Sound-based Distance Estimation for Indoor Navigation in the Presence of Ego Noise

Usama Saqib and Jesper Rindom Jensen
Audio Analysis Lab, CREATE
Aalborg University
Aalborg, Denmark
{ussa,jrj}@create.aau.dk

***Abstract*—An off-the-shelf drone for indoor operation would come with a variety of different sensors that are used concurrently to avoid collision with, e.g., walls, but these sensors are typically uni-directional and offers limited spatial awareness. In this paper, we propose a model-based technique for distance estimation using sound and its reflections. More specifically, the technique is estimating Time-of-Arrivals (TOAs) of the reflected sound that could infer knowledge about room geometry and help in the design of sound-based collision avoidance. Our proposed solution is thus based on probing a known sound into an environment and then estimating the TOAs of reflected sounds recorded by a single microphone. The simulated results show that our approach to estimating TOAs for reflector position estimation works up to a distance of at least 2 meters even with significant additive noise, e.g., drone ego noise.**

***Index Terms*—robotics, room geometry estimation, acoustic impulse response**

I. Introduction

One of the key issues when it comes to indoor operation of Unmanned Aerial Vehicles (UAVs), also known as drones, is the estimation of the physical boundaries' (e.g., walls) position in order to avoid collision. A common approach to estimating such positions is to use active sensors such as ultrasonic or infrared. Alternatively, camera-based technology combined with advanced computer vision techniques such as Simultaneous Localization And Mapping (SLAM) can be used for landmark or wall position estimation [1]. These techniques, however, have certain limitations. For instance, computer vision based techniques are susceptible to changing lightening conditions and does not work well under low-light conditions. Also, SLAM-based algorithms tends to have difficulty tracking a plain, white surface or landmarks making it harder for SLAM algorithm to estimate a wall position [2]. Moreover, such sensors have a limited field-of-view, so multiple sensors are required to cover all directions around the drone to avoid collisions with walls or other acoustic reflectors, e.g., glass windows. However, localization of a reflector position can be achieved using sound, by estimating the Time-of-Arrivals (TOAs) of acoustic reflections. This is a known estimation problem within the area of acoustic signal processing, which can potentially be implemented on moving robotic platforms or drones. TOAs estimation can thus be important in, e.g., robot and drone (UAV) applications, where it can facilitate acoustic SLAM (ASLAM) [3] and room geometry estimation (RGE) [4]. Moreover, if knowledge of TOAs is obtained, then distance estimation to acoustic reflectors is a straight-forward process given that the speed of sound is known.

In acoustic signal processing, the sound recorded by a microphone consists of a direct path component, first-order early reflections and later reflectins. This acoustic signal propagation from a loudspeaker to a microphone in a room is described by the room impulse response (RIR). The RIR contains information about the TOAs of acoustic reflections, which can be extracted. In the following, we review recent examples of methods utilizing this approach. For instance, in [5], a cell phone is used to probe the walls at different locations of the room with a chirp signal. The sound signals are reflected by the wall which are then correlated against the source signal to find TOAs which in turn helps determining the distances of the reflectors. The distance estimation was done by successfully extracting TOAs from a RIR. This knowledge helps the authors generate a map of the environment. Similarly, in [6], a single collocated microphone and loudspeaker arrangement was placed on a moving robotic platform to estimate distance between the robot and the reflecting surface from TOAs obtained from RIR. The authors in [6] proposes two estimators to calculate distance from TOAs; one involving multilateration techniques that uses the measured TOA values to construct a tangent line of the circle that indicates the position of the wall while the other approach is a Bayesian approach that gives a general solution to the RGE. Common for these state-of-the-art methods is that they require information about the TOA's of the early reflections. Typically, it is assumed that these estimates can be simply obtained through peak picking on an estimated RIR [7]- [10]. This approach is problematic in practice, however, because the individual peaks corresponding to the true TOA's can be small due to dispersion, diffusion, etc., and additive noise (e.g., drone ego noise) can introduce spurious peaks in the estimated RIR [11]. Moreover, the accuracy of the TOA estimates will be limited by the sampling rate [12], unless heuristic interpolation methods are used.

Since a moving drone is always accompanied by ego noise due to the motion of the rotors, we therefore propose and alternative approach to TOA estimation. This is a model-based approach for estimating TOA's based on a model for the early reflections. This enable us to derive a statistically optimal estimator for obtaining TOA estimates directly from

978-9-0827-9703-9/19/$31.00

observed microphone recordings instead of the traditional peak picking on an estimated RIR. This is inspired by the work in [13] on DOA estimation in reverberant environments. When it is desired to estimated multiple TOA's, e.g., to estimate the distance to multiple reflectors, our proposed estimator becomes computational complex due to its multidimensional nature. To tackle this, we propose an iterative estimation procedure based on the RELAX procedure [14].

The remaining part of this paper is organized as follows: Section II formulates the signal model and the problem. Section III describe the proposed TOA estimator based on the model, Section IV describe an iterative procedure for handling multiple reflections, while Section V evaluates the performance and robustness of the proposed solution. Furthermore, Section VI contains our conclusions and future work.

II. Signal Model and Problem Formulation

Consider the setup where a single loudspeaker is situated at $r_s \triangleq [x_s, y_s, z_s]$ that emits a known signal $s(n)$ which is recorded by a microphone placed at location $r_m \triangleq [x_m, y_m, z_m]$. The microphone and sound source are assumed to be collocated and placed inside a room. The observed signal recorded by microphone $y(n)$ is then modeled as follows:

$$y(n) = s(n) * h(n) + v(n) = x(n) + v(n) \tag{1}$$

where $h(n)$ is the impulse response of the room measured from r_s to r_m, $x(n) = s(n) * h(n)$ is the sound source signal including reverberation, $v(n)$ is additive background noise, e