections. Often, we can combine the late reflections, $d_{m,k}(n)$, with the background noise as shown in (4) [38]. If we collect N time samples from each microphone and assume stationarity across those samples, we can vectorize our data and extend our signal model as shown:

$$\begin{aligned} \mathbf{y}_{m,k}(n) = &\sum_{r=1}^{R} g_{m,r,k} \mathbf{s}(n - \tau_{ref,r,k} - \eta_{m,r,k}) \\ &+ \mathbf{d}_{m,k}(n) + \mathbf{v}_{m,k}(n), \\ &= \mathbf{x}_{m,k}(n) + \mathbf{v}'_{m,k}(n) \end{aligned} \tag{5}$$

$$= [y_{m,k}(0) \quad y_{m,k}(1) \quad \cdots \quad y_{m,k}(N-1)]^T, \tag{6}$$

where the time-stacked probe signal, $\mathbf{s}(n)$, early reflections, $\mathbf{x}_{m,k}(n)$, and noise, $\mathbf{v}'_{m,k}(n)$, are defined similarly to $\mathbf{y}_{m,k}(n)$.

Hence, the signal formulation above yield an interesting problem to solve, namely, how to estimate $\tau_{ref,r,k}$ and $\eta_{m,r,k}$ of an acoustic reflector that will aid in simultaneously localizing and mapping an indoor environment. However, this requires us to estimate R unknown TOA and MR TDOAs from the observations $\mathbf{y}_{m,k}(n)$, at position, w_k. If we assume a known array configuration, however, we can reduce the dimensionality of this problem by incorporating the geometry of the loudspeaker and the microphone array.

B. Array Model

The array model can be chosen to be of any geometry but in this paper, we use a uniform circular array (UCA) with a loudspeaker placed at the center of the array. Although any reference point could be chosen to solve the TOA and DOA problems, we assume the center of the UCA to be the reference point. Assuming that the reflectors are in the far-field of the array and given the geometry of the microphones and the loudspeaker where the center of the microphone array is chosen as the reference point, we can then write the TDOAs of the acoustic echoes as follows:

$$\eta_{m,r}(\boldsymbol{\zeta}_r) = d \sin\psi_r \cos(\phi_r - \beta_m) \frac{f_s}{c}, \tag{7}$$

where $\boldsymbol{\zeta}_r = [\psi_r \ \ \phi_r]^T$, and ψ_r and ϕ_r are the elevation and azimuth angles, respectively, while d is the radius of the UCA. Furthermore, $\beta_m = \frac{2\pi i}{M} + \alpha$ is the angular position of the mth element on the UCA circle counted in an anti-clockwise manner from the x-axis and α is the offset angle. Moreover, f_s is the sampling frequency and c is the speed of sound. The TDOA model in (7) can be combined with the observation model in (5) to simplify the dimension of the estimation problem from MR to $2R$. The problem of interest is thus to estimate the unknown orientation parameters, i.e., ψ_r and ϕ_r, and the distance-related parameter, τ_r, based on

the posed array (7) and observation (5) models. Additionally, a classification of the estimates as either belonging to an actual acoustic reflector or an empty space is needed to generate a spatial map of the acoustic reflectors. Finally, the parameter estimates of the acoustic echoes need to be mapped into the acoustic reflector positions based on the robots movement and orientation.

III. Non-Linear Least Squares (NLS) Estimator

We can resolve the problem of estimating the unknown parameters in (5), i.e., $\tau_{ref,r,k}$ and $\eta_{m,r,k}$, by using a non-linear least squares (NLS) estimator, which is statistically optimal under the assumed white Gaussian noise conditions. Mathematically, this can be formulated as

$$\{\widehat{\mathbf{g}}_k, \widehat{\boldsymbol{\tau}}_k, \widehat{\boldsymbol{\zeta}}_k\} = \underset{\mathbf{g},\boldsymbol{\tau},\boldsymbol{\zeta},}{\arg\min} \sum_{m=1}^{M} \left\| \mathbf{y}_{m,k}(n) - \sum_{r=1}^{R} g_{m,r,k}\mathbf{s}(n - \tau_{ref,r,k} - \eta_{m,r,k}(\boldsymbol{\zeta}_r)) \right\|_2^2, \tag{8}$$

where

$$\boldsymbol{\tau} = \begin{bmatrix}\tau_1 & \tau_2 & \cdots & \tau_R\end{bmatrix}^T, \tag{9}$$

$$\mathbf{g} = \begin{bmatrix}\mathbf{g}_1^T & \mathbf{g}_2^T & \cdots & \mathbf{g}_R^T\end{bmatrix}^T, \tag{10}$$

$$\mathbf{g}_r = \begin{bmatrix}g_{1,r} & g_{2,r} & \cdots & g_{M,r}\end{bmatrix}^T, \tag{11}$$

$$\boldsymbol{\zeta} = \begin{bmatrix}\boldsymbol{\zeta}_1^T & \boldsymbol{\zeta}_2^T & \cdots & \boldsymbol{\zeta}_R^T\end{bmatrix}^T, \tag{12}$$

with $\widehat{x}$ denoting an estimate of x, and x_k denoting a parameter x related to the kth robot position. The displacement k of the robot can be estimated using an accelerometer or can be pre-programmed within the robot so that the robot follows a predefined trajectory. We can also solve (8) by converting it into frequency domain because 1) it will reduce the computational load when estimating the desired parameters [39], and 2) by working in the frequency domain, we will have the flexibility to work in specific frequency ranges or account for frequency dependency. For instance, if we want to work in the ultrasonic range then we can select and utilize only the frequency bins corresponding to these high frequencies. This may also help us design probe signals that are non-intrusive to human hearing but this is left for future iteration of this research. Using Parseval's theorem and omitting the frequency dependency in the notation, we can transfer (8) to the frequency domain, which yields the following:

$$\{\widehat{\mathbf{g}}_k, \widehat{\boldsymbol{\tau}}_k, \widehat{\boldsymbol{\zeta}}_k\} = \underset{\mathbf{g},\boldsymbol{\tau},\boldsymbol{\zeta}}{\arg\min}\, J(\mathbf{g}, \boldsymbol{\tau}, \boldsymbol{\zeta}), \tag{13}$$

where

$$J(\mathbf{g}, \boldsymbol{\tau}, \boldsymbol{\zeta}) = \sum_{m=1}^{M} \left\| \mathbf{Y}_{m,k} - \sum_{r=1}^{R} g_{m,r}\mathbf{Z}(\tau_r, \boldsymbol{\zeta}_r) \odot \mathbf{S} \right\|^2, \tag{14}$$

$$\mathbf{Y}_{m,k} = \begin{bmatrix}Y_{m,k}(0) & \cdots & Y_{m,k}(F-1)\end{bmatrix}^T, \tag{15}$$

$$\mathbf{Z}(\tau, \boldsymbol{\zeta}) = \begin{bmatrix}1 & e^{-j(\tau+\eta(\boldsymbol{\zeta}))2\pi\frac{1}{F}} & \cdots & e^{-j(\tau+\eta(\boldsymbol{\zeta}))2\pi\frac{F-1}{F}}\end{bmatrix}^T \tag{16}$$

with F denoting the number of frequency bins, $Y_{m,k}(f)$ denoting the DFT of $y_{m,k}(n)$ in frequency bin f. Moreover, $\mathbf{S}$ is the DFT vector of $\mathbf{s}(n)$ defined similarly to $\mathbf{Y}_{m,k}$. This estimation problem is multidimensional and thus computationally expensive in practice. To minimize the computational complexity, the multidimensional estimator could instead by implemented using various cyclic methods like the RELAX method proposed in [40] and later used in [30] to iteratively estimate the values of $\widehat{\boldsymbol{\tau}}_k$ and $\widehat{\mathbf{g}}_k$. In the special case where we are only concerned with estimating one acoustic reflection, and assuming that the direct-path component has been removed via preprocessing, we can set $R = 1$. Additionally, if we assume that the gain of each microphone is the same, then we can solve (13) for the gain $\widehat{g}_k$ by taking the derivative of the cost function, yielding:

$$\begin{aligned}\frac{\partial J(g_k, \tau_k)}{\partial g_k} &= \frac{\partial}{\partial g_k}(\mathbf{Y}^H\mathbf{Y} \\ &\quad - g_k\mathbf{Y}^H\overline{\mathbf{Z}}(\tau_k) - g_k\overline{\mathbf{Z}}^H(\tau_k)\mathbf{Y} \\ &\quad + g_k^2\overline{\mathbf{Z}}^H(\tau_k)\overline{\mathbf{Z}}(\tau_k)) \\ &= -\mathbf{Y}^H\overline{\mathbf{Z}}(\tau_k) - \overline{\mathbf{Z}}^H(\tau_k)\mathbf{Y} \\ &\quad + 2g_k\overline{\mathbf{Z}}^H(\tau_k)\overline{\mathbf{Z}}(\tau_k)\end{aligned} \tag{17}$$

where $\overline{\mathbf{Z}}(\tau_k) = \mathbf{Z}(\tau_k) \odot \mathbf{S}$ is the frequency domain probe signal delayed by τ_k samples. Solving for the linear gain parameter g_k gives:

$$\widehat{g}_k = \frac{\mathbf{Y}_k^H\overline{\mathbf{Z}}(\tau_k) + \overline{\mathbf{Z}}^H(\tau_k)\mathbf{Y}_k}{2\overline{\mathbf{Z}}^H(\tau_k)\overline{\mathbf{Z}}(\tau_k)}. \tag{18}$$

By inserting this back into (13), we get

$$\widehat{\tau}_k = \underset{\tau}{\arg\min} \left\| \mathbf{Y}_k - \frac{\mathbf{Y}_k^H\overline{\mathbf{Z}}(\tau) + \overline{\mathbf{Z}}^H(\tau)\mathbf{Y}_k}{2\overline{\mathbf{Z}}^H(\tau)\overline{\mathbf{Z}}(\tau)}\overline{\mathbf{Z}}(\tau) \right\|^2 \tag{19}$$

$$= \underset{\tau}{\arg\max}\, \mathbb{R}\{\mathbf{Y}_k^H\overline{\mathbf{Z}}(\tau)\} \tag{20}$$

where the operator $\mathbb{R}$ represents taking the real part of the signal. The expression in (20) estimates TOA for a single reflector at position, w_k. That is, for the special case with one acoustic echo, the NLS estimator in (20) can be interpreted as a cross-correlation based technique, which is widely used within robotics for source localization [41].

Based on the above problem formulation and methods in Section II and Section III, respectively, we propose two algorithm that could aid different robotic platform for spatial map construction: A multichannel localization and mapping (McLAM) algorithm and a singlechannel localization and mapping algorithm (ScLAM).

IV. Multi-channel Localization and Mapping (McLAM)

When constructing a spatial map of an environment using sound, a robotic platform requires both DOA and TOA information of the acoustic echoes while distinguishing estimates belonging to an acoustic reflector from spurious estimates. Furthermore, this should be carried out under the presence of a strong direct-path component originating from the sound

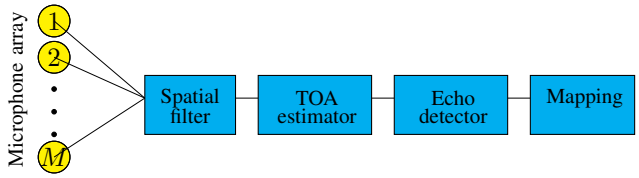


Fig. 3: Proposed McLAM Architecture

source, which detrimentally influence the estimation of the acoustic parameters. To address these problems, we propose mounting a microphone array on a robotic platform so that both the DOA and TOA of the acoustic echoes could be estimated, while suppressing the direct-path component. The multi-channel localization and mapping (McLAM) architecture has four important components as shown in Fig. 3. First, we introduce a spatial filter, i.e., a beamformer, to determine the DOAs of the acoustic echoes impinging from the reflectors, e.g., walls. Second, we feed the filtered observation into an NLS estimator to find the TOAs of the acoustic echoes. Then, we introduce a binary classifier to distinguish between spurious and real estimates, to exclude spurious estimates in the subsequent mapping of the acoustic reflectors, which constitutes the final block.

A. Spatial Filter block

The DOA information of an acoustic echo can, for example, be determined using the traditional spatial filtering techniques, e.g., beamforming [33], as considered in this paper. Later, a TOA estimate technique is applied, so that acoustic echoes corresponding to the distance of acoustic reflectors are estimated. Apart from DOA estimation, the other advantage of using spatial filtering before TOA estimation is that it can suppress the direct-path component that can affect the parameter estimation. Beamforming is based on the spatial weighting of the signals recorded by a microphone array such that the output signal is the weighted summation of all the signals to extract the signal impinging from a particular DOA [42]. In this way, we first employ a beamformer for estimating the angle of an acoustic echo using the steered response power approach. Subsequently, the echo is extracted by applying a beamformer steered towards the estimated angle to produce the output signal for the later TOA estimation using an NLS estimator (20) [30]. We therefore seek to estimate the signal of interest (SOI), while minimizing the influence of the direct-path component of the probe sound and other noise sources, e.g., from the rotors of a drone. With this aim, we consider the use of an adaptive beamformer [33]. In addition, this idea also builds on the statistical foundation of the EM method [26], which indicates that this is the optimal way of solving the problem of localizing acoustic reflectors in an indoor environment.

Due to the broadband nature of the signals involved, we implement the beamformer in the frequency domain. Therefore, the observations in (1) were first converted into frequency

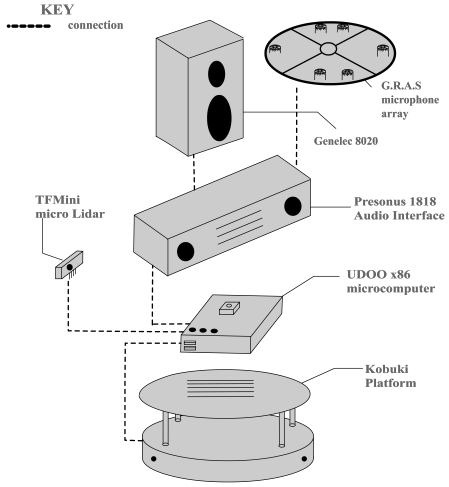


Fig. 4: An overview of components required to built a multi-channel robotic platform used for this research

domain as shown:

$$\begin{aligned}\mathbf{Y}_k &= \mathbf{X}_k + \mathbf{V}'_k \\ &= \mathbf{d}(\boldsymbol{\zeta}_{r,k})S_{r,k} + \mathbf{U}_k \\ &= \begin{bmatrix} Y_{1,k}(\omega) & Y_{2,k}(\omega) & \cdots & Y_{M,k}(\omega) \end{bmatrix}^T,\end{aligned} \tag{21}$$

where $\mathbf{X}_k$ and $\mathbf{V}'_k$ is defined similarly to $\mathbf{Y}_k$. Moreover, $\mathbf{U}_k$ contains the remaining $R-1$ early reflections as well as the late reverberation and background noise, and $S_{r,k}$ is the complex amplitude of the rth reflection at frequency ω. Assuming a UCA with the center of the array chosen as the reference point, the steering vector can be written as follows:

$$[\mathbf{d}(\boldsymbol{\zeta}_k)]_m = e^{-j\frac{\omega}{c}d\sin(\psi_k)\cos(\phi-\beta_m)}. \tag{22}$$

Here, $\boldsymbol{\zeta}_k$ is the look direction of the beamformer. The objective of the beamformer is then to recover the desired signal $S_{r,k}$ given the observation $\mathbf{Y}_k$, i.e.,

$$\overline{Y}_{\boldsymbol{\zeta}_k} = \mathbf{w}^H\mathbf{Y}_k, \tag{23}$$

where $\mathbf{w} \in \mathbb{C}^M$ and $\overline{Y}_{\boldsymbol{\zeta}_k}$ is the recovered signal from the observed signal from direction $\boldsymbol{\zeta}_k$ at position w_k, which should be an estimate of S_k. Here, several beamforming filters could be used, while, in this paper, we consider three types of beamformers which, e.g., facilitates a trade off between computational efficiency, estimation accuracy, and direct-path component suppression: 1) the minimum power distortionless response (MPDR) beamformer, 2) the delay-and-sum (DSB) beamformer, and 3) the linearly constrained minimum variance (LCMV) beamformer [43]. The MPDR beamformer is derived by minimizing the power of the of the output of the beamformer $\overline{Y}_{\boldsymbol{\zeta}_k}$ subject to a distortionless constraint, i.e.,

$$\begin{aligned}\mathbf{w}_{\text{MPDR}} &= \arg\min \mathbf{w}^H\mathbf{R}_{Y_k}\mathbf{w} \\ &\quad \text{subject to } \mathbf{w}^H\mathbf{d}(\boldsymbol{\zeta}_k) = 1.\end{aligned} \tag{24}$$

The solution to this is then well known to be given by [30]

$$\mathbf{w}_{\text{MPDR}} = \frac{\mathbf{R}_{Y_k}^{-1}\mathbf{d}(\boldsymbol{\zeta}_k)}{\mathbf{d}^H(\boldsymbol{\zeta}_k)\mathbf{R}_{Y_k}^{-1}\mathbf{d}(\boldsymbol{\zeta}_k)}, \tag{25}$$

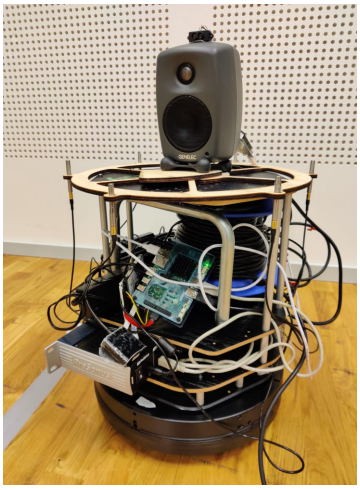


Fig. 5: The multi-channel proof of concept robotic platform.

where $\mathbf{R}_{Y_k} = E[\mathbf{Y}_k \mathbf{Y}_k^H]$ is the $M \times M$ covariance matrix of the observed signal, $E[\cdot]$ is the mathematical expectation operator, and $\mathbf{w}_{\text{MPDR}}$ is the complex weight vector corresponding to the MPDR beamformer. If the observed signal is assumed to be white Gaussian noise, e.g., $\mathbf{R}_{Y_k} = \mathbf{I}_M$, where $\mathbf{I}_M$ is the $M \times M$ identity matrix, the MPDR design resembles the DSB, i.e.,

$$\mathbf{w}_{\text{DSB}} = \frac{\mathbf{d}(\boldsymbol{\zeta}_k)}{M}. \tag{26}$$

Similarly, the LCMV beamformer is derived by extending the MPDR beamformer with additional constraints such that the optimization problem is solved as shown:

$$\begin{aligned} \mathbf{w}_{\text{LCMV}} = &\arg\min \mathbf{w}^H \mathbf{R}_{Y_k} \mathbf{w} \\ &\text{subject to } \mathbf{w}^H \mathbf{D} = \mathbf{f}^T. \end{aligned} \tag{27}$$

Here, $\mathbf{D}$ is a matrix containing all the steering vector for the C different constraints in $\mathbf{f} \in \mathbb{R}^L$. In this paper, we choose $\mathbf{f}$ as

$$\mathbf{f} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T \tag{28}$$

By utilizing this, we can reject the interference of the direct-path component by introducing a null in the direction of the loudspeaker, i.e., the center of the UCA. The solution to the LCMV beamforming problem is

$$\mathbf{w}_{\text{LCMV}} = \mathbf{R}_{Y_k}^{-1} \mathbf{D} [\mathbf{D}^H \mathbf{R}_{Y_k}^{-1} \mathbf{D}]^{-1} \mathbf{f}. \tag{29}$$

B. TOA estimator block

The output of these beamformers are subsequently feed to the non-linear least squares (NLS) Estimator for TOA estimation in (20). This estimator is statistically optimal when estimating τ and g for a single reflection while the background noise is white Gaussian. By preprocessing the observation with the adaptive beamformer, this assumption is better met since we can reduce the impact of directional and colored noise [44]. The resulting NLS estimator is then given by

$$\{\widehat{g}_k, \widehat{\tau}_k, \widehat{\boldsymbol{\zeta}}_k\} = \underset{g,\tau,\boldsymbol{\zeta}}{\arg\min} \left\| \overline{Y}_{\boldsymbol{\zeta}} - g\mathbf{Z}(\tau) \odot \mathbf{S}) \right\|^2, \tag{30}$$

where $\overline{Y}_{\boldsymbol{\zeta}}$ is the output of the beamformer (23) extracted from direction $\boldsymbol{\zeta}$ at a position w_k at frequency ω, while $\odot$ is the element-wise multiplication operator. By solving for the linear parameters in (30), we get the concentrated estimator for the TOA and DOAs:

$$\{\widehat{\tau}_k, \boldsymbol{\zeta}_k\} = \underset{\tau,\boldsymbol{\zeta}}{\arg\max}\, \mathbb{R} \left\{ \overline{Y}_{\boldsymbol{\zeta}}^H \overline{\mathbf{Z}}(\tau) \right\} \tag{31}$$

where $\overline{\mathbf{Z}}(\tau) = \mathbf{Z}(\tau) \odot \mathbf{S}$ and the operator $\mathbb{R}$ represents taking the real part of the signal.

C. Echo detector block

If the robotic platform is expected to move autonomously based on echolocation, a problem of great importance is to detect whether the observed signal received by the microphones represent an acoustic reflector, or if it only contains noise, e.g., the ego-noise of the robotic platform. This is because the TOA estimator in (31) provides estimates even when no acoustic reflector is present, which may lead to spurious localization estimates. To prevent false estimation when no acoustic reflector is present, several approaches could be applied including machine learning approaches [45], and deep learning [46] to categorize acoustic reflectors. Another approach could be to include a Generalized Likelihood Ratio Test (GLRT) detector [31] within our framework to distinguish whether the observed signal contains an acoustic reflection or not. Compared to the data-driven machine learning approaches, the GLRT is based on a priori model assumptions, and does thus not require training data. In this paper, we therefore employ the GLRT detection approach as discussed in the following.

If we assume the acoustic reflection to be in the far-field of the array, the decision about whether an observation contains an acoustic reflection can be formulated as a detection problem [31]:

$$\mathcal{H}_0 : \mathbf{y}_{m,k}(n) = \mathbf{v}'_{m,k}(n) \tag{32}$$

$$\mathcal{H}_1 : \mathbf{y}_{m,k}(n) = g_k \mathbf{s}(n - \tau_{m,k}) + \mathbf{v}'_{m,k}(n), \tag{33}$$

for $m = 1, \ldots, M$, where $\mathcal{H}_0$ is the null hypothesis referring to a situation when the observation only includes white Gaussian background noise and late reverberation, $v'_{m,k}(n)$, with variance σ^2, while $\mathcal{H}_1$ refers to the situation when the observation includes a reflected version of the known probe signal $s(n)$ in noise. Here, we assume that the direct-path component is absent, i.e., suppressed via preprocessing. The GLRT is then given by

$$\frac{p(\mathbf{y}_k; \widehat{g}_k, \mathcal{H}_1)}{p(\mathbf{y}_k; \mathcal{H}_0)} > \gamma, \tag{34}$$

$$\mathbf{y}_k = \begin{bmatrix} \mathbf{y}_{1,k}^T(0) & \cdots & \mathbf{y}_{M,k}^T(0) \end{bmatrix}^T, \tag{35}$$

It can then be shown that, in order to detect if the observation belongs to $\mathcal{H}_1$, we can use a threshold that depends on the power of the attenuated probe signal, the noise variance, and γ. If the power, $T(\mathbf{y}_k)$, of a matched filtering between the probe signal and the observed signal at the reference microphone exceeds this threshold, we decide $\mathcal{H}_1$, i.e., if

$$T(\mathbf{y}_k) = \mathbf{y}_k^H(n) \mathbf{H}(\boldsymbol{\tau}_k) \mathbf{s}(n) > \widehat{g}_k \frac{\epsilon}{2} + \frac{\sigma^2 \ln \gamma}{2\widehat{g}_k} \tag{36}$$

with

$$\mathbf{H}(\boldsymbol{\tau}_k) = \begin{bmatrix}\mathbf{D}^T_{\tau_{1,k}} & \cdots & \mathbf{D}^T_{\tau_{M,k}}\end{bmatrix}^T, \tag{37}$$

$$\epsilon = M\|\mathbf{s}(n)\|^2, \tag{38}$$

$$\widehat{g}_k = \frac{2\mathbf{y}_k^H(n)\mathbf{H}(\boldsymbol{\tau}_k)\mathbf{s}(n)}{M\|\mathbf{s}(n)\|^2}, \tag{39}$$

where $\mathbf{D}_\tau$ is a cyclic shift register that delays a signal by τ samples. The GLRT derivation is shown in Appendix **??**.

D. Mapping block

In this block, the DOA and TOA estimates are used alongside the robot's position within an environment to localize the position of an acoustic reflector. The aspect of the robot's navigation and path planning is beyond the scope of this paper, however, by utilizing common on-board sensors, e.g., Initial Mass Units (IMUs), of the robotic platform, we can estimate the robot's position. By combining this information with the estimates of the acoustic echoes obtained using, e.g., the methods considered in this paper, a spatial map of the environment can be generated for the robotic platform. The resulting spatial map may then enable the robotic platform to plan its path and move autonomously within the environment.

To estimate the position of the acoustic reflector from the estimated TOA, $\widehat{\tau}_k$, we assume that the sound propagates in plane waves (i.e., the source is in the far-field of the array). If we assume the speed of sound to be fixed then also assume estimation of a single acoustic reflector then the distance of the acoustic reflector with respect to the robotic platform is estimated as $\delta_k = \frac{c\cdot\tau_k}{2}$. Additionally, the direction of the acoustic reflector at position w_k is determined from the DOA estimates ψ and ϕ. If we assume a $2D$ scenario, where the reflections and the hardware is located in the same plane, we can utilize the far-field assumption and the choice of our reference point to conduct the mapping as:

$$\begin{aligned} p_{x_k} &= w_{x_k} + \delta_k \cos\phi_k \\ p_{y_k} &= w_{y_k} + \delta_k \sin\phi_k \end{aligned} \tag{40}$$

The procedure is then to estimate the acoustic reflector positions for each of the known robot positions, w_k, along its trajectory. The estimated acoustic reflector positions are then concatenated in the set $\mathcal{P} = \{p_1, \ldots, p_K\}$ with $p_k = (p_{x_k}, p_{y_k})$ for $k = 1, \ldots, K$.

The spatial filtering, the TOA estimator, the echo detector and the mapping block are then combined to form the basis of our proposed McLAM method. The algorithm describing the proposed McLAM method is outlined in Algorithm 1. However, in some applications, only one microphone and loudspeaker pair may be available for the mapping. In the following section, we therefore consider, how the hardware directivity properties may be exploited to localize the acoustic reflectors.

Algorithm 1: Proposed method McLAM.

Input : Trajectory $\mathcal{W} = \{(w_{x_1}, w_{y_1}), \ldots, (w_{x_K}, w_{y_K})\}$;
Output: Reflector position estimates $\mathcal{P} = \{(p_{x_1}, p_{y_1}), \ldots, (p_{x_K}, p_{y_K})\}$;
Initialization:
  $\mathcal{P} = \{\}$, $\mathbf{DOA} = \{\}$, $\mathbf{TOA}, = \{\}, \boldsymbol{\Phi} = [0°; 360°]$;
for $k = 1, \ldots, K$ **do**
  Probe the environment with $\mathbf{s}(n)$;
  Record echoes in $\mathbf{y}_k$;
  Transform signals to frequency domain $\mathbf{s}(n), \mathbf{y}_k(n) \xrightarrow{\text{FFT}} \mathbf{S}, \mathbf{Y}_k$;
  for $\phi \in \boldsymbol{\Phi}$ **do**
    Compute $\mathbf{w}$, e.g., using (24);
    $\overline{Y}_{\phi,k}(\omega) = \mathbf{w}^H\mathbf{Y}_k$;
  end
  $\{\widehat{\tau}_k, \widehat{\phi}_k\} = \arg\max_{\tau,\phi} \mathbb{R}\left\{\overline{Y}^H_{\phi,k}\overline{\mathbf{Z}}(\tau_k)\right\}$;
  $\widehat{\phi}_k \xrightarrow{\text{update}} \mathbf{DOA}$;
  $\widehat{\tau}_k \xrightarrow{\text{update}} \mathbf{TOA}$;
  Apply the echo detector in (36);
  if $\mathbf{y}_k^H(n)\mathbf{H}(\boldsymbol{\tau}_k)\mathbf{s}(n) > \widehat{g}_k\frac{\epsilon}{2} + \frac{\sigma^2\ln\gamma}{2\widehat{g}_k}$ **then**
    Compute p_k using (40);
    $p_k \xrightarrow{\text{update}} \mathcal{P}$;
  end
end

V. Single Channel Localization and Mapping (ScLAM)

In some applications, robotic platforms, e.g, intended for HRI may consist of only a single loudspeaker and microphone. In such a scenario, it is therefore necessary to reduce the McLAM algorithm to a single channel localization and mapping (ScLAM) algorithm. However, using such a single-channel approach has certain limitations, for instance, it cannot generally not be used to estimate the DOA of the acoustic echoes because of the lacking spatial information. Some possible ways of combating this are to, e.g., exploit the movement of the robot [19], or, as considered in this paper, to exploit the directionality of the employed hardware [12].

As in the McLAM, the loudspeaker probes the room with a known sound, $s(n)$, which is recorded by a microphone as the robot moves via positions w_k, for $k = 1, \ldots, K$. The NLS estimator described in (20) estimates τ_k for every robot position, w_k. Consider the platform moving in a predefined trajectory $\mathcal{W} = \{w_1, \ldots, w_K\}$ with $w_k = (w_{x_k}, w_{y_k})$, such that the platform moves from w_k to w_{k+1} etc. For every position, w_k, the platform will thus probe the environment with $\mathbf{s}(n)$ and record the observed signal $\mathbf{y}_k(n)$. The probed and observed signals are then converted into the frequency domain before passing them to the NLS estimator. In practice, the analysis window for the TOA could be restricted to a search interval from $\tau_{\min}$ up to $\tau_{\max}$ samples. This leads to

$$\widehat{\tau}_k = \underset{\tau\epsilon[\tau_{\min};\tau_{\max}]}{\arg\max}\ \mathbb{R}\{\mathbf{Y}_k^H\overline{\mathbf{Z}}(\tau)\} \tag{41}$$

In ScLAM, the position of the acoustic reflector is then inferred from the estimated TOA, $\widehat{\tau}_k$, by exploiting the typical directionality of a loudspeaker. More specifically, we assume the acoustic reflector to be located at the distance correspond-

Algorithm 2: Proposed method ScLAM.

input : Trajectory $\mathcal{W} = \{(w_{x_1}, w_{y_1}), \ldots, (w_{x_K}, w_{y_K})\}$, Initialization $\mathcal{P} = \{\}$, $\mathbf{TOA}, = \{\}$;
output : Reflector position estimates $\mathcal{P} = \{(p_{x_1}, p_{y_1}), \ldots, (p_{x_K}, p_{y_K})\}$;
for $k = 1, \ldots, w_k$ **do**
  Acquire direction of robot movement: $\theta_{r,k}$;
  Acquire direction of loudspeaker: $\theta_{l,k}$;
  Probe the environment with $\mathbf{s}(n)$;
  Record echo: $\mathbf{y}_k$;
  Transform signals to frequency domain $\mathbf{s}(n), \mathbf{y}_k(n) \xrightarrow{\text{FFT}} \mathbf{S}, \mathbf{Y}_k$;
  $\widehat{\tau}_k = \arg\max_{\tau_k} \mathbb{R}\{\mathbf{Y}_k^H \overline{\mathbf{Z}}(\tau)\}$;
  $\{\widehat{\tau}_k\} \xrightarrow{\text{update}} \mathbf{TOAs}$;
  Apply the echo detector in; (36);
  if $\mathbf{y}_k^H(n)\mathbf{H}(\boldsymbol{\tau}_k)\mathbf{s}(n) > \widehat{g}_k \frac{\epsilon}{2} + \frac{\sigma^2 \ln \gamma}{2\widehat{g}_k}$ **then**
    $\tau_k \xrightarrow{\text{remove}} \mathbf{TOAs}$;
    p_k using (43) $\xrightarrow{\text{update}} \mathcal{P}$;
  end
end

ing to the estimated τ_k in the direction of the loudspeaker. Additionally, the direction in which the robot platform is moving, $\theta_{\text{rob},k}$, at position w_k, is related to the direction that the loudspeaker is facing, θ_{l_k}, by a fixed offset angle, $\Delta\theta$, i.e.,

$$\theta_{l_k} = \theta_{\text{rob},k} + \Delta\theta. \tag{42}$$

Based on the above information, the coordinates of the position of the acoustic reflector is then estimated as follows:

$$\begin{aligned} p_{x_k} &= w_{x_k} + \delta_k \cos\theta_{l_k}, \\ p_{y_k} &= w_{y_k} + \delta_k \sin\theta_{l_k}. \end{aligned} \tag{43}$$

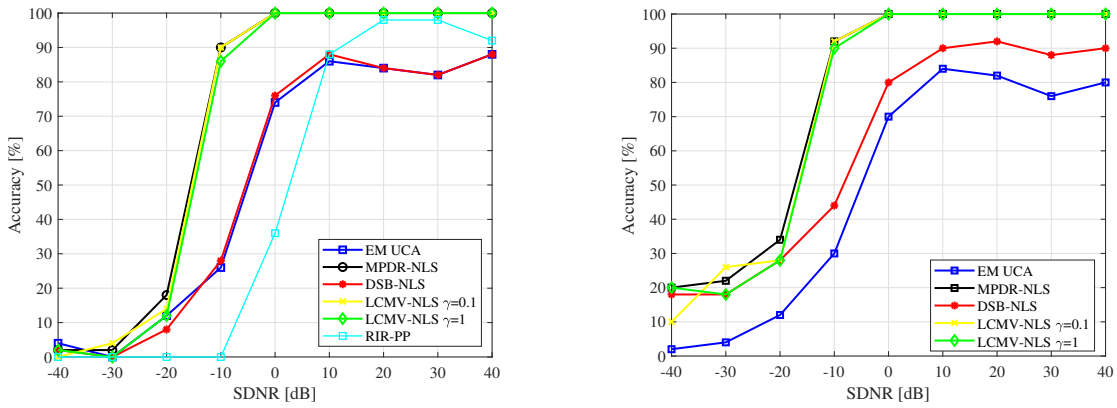
The resulting ScLAM algorithm is then proposed in Algorithm 2, which can be used used to construct a spatial map of a 2D environment with a single-channel loudspeaker/microphone setup.

VI. Robotic Platform Overview

The proposed methods discussed in Section IV and Section V were implemented on an embedded platform running a Windows 10 Operating System. The microcomputer used for the proof-of-concept robotic platform is an UDOO x86, which is a single board development platform. On the platform, we use MATLAB to implement the proposed McLAM and ScLAM methods in Algorithm 1 and 2, respectively. Moreover, for multichannel audio data acquisition, Playrec [47] was used to probe and record the acoustic signals. The base of the robot used for moving the microphone and loudspeaker array as shown in Fig. 4 is a Kobuki (TMR-K01-W1), which is a wheeled platform with on-board sensors such as accelerometer, odometer, etc., for precise control and movement. The Kobuki platform has a built-in microcontroller (Arduino) that can be programmed with a predefined trajectory to conduct experiments. The microphone and loudspeaker array is connected to a Presonus (1818VSL) audio interface, which was subsequently connected to the UDOO x86 microcomputer. The sampling frequency of the audio interface was set to $48,000$ Hz. Furthermore, a pre-calibrated laser range sensor (TFMini micro Lidar), was also attached to an external microcontroller (Arduino Uno) which was then connected to the UDOO microcomputer to receive a ground truth distance value for the experiments. The laser range finder helps in evaluating the performance of the proposed method at varying distances under different noise conditions. The recorded data was processed by the UDOO x86 microcomputer in real-time as the robot was moving along its trajectory. The final assembly is shown in Fig. 5 where the microphone and loudspeaker array is attached on top of the Kobuki base. The microphones are organized as a UCA with a radius of 0.2 m.

VII. Simulated Results

In this section, we evaluate the performance of the proposed method presented in the earlier sections. We evaluate the performance of the ScLAM and McLAM using simulation data and later, implement the methods on a proof-of-concept hardware platform that was built to test the proposed method in a lab setting. In the first experiments, the performance of the considered TOA/DOA estimators in terms of their accuracy were evaluated and compared against existing methods under different background noise levels. Similarly, in the second experiment, we evaluate the TOA/DOA accuracy of the proposed methods against varying distances from the acoustic reflector. The simulated experiments were conducted using the room impulse response generator [48]. The dimension of the simulated room was set to $8 \times 6 \times 5$ m., the reverberation time (T_{60}) was set to 0.6 s, while the speed of sound was fixed at 343 m/s. The loudspeaker was positioned at the center of an UCA with a radius of 0.2 m and $M = 6$ microphones. A white Gaussian noise sequence was used as the known probe signal, $s(n)$, consisting of $1,500$ samples from a Gaussian distribution. Using such a broadband signal minimizes the effect of spatial aliasing [49] and was also used in [26] to simplify the EM estimator. However, any type of known broadband signal could be used to probe the environment, e.g., a chirp signal or a maximum length sequence (MLS) [50]. Additionally, we have zero-padded the probe signal to get a total length of $20,000$ samples so that we get a longer analysis window which will ensure that all the reflections are captured in the observed signal. The sampling frequency f_s was set to $48,000$ Hz. The background noise for the evaluation was composed of three components: a cylindrical diffuse noise $\mathbf{e}_{m,k}$, the sensor noise, $\mathbf{f}_{m,k}$, and an interfering source, $\mathbf{i}_{m,k}$, e.g., external and directional noise source. The diffuse cylindrical noise was generated using the method in [51] with the rotor noise of a drone from the DREGON database [52]. The audio file used to generate the cylindrical noise has a rotor speed of 70 revolutions per second (RPS). The thermal sensor noise was simulated as a white Gaussian noise while the interfering source is modelled as a point source. These noises were then added to the observed probe signal before



(a) TOA evaluation of the proposed method and state-of-the-art

(b) DOA evaluation of the proposed method and state-of-the-art

Fig. 6: Comparison of the proposed method against state-of-the-art

estimating the parameters of interest from the observations, which can be mathematically written as:

$$\mathbf{y}_{m,k}(n) = \mathbf{x}_{m,k}(n) + \mathbf{v}'_{m,k}(n), \tag{44}$$

$$= \mathbf{x}_{m,k}(n) + \mathbf{e}_{m,k}(n) + \mathbf{f}_{m,k}(n) + \mathbf{i}_{m,k}(n). \tag{45}$$

The noise was added to achieve certain signal-to-diffuse noise ratios (SDNR's), signal-to-sensor noise ratios (SSNR's), and signal-to-inteference noise ratios (SINR's). These are defined, for the microphones $m = 1, \ldots, M$, as

$$\text{SDNR}_m = \frac{\sigma^2_{x_m}}{\sigma^2_{e_m}}, \tag{46}$$

$$\text{SSNR}_m = \frac{\sigma^2_{x_m}}{\sigma^2_{f_m}}, \tag{47}$$

$$\text{SINR}_m = \frac{\sigma^2_{x_m}}{\sigma^2_{i_m}}, \tag{48}$$

where σ^2_y denotes the variance, $\sigma^2_y = E[y^2(n)]$ of a zero-mean signal $y(n)$. In the following experiments, we then compare our proposed method with existing TOA/DOA methods found in the literature. This includes, the multi-channel expectation-maximization method (EM-MC) method proposed in [26] and the common approach to extracting TOAs from the estimated RIR using dual-channel method [53] through the peak-picking approach (RIR-PP). This is done by computing $\widehat{H}(f) = Y(f)/S(f)$ and then taking the inverse DFT to get $\widehat{h} = \mathcal{F}^{-1}\{\widehat{H}(f)\}$. These methods were compared with different variations of the proposed beamforming and NLS-based approach, utilizing DS (DS-NLS), MPDR (MPDR-NLS), and LCMV (LCMV-NLS) beamforming, respectively.

Although the proposed methods can be extended and applied to $3D$ scenarios, we focus on the construction of $2D$ maps in our experiments and therefore set $\psi = 0$. The generalization to $3D$ is left for future research. In contrast to earlier works in [26], [30], the direct-path component is accounted for and thus included within the simulations. Within the experiments, we assume that the robotic platform is closer to one acoustic reflector. Therefore, we choose $R = 1$ to estimate the TOA and the DOA of the nearby acoustic reflector. In order to estimate multiple reflections $R > 1$, we can adopt several iterative methods, such as, RELAX and EM method [40], [54]

A. Implementation of the proposed DOA estimator

To implement the beamformers, we use the overlap-add technique [33]. The output of the microphone was divided into overlapping frames with a frame width of 960 samples (20 ms with a sampling rate of 48 kHz) with a window overlap of 50 %. Later, each frame is multiplied with a Hanning window. These frames are then transformed using a short-time Fourier transform (STFT). For each frequency bin, a beamformer was designed and applied to the received signals $\mathbf{Y}_k$. Furthermore, for each sub-band, the observed signal covariance matrix, needed in forming the MPDR and LCMV beamformers, is estimated as

$$\mathbf{R}_{Y_k} = \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{Y}_k \mathbf{Y}_k^H. \tag{49}$$

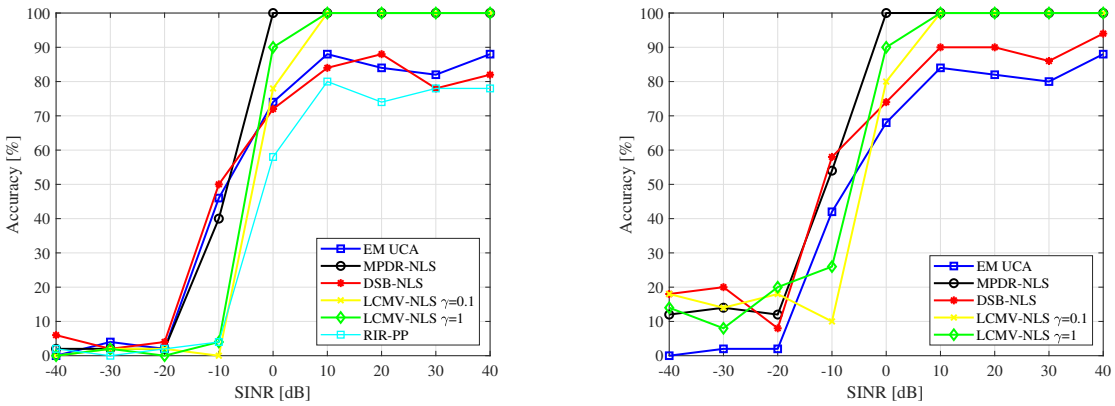
Moreover, to make the beamformers robust against, e.g., miscalibration and reverberation, we regularize the covariance matrix of the observed signal as in [55]

$$\overline{\mathbf{R}}_{Y_k} = (1-\beta)\mathbf{R}_{Y_k} + \beta\frac{\text{Tr}\{\mathbf{R}_{Y_k}\}\mathbf{I}}{M} \tag{50}$$

where β is the regularization parameter, $\text{Tr}(\cdot)$ is the trace of a matrix, and $\mathbf{I}_M$ is the $M \times M$ identity matrix. When evaluating the performance of our estimator, a value of $\beta = 0.1$ was selected for the MPDR beamformer. The noise covariance matrix, $\mathbf{R}_{Y_k}$, in (49) is then replaced by regularized noise variance matrix (50), $\overline{\mathbf{R}}_{Y_k}$. For the LCMV beamformer, we added an additional regularization using γ to mitigate poor matrix conditioning for certain constraint and frequency combinations. This was done as $\mathbf{w}_{\text{LCMV}} = \mathbf{R}_{Y_k}^{-1}\mathbf{D}[\mathbf{A}(\gamma)]^{-1}\mathbf{f}$, where

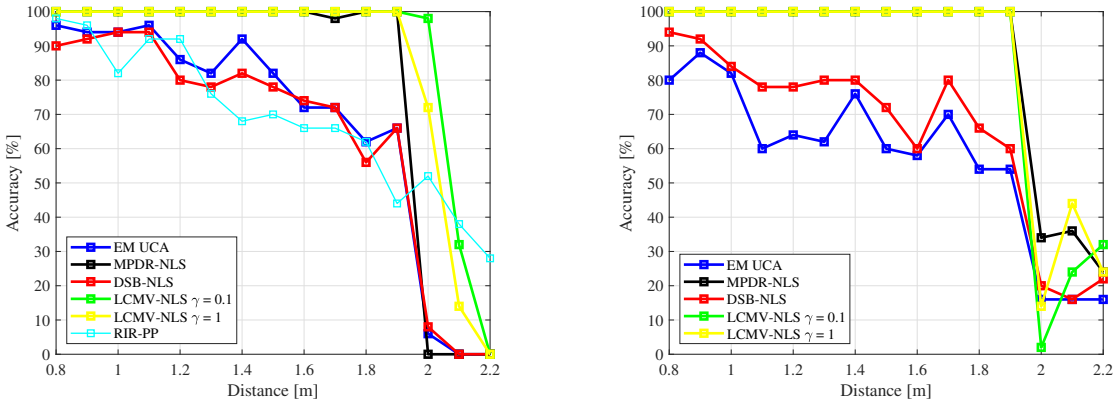
$$\mathbf{A}(\gamma) = (1-\gamma)\mathbf{D}^H\mathbf{R}_{Y_k}^{-1}\mathbf{D} + \gamma\frac{\text{Tr}\{\mathbf{D}^H\mathbf{R}_{Y_k}^{-1}\mathbf{D}\}\mathbf{I}}{M}. \tag{51}$$

Values of $\gamma = 0.1$ and $\gamma = 1$ were selected empirically and used in the simulations. To initiate the method, we probe the environment with a known sound. The observed signals recorded by the microphone array were first processed to



(a) TOA evaluation (b) DOA evaluation

Fig. 7: Evaluation of proposed McLAM method and state-of-the-art against different SINR



(a) TOA evaluation (b) DOA evaluation

Fig. 8: Evaluation of proposed McLAM method and state-of-the-art against different distances

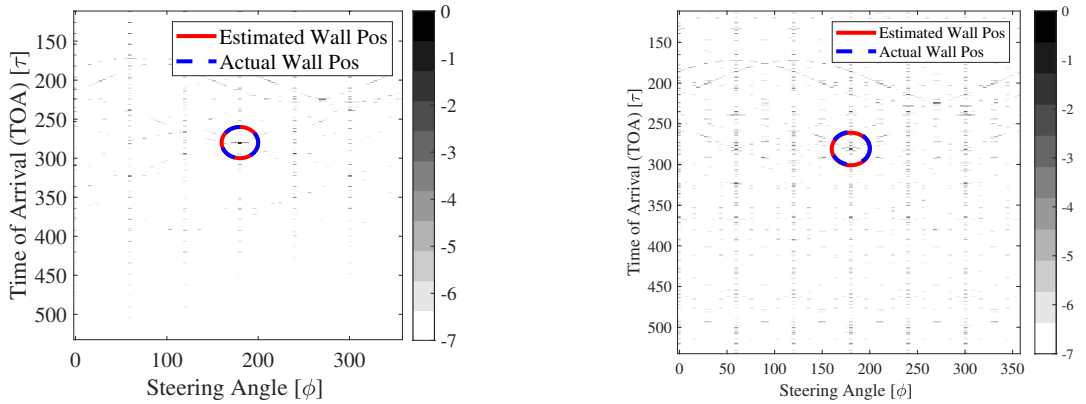
determine the DOA of the acoustic echoes. To estimate the DOA and the TOA of the acoustic echoes, a uniform grid of DOAs over the interval $[0°; 360°]$ and a uniform grid of TOAs corresponding to a distance interval from 0.5 m up to 3 m were considered. The estimators were then evaluated over these grids of candidate DOAs and TOAs. The reason for selecting 0.5 m as the lower bound was done to search for acoustic echoes that are outside the UCA which has a radius of 0.2 m, and so that the direct-path component is not included within the estimation window. Moreover, the upper bound of 3 m were selected because the performance of the proposed method degrades after 3 m according to [26].

B. Comparison of the proposed methods

In our first experiment, we compare our proposed method with the existing TOA/DOA methods. We compare the proposed methods against different SDNRs while placing the setup at a distance of 1 m close to an acoustic reflector. The performance of the proposed methods are shown in Fig. 6. The accuracy is defined as the percentage of estimated TOAs that are within ± 10 % of the true TOA/DOA parameter of the first order acoustic echo computed using the image-source method [56]. This was measured for different SDNRs while the SSNR was fixed to 40 dB and the interfering source was absent in this experiment. For each SDNR values the accuracy was measure over 50 Monte-Carlo simulations. As seen in Fig. 6, the proposed methods, MPDR-NLS and LCMV-NLS, outperforms the existing TOA/DOA methods, EM-MC and RIR-PP, for SDNR levels greater than -10 dB. The DSB-NLS method offers similar performance to EM-MC both in terms of TOA and DOA estimation for most SDNRs as seen in Fig. 6(b).

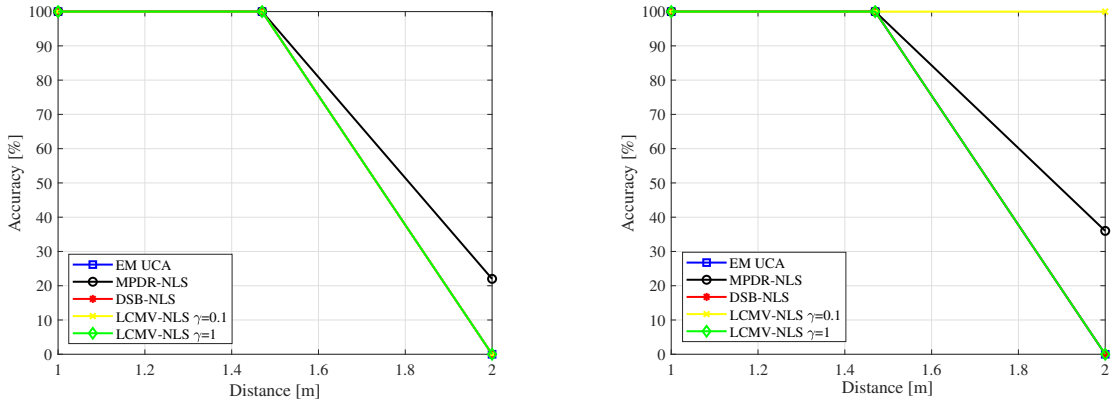
C. Evaluation of the proposed method in the presence of a point source interference

In this experiment, we investigate a scenario where the robot is placed within an environment in the presence of an external interfering source, e.g, a human-speaker, machinery, a radio, etc. In such a scenario, the proposed method will be affected from external elements present in the environment. Therefore, the objective of this experiment is to evaluate both the TOA and the DOA performance of the proposed method against different SINR values. The interfering source was modelled as a point source for this experiment. More specifically, within this experiment, the robotic platform was placed close to an acoustic reflector at a position, $[1, 3, 2.5]$ m within an environment of dimension $8 \times 6 \times 5$ m^3. Furthermore, the external interfering point source was positioned at a location $[2, 1, 2.5]$ m such that the acoustic reflector is at a fixed angle of $180°$ while the point source is placed at an angle



(a) MPDR-NLS costfunction (b) DSB-NLS costfunction

Fig. 9: Acoustic image of MPDR-NLS and DSB-NLS beamformer



(a) TOA evaluation (b) DOA evaluation

Fig. 10: Evaluation of proposed McLAM method and state-of-the-art against varying Distances

of $300°$ with respect to the robotic platform. The performance is shown in Fig. 7. The SINR level selected for this experiment are within the interval $[-40; 40]$ dB while the SDNR and SSNR were both set to 40 dB. Moreover, some additional consideration was taken into account when modelling the interfering point source. For instance, if a human talker is considered as a point source, then it is natural for the human to move within the environment. To model this, the position of the point source was randomize in both the x-axis and y-axis. The interval selected to model point source movement for both x-axis and y-axis are $[1; 3]$ m and $[1; 2]$ m, respectively. As seen in Fig. 7, the TOA of the MPDR-NLS and LCMV-NLS offer more robustness at low SINR compared to EM-MC, RIR-PP and DSB-NLS. Similar performance is seen in the DOA estimation. The accuracy is defined similarly to the previous method with a tolerance of $\pm 10\%$ of true TOA and DOA.

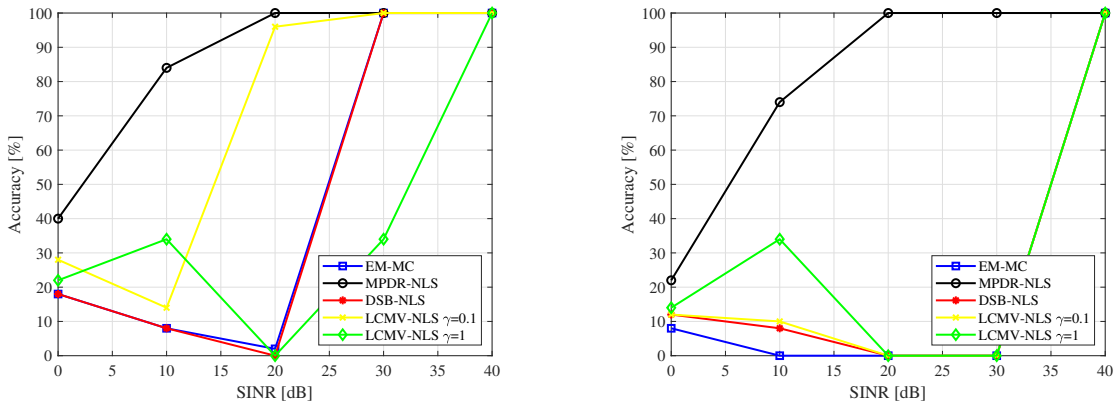
D. Evaluation of proposed methods against distance

In this experiment, we consider a scenario where the robotic platform is placed closer to an acoustic reflector and its distance with respect to the acoustic reflector was changed after every 50 iteration. With this setup, the performance of the proposed method and existing methods over distance interval $[0.8; 2.2]$ m was investigated. Here, the SDNR and SSNR values were set to 40 dB while the interfering source was absent. As seen in Fig. 8, the MPDR-NLS, and the LCMV-NLS variants outperform other methods in terms of TOA estimation and accurately estimate the DOA of the acoustic reflector as it can detect an acoustic reflector up to a distance of around 2 m. This is because at larger distance the acoustic echoes loses its energy quadratically due to inverse square law.

E. Visualizing Acoustic Echoes

Microphone array imaging has been around for quite sometime and are used in aviation [57] for structural analysis as well as to study low frequencies [58]. Similarly, our proposed method could also be used to generate an acoustic image of acoustic echoes which could aid researcher to analyse the direction and distance of acoustic reflectors or be used as input data for the development of deep learning based methods.

To generate an acoustic image using the methods proposed in this paper, we consider estimation of the reflector in $2D$ only, i.e., we only estimate ϕ. For each beamformer we considered a grid of candidate steering angles with a resolution of $4°$ in the interval $[0°; 360°]$. The output of the beamformer is then passed to the NLS estimator in (20), which then estimates τ from candidate grid of delays in $[\tau_{min}; \tau_{max}]$. The resulting $2D$ cost functions are shown in Fig. 9(a) and (b), respectively for one of these experiments. Both plots in Fig. 9(a) and (b) were generated at an SINR of 40 dB with the observed signal



(a) TOA estimation (b) DOA estimation

Fig. 11: Evaluation of the proposed method using proof of concept robotic platform against different SDNR

LIDAR = 1 m	**SINR = 0 dB**		**SINR = 10 dB**		**SINR = 20 dB**		**SINR = 30 dB**		**SINR = 40 dB**	
Methods	μ **[m]**	**RMS error [m]**	μ **[m]**	**RMS error [m]**	μ **[m]**	**RMS error [m]**	μ **[m]**	**RMS error [m]**	μ **[m]**	**RMS error [m]**
MPDR-NLS	1.0558	0.2843	0.9797	0.0992	0.9718	0.0281	0.9890	0.0118	0.9861	0.0138
DSB-NLS	1.0231	0.2802	0.8647	0.8896	0.1103	0.1174	0.9075	0.0924	0.9861	0.0138
EM-MC	1.0229	0.2803	0.8647	0.1485	0.8908	0.1094	0.9075	0.0924	1.0040	0.0039
LCMV-NLS $\gamma = 0.1$	0.9899	0.2603	0.8758	0.1647	1.0387	0.0556	1.0647	0.0647	0.9861	0.0138
LCMV-NLS $\gamma = 1$	1.0325	0.2819	0.8813	0.1409	0.7996	0.2134	0.8084	0.2042	1.0254	0.0254

TABLE I: Evaluation of the proposed McLAM against ground truth and SDNRs

SINR = 40 dB	**MPDR-NLS**		**DSB-NLS**		**EM-MC**		**LCMV-NLS** $\gamma = 0.1$		**LCMV-NLS** $\gamma = 1.0$	
LiDAR [m]	μ **[m]**	**RMSE [m]**	μ **[m]**	**RMSE [m]**	μ **[m]**	**RMSE [m]**	μ **[m]**	**RMSE [m]**	μ **[m]**	**RMSE [m]**
1.01	0.9861	0.0138	0.9861	0.0138	1.004	0.0039	0.9861	0.0138	1.0254	0.0254
1.47	1.4327	0.0372	1.5899	0.1199	1.4542	0.0158	1.4327	0.0372	1.5899	0.1199
2	1.4480	0.6142	0.7610	1.239	0.7610	1.2390	1.3321	0.6716	1.2734	0.8174

TABLE II: Evaluation of the proposed McLAM against ground truth and Distances

including the direct-path component. As seen, the cost function of the MPDR beamformer Fig.9(a) shows a peak at times and angles corresponding to the TOAs and the DOAs at which the beamformer received the acoustic echo, these regions are marked by a red circle. In comparison, the DSB cost function in Fig.9(b) is very noisy, despite evaluating under the high SINR of 40 dB, which makes it difficult to extract the true TOAs and DOAs. This is partly caused by the presence of the direct-path component and the ego-noise, which cannot be sufficiently suppressed by the DSB.

F. Computational cost

The computational cost of the proposed methods were measured using MATLAB's built-in function *timeit*. These were tested on a standard desktop computer running a Microsoft Windows 10 operating system with an Intel Core i7 CPU with a 3.40 GHz processing speed and 16 GB of RAM. A Monte Carlo Simulation of 50 trials were conducted and an average time was calculated. The measured computational time of EM-MC, RIR-PP, LCMV-NLS and MPDR-NLS are 63.25 s, 0.024, 59.75, and 60.65 s, respectively, for $R = 1$ for SINR = 40 dB. The proposed algorithms are computationally expensive when implemented within a robotic platform compared to lidar technologies. To address these issues, one tweak that would allow faster computation on the robot is to probe the environment and use echo detector first to determine whether the robot is closer to an acoustic reflector before proceeding with the proposed DOA/TOA estimates. This accelerate processing and prevents the robot from estimating the parameters when not in the presence of an acoustic reflector. Moreover, tracking, e.g., in the form of gradient searches, may be employed instead of performing a full grid search for every new robot position.

VIII. Experiments using Proof-of-Concept Robotic Platform

In this section, we evaluate the performance of the proposed algorithm (McSLAM) using a robotic platform under different SINRs and distances. The objective of these experiments are to compare our simulated data with the real data to test the performance of the proposed method in real scenarios. Two sets of experiments were conducted using the proof-of-concept robotic platform described earlier in Section VII. The first set of experiments were performed under different SINR and distances while the second set of experiments were performed as a qualitative test to show the mapping ability of the robotic platform while comparing the MPDR-NLS algorithm against the lidar data (ground truth). The data is also summarized in Table I and Table II.

A. Evaluation of the proposed method against different SINRs and distances

Similar to the experiments performed in Section VII-B, the proof-of-concept robotic platform was placed against an acoustic reflector within Aalborg University's Sound Lab. In the first part of the experiment, the robotic platform was placed at varying distances while the SINR value was set fixed to 40 dB. The platform was placed at an interval of $[1, 1.5, 2]$ m. At each distance, the robotic platform probed the environment with a known sound and 50 samples were collected at each distances. The TOA/DOA obtained from the robotic platform are shown in Fig. 10. As seen in the figures, the proposed McLAM algorithm gives an accurate TOA estimate up to a distance of 1.5 m for all combination of spatial filter. The accuracy is defined as the number of estimates that are $\pm 10\%$ of the true TOAs obtained from the lidar data. DOA accuracy is defined similarly.

The next experiment was performed to test the proposed method against different SINR values of the environment. The SINR value of the environment was changed by using a separate loudspeaker playing an audio file from YouTube called Cocktail party[2]. The loudspeaker was placed 6.3 m away from the robotic platform while the robotic platform was fixed at a distance of 1 m away from the acoustic reflector. The SINR of the environment was estimated by dividing the variance of the probed signal, σ_x^2, with the variance of the background noise, σ_v^2. The background noise $\mathbf{v}(n)$ was recorded by the robot before probing the environment. By tuning the volume of the loudspeaker, we then select 5 SINR values, $[0, 10, 20, 30, 40]$ dB. The results for this experiment is shown in Fig. 11. Here, we see that the proposed MPDR-NLS is robust under low SINR value of 10 dB for both TOA and DOA estimation with 80% accuracy. The changes seen in this experiments are discussed in Section IX.

B. Application Examples

Two qualitative experiments were performed to test the performance of the proposed method (MPDR-NLS) in constructing a spatial map of an indoor environments. Two environments were selected to perform this task: 1) a typical office environment with a glass partition and 2) Aalborg University's Sound Lab. These experiments are similar to the one performed in our earlier work with ScLAM [12]. In the first experiment, the McLAM algorithm was used to move within an office environment in a predefined trajectory (straight line). The objective of this experiment was to compare the proposed method against lidar, e.g., in detecting a glass surface. The robot moves a distance of 0.5 m and stops momentarily to probe the environment with a known sound before moving to a new location. The robot repeats this process for $k = 1, \dots K$, positions. The results are shown in Fig. 12. As seen from the experiment, the proposed method is capable of detecting a glass surface compared to the commonly used lidar sensor. This shows that the proposed method is suitable for constructing a spatial map of a typical office environment.

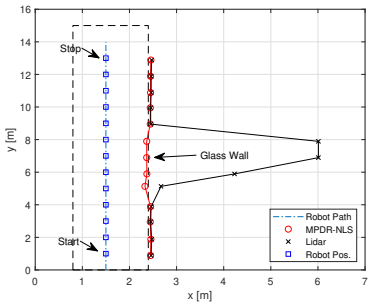
[2]https://youtu.be/IKB3Qiglyro

In the second experiment, Algorithm 1 was used within Aalborg University's Sound Lab, which has a dimension of $5.4 \times 6.38 \times 4.05$ m^3, to construct a spatial map. The objective of this experiment was to move the robot in a more elaborate path within a $3D$ space such that the robot encounters acoustic reflectors as well as empty space along its trajectory. This was done to construct a spatial map of an enclosed environment and also to test the echo detector method presented in Section IV-C. To accomplish this task, the room was divided in to a grid of 20 square boxes, each box has a size of 1 m^2 so that we can ensure that the robot moves along its predefined trajectory and robot's location with respect to the acoustic reflector is always known. Autonomous navigation is also possible but this would require additional on-board sensors, e.g., using Initial Mass Unit (IMU), odometer, gyroscope, etc., to estimate the robot's current position which can then be combined with our estimates to generate a spatial map. As the robot moves within the square grids and follows a predefined trajectory as shown in Fig. 13(b), the robot probes the environment with a known sound. The recorded sound is spatially filtered using MPDR beamformer which is later feed to a NLS estimator for TOA estimation. Later, the estimated data is passed to a echo-detector to determine whether it belongs to an acoustic reflector or is an spurious estimate. Finally, the estimated data are combined with the trajectory of the robotic platform to localize acoustic reflectors. As seen in Fig. 13(b), if the robot moves without the echo detector then it will estimate spurious estimates even when the robot is away from any reflecting surfaces. However, these spurious estimates are removed when echo detector is applied as seen in Fig. 13(c)

IX. Discussion

In the experimental section, the performance of the different methods were evaluated in both simulated and practical environment. According to the simulation results in Fig. 6 and Fig. 8, the MPDR-NLS, LCMV-NLS $\gamma = 0.1$ and LCMV-NLS $\gamma = 1$ methods detects an acoustic reflector up to a distance of around 2 m under the SDNR of -10 dB. Similarly, in practical scenario, the methods could detect an acoustic reflector up to a distance of around 1.5 m as seen in Fig. 10. However, in Fig. 11, only the MPDR-NLS is seen to provide good accuracy at low SINR for TOA/DOA estimation while LCMV-NLS $\gamma = 0.1$ is the second best choice for TOA estimation compared to its other variant LCMV-NLS $\gamma = 1$ which performs less then EM-MC and DSB-NLS when evaluating under different SINR. From these experiments, we can deduce that the MPDR-NLS estimator provides better performance compared to other methods. The results from practical experiments are also detailed in Table. II. The RMSE of all beamformer variants are robust when the distance of the acoustic reflector is less than 1.5 m with respect to the robot. At higher distances, the RMSE increases while the RMSE decreases with higher SDNR values. One noticeable difference we see between simulated and practical evaluation is the low accuracy in practical experiment. There could be various reasons for a lower accuracy in real scenarios, for instance, in our proposed method, we do not take sensor calibration into

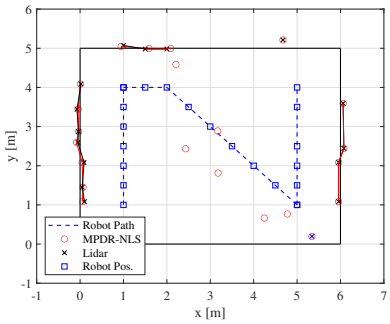
(a) Layout of office with glass surfaces.



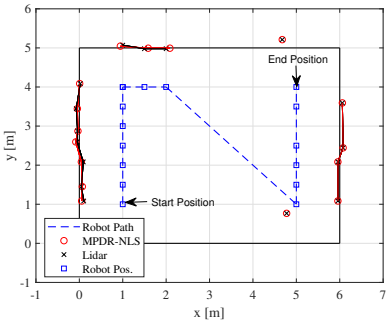
(b) 2D map of an office with glass surfaces.

Fig. 12: Detecting of glass surface at Aalborg University.

(a) Layout of the Sound Lab.



(b) 2D map of the Sound Lab without GLRT.



(c) 2D map of the Sound Lab with GLRT.

Fig. 13: Generating a spatial map of the Sound Lab.

account as well as sensor drift due to an increase in thermal temperature that can arise when the sensors are operating for a long period of time [59].

Additionally, in the simulation, we see that the presence of a point interfering source does not limit the proposed methods' robustness as seen in Fig. 7. This goes to show that an MPDR-NLS and both implementations of the LCMV-NLS method are effective in localizing the acoustic reflector of the environment when compared to other beamformer variants in the presence of an interfering source. In the qualitative experiments, we exploit robot's movement to construct a spatial map of an environment. Here, we compare the current technologies, e.g., lidar, with the proposed McLAM algorithm. As seen in Fig. 13, our proposed method successfully construct a spatial map of an environment. However, one obvious limitation of the proposed method is that lidar accurate distance measurements over longer distance. This is because sound intensity decreases quadratically over distance due to the inverse squares law. One major advantage of our proposed method, on the other hand, is that it provides is that it can be used to detect transparent surfaces as seen in Fig. 12 that are typically found in an office environment, hence our proposed method could compliment existing technologies for spatial map generation. Additionally, we also test our echo-detector in our qualitative experiment. As seen in Fig. 13(b), without echo-detector enabled, spurious estimates are seen when the robotic platform is at an empty space. However, as seen in Fig. 13(c), with the echo-detector enabled, the spurious estimates are removed.

Moreover, both variant of LCMV beamformers behave differently using real data and offers lower performance compared to the simulated results. This could be due to mismatch of the microphones/loudspeaker positions in the array that leads, in which case the null constraint of the LCMV beamformers are not aligned with the direct-path component. Moreover, the LCMV beamformer implicitly assumes the loudspeaker to be a point source, which will not hold for larger loudspeakers located close to the array in practice. In our future work, we plan to incorporate these inaccuracies within our models and methods to improve their robustness.

X. Conclusion

In this paper, we proposed a non-traditional method of constructing a spatial map of an indoor environment using the concept of echolocation. Instead of working in the ultrasonic range, a novelty in this paper is that our proposed method could work in the audible frequency range and any kind of broadband probe signal could be used for DOA/TOA estimation. In addition to this, our proposed algorithms could be used on existing robotic platform, e.g., NAO robots, as these consist of microphones and loudspeaker that work in audible frequency range. As seen from the experimental results, our proposed framework could utilize different beamformers for DOA estimation, which could be combined with our NLS estimator for TOA estimation and an echo detector, to construct a spatial map. One obvious advantage of our framework is that each modules in Fig. 3 could be separately improved over time in order to increase the performance of the acoustic echo localization. Our proposed methods can detect acoustic reflectors up to a distance of 1.5 m at an SINR of 40 dB and robustly estimate TOA at an SINR of 10 dB with 80%

accuracy in realistic scenario. Furthermore, an echo detector is proposed based on the statistics of the background noise that could aid a robot in classifying estimates stemming from acoustic reflectors from spurious estimates. This can help a robot map and environment to facilitate its autonomous planning and movement. In the qualitative experiments, we see that, compared to the commonly used lidar technology, our proposed method can detect transparent surfaces as seen in Fig. 12 and it can also construct a spatial map of an indoor environment as seen in Fig. 13.

In a future iteration of this research, we aim to include a loudspeaker's directivity and transfer function within the signal model. This will enable our algorithms to work more efficiently and help us understand and develop sophisticated sound propagation model that could enable even more accurate construction of spatial maps in an indoor environment. Additionally, we aim to decrease the computation load of the proposed method to make it run faster on embedded devices.

REFERENCES

[1] F. Rovira-Más, V. Saiz-Rubio, and A. Cuenca-Cuenca, "Augmented perception for agricultural robots navigation," *IEEE Sensors Journal*, pp. 1–1, 2020.

[2] N. Melenbrink, J. Werfel, and A. Menges, "On-site autonomous construction robots: Towards unsupervised building," *Automation in Construction*, vol. 119, p. 103312, 2020.

[3] J. Lima, V. Oliveira, T. Brito, J. Gonçalves, V. H. Pinto, P. Costa, and C. Torrico, "An industry 4.0 approach for the robot@factory lite competition," in *IEEE Int. Conf. on Autonomous Robot Systems and Competitions*, 2020, pp. 239–244.

[4] X. Huang, Q. Cao, and X. Zhu, "Mixed path planning for multi-robots in structured hospital environment," *The Journal of Engineering*, vol. 2019, no. 14, pp. 512–516, 2019.

[5] "IEEE standard for robot map data representation for navigation," *IEEE Standard for Robot Map Data Representation for Navigation*, pp. 1–54, 2015.

[6] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part 1," *IEEE robotics and automation magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[7] A. I. M. G. P. Huang and S. I. Roumeliotis, "A quadratic-complexity observability-constrained unscented kalman filter for SLAM," *IEEE Transactions on Robotics*, vol. 29, no. 5, pp. 1226–1243, 2013.

[8] W. L. G. Deng, J. Li and H. Wang, "SLAM: Depth image information for mapping and inertial navigation system for localization," in *Asia-Pacific Conference on Intelligent Robot Systems*. IEEE, 2016, pp. 187–191.

[9] T. R. Wanasinghe, R. G. Gosine, O. De Silva, G. K. I. Mann, L. A. James, and P. Warrian, "Unmanned aerial systems for the oil and gas industry: Overview, applications, and challenges," *IEEE Access*, vol. 8, pp. 166 980–166 997, 2020.

[10] R. Worley, Y. Yu, and S. Anderson, "Acoustic echo-localization for pipe inspection robots," in *IEEE Int. Conf. on Multisensor Fusion and Integration for Intell. Syst.*, 2020, pp. 160–165.

[11] D. Hahnel, R. Triebel, W. Burgard, and S. Thrun, "Map building with mobile robots in dynamic environments," in *Proc. IEEE Int. Conf. Robotics, Automation.*, vol. 2. IEEE, 2003, pp. 1557–1563.

[12] U. Saqib and J. R. Jensen, "A model-based approach to acoustic reflector localization using robotic platform," in *Proc. IEEE Int. Conf. Intell., Robot, Automation.* IEEE, 2018, pp. 1–8.

[13] H. Wei, X. Li, Y. Shi, B. You, and Y. Xu, "Multi-sensor fusion glass detection for robot navigation and mapping," in *WRC Symposium on Advanced Robotics and Automation*. IEEE, 2018, pp. 184–188.

[14] C. Hui and M. Shiwei, "Visual SLAM based on EKF filtering algorithm from omnidirectional camera," in *IEEE 11th International Conference on Electronic Measurement and Instruments*, vol. 2. IEEE, 2013, pp. 660–663.

[15] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5610–5614.

[16] I. Eliakim, Z. Cohen, G. Kosa, and Y. Yovel, "A fully autonomous terrestrial bat-like acoustic robot," *PLOS Computational Biology*, vol. 14, no. 9, 2018.

[17] T. G. Muir and D. L. Bradley, "Underwater acoustics: A brief historical overview through world war 2," *Acoustics today*, vol. 12, no. 3, 2016.

[18] L. Kleeman and R. Kuc, "Mobile robot sonar for target localization and classification," *The International Journal of Robotics Research*, vol. 14, no. 4, pp. 295–318, 1995.

[19] M. Kreković, I. Dokmanić, and M. Vetterli, "EchoSLAM: Simultaneous localization and mapping with acoustic echoes," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 11–15.

[20] J. V. M. L. Nguyen and X. Qiu, "Can a robot hear the shape and dimensions of a room?" *Proc. IEEE Int. Conf. Intell., Robot, Automation.*, 2019.

[21] M. Boutin and G. Kemper, "A drone can hear the shape of a room," *SIAM Journal on Applied Algebra and Geometry*, vol. 4, no. 1, pp. 123–140, 2020.

[22] I. J. Kelly and F. M. Boland, "Detecting arrivals in room impulse responses with dynamic time warping," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1139–1147, 2014.

[23] G. Moschioni, "A new method for measurement of early sound reflections in theaters and halls," in *IMTC/2002. Proceedings of the 19th IEEE Instrumentation and Measurement Technology Conference*, vol. 1, 2002, pp. 425–430 vol.1.

[24] Y. E. Baba, A. Walther, and E. A. P. Habets, "3d room geometry inference based on room impulse response stacks," *J. Audio, Speech, Language Process.*, vol. 26, no. 5, pp. 857–872, 2018.

[25] J. Steckel and H. Peremans, "BatSLAM: Simultaneous localization and mapping using biomimetic sonar," *PLOS ONE*, vol. 8, no. 1, pp. 1–11, 01 2013.

[26] U. Saqib, S. Gannot, and J. R. Jensen, "Estimation of acoustic echoes using expectation-maximization methods," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, pp. 1–15, 2020.

[27] L. B. Nelson and H. V. Poor, "Iterative multiuser receivers for CDMA channels: an EM-based approach," *IEEE Trans. Commun.*, vol. 44, no. 12, pp. 1700–1710, Dec 1996.

[28] M. C. Vanderveen, C. B. Papadias, and A. Paulraj, "Joint angle and delay estimation (JADE) for multipath signals arriving at an antenna array," *IEEE Commun. Lett.*, vol. 1, no. 1, pp. 12–14, Jan 1997.

[29] J. Verhaevert, E. V. Lil, and A. V. de Capelle, "Direction of arrival (DOA) parameter estimation with the SAGE algorithm," *Signal Processing*, vol. 84, no. 3, pp. 619–629, 2004.

[30] U. Saqib and J. R. Jensen, "Sound-based distance estimation for indoor navigation in the presence of ego noise," in *Proc. European Signal Processing Conf.*, 2019.

[31] S. M. Kay, *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993, vol. 2.

[32] J. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency- domain steered beamformer approach," vol. 1, 2004, pp. 103–1038 Vol.1.

[33] C. Pan, J. Chen, and J. Benesty, "Performance study of the MVDR beamformer as a function of the source incidence angle," *J. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 67–79, 2014.

[34] H. Sun, E. Mabande, K. Kowalczyk, and W. Kellermann, "Joint DOA and TDOA estimation for 3d localization of reflective surfaces using eigenbeam MVDR and spherical microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 113–116.

[35] R. C. Maher and E. Hoerr, "Audio forensic gunshot analysis and multilateration," in *Audio Engineering Society Convention 145*. Audio Engineering Society, 2018.

[36] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," vol. 88, no. 4, p. 972–983, Apr. 2008.

[37] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "A study on how pre-whitening influences fundamental frequency estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6495–6499.

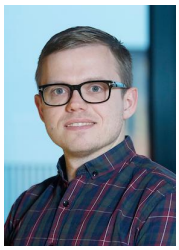
[38] S. Braun, A. Kuklasiński, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators," *J. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1056–1071, 2018.

[39] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[40] J. Li and P. Stoica, "Efficient mixed-spectrum estimation with applications to target feature extraction," *IEEE transactions on signal processing*, vol. 44, no. 2, pp. 281–295, 1996.

[41] S. Hirata, M. K. Kurosawa, and T. Katagiri, "Real-time ultrasonic distance measurements for autonomous mobile robots using cross correlation by single-bit signal processing," in *Proc. IEEE Int. Conf. Robotics, Automation.*, 2009, pp. 3601–3606.
[42] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 2911–2917, 2008. [Online]. Available: https://doi.org/10.1121/1.2987429
[43] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.
[44] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multispeaker LCMV beamformer and postfilter for source separation and noise reduction," *J. Audio, Speech, Language Process.*, vol. 25, no. 5, pp. 940–951, 2017.
[45] L. Liang, F. Kong, C. Martin, T. Pham, Q. Wang, J. Duncan, and W. Sun, "Machine learning–based 3-d geometry reconstruction and modeling of aortic valve deformation using 3-d computed tomography images," *International journal for numerical methods in biomedical engineering*, vol. 33, no. 5, p. e2827, 2017.
[46] W. Yu and W. B. Kleijn, "Room geometry estimation from room impulse responses using convolutional neural networks," *arXiv e-prints*, p. arXiv:1904.00869, Apr. 2019.
[47] R. Humphrey, "Playrec: Multi-channel matlab audio," *URL http://www.playrec.co.uk*, 2007.
[48] E. A. P. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Tech. Rep., 2010, ver. 2.0.20100920. [Online]. Available: https://github.com/ehabets/RIR-Generator
[49] J. Dmochowski, J. Benesty, and S. Affes, "On spatial aliasing in microphone arrays," *IEEE Trans. Signal Process.*, vol. 57, no. 4, pp. 1383–1395, 2009.
[50] D. Florencio and Z. Zhang, "Maximum a posteriori estimation of room impulse responses," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 728–732.
[51] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 2911–2917, 2008.
[52] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "DREGON: Dataset and methods for UAV-embedded sound source localization," in *Proc. IEEE Int. Conf. Intell., Robot, Automation.*, Oct 2018, pp. 5735–5742.
[53] H. Herlufsen, "Dual channel FFT analysis (part I)," in *Brüel & Kjær Technical Review*, no. 1984-1, 1984.
[54] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the em algorithm," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 4, pp. 477–489, 1988.
[55] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise reduction in speech processing*. Springer Science & Business Media, 2009, vol. 2.
[56] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.
[57] M. Legg and S. Bradley, "Automatic 3d scanning surface generation for microphone array acoustic imaging," *Applied acoustics*, vol. 76, pp. 230–237, 2014.
[58] E. G. Williams, J. D. Maynard, and E. Skudrzyk, "Sound source reconstructions using a microphone array," *The Journal of the Acoustical Society of America*, vol. 68, no. 1, pp. 340–344, 1980.
[59] B. Kim and H. Lee, "Acoustical-thermal noise in a capacitive mems microphone," *IEEE Sensors Journal*, vol. 15, no. 12, pp. 6853–6860, 2015.

Usama Saqib received his B.Sc degree in Electrical Engineering from American University of Sharjah, U.A.E. in 2010, and a M.Sc degree in Embedded Systems Engineering from University of Bedfordshire, U.K., in 2015. He is currently working towards his Ph.D. degree from Aalborg University, Denmark. His research interest includes combining audio processing techniques with robotics.



Jesper Rindom Jensen is an associate professor at Aalborg University, Denmark. He is part of Audio Analysis Lab at CREATE where his research interests include signal processing theory and methods for, e.g., robot and drone audition, and microphone arrays. Examples of more specific research interests within this scope are enhancement, separation, localization, tracking, parameter estimation, signal analysis, and modeling. He has published more than 80 papers on these topics in top-tier, peer-reviewed conference proceedings and journals. Moreover, he is the co-author of two books, namely, "Speech Enhancement – A Signal Subspace Perspective" and "Signal Enhancement with Variable Span Linear Filters".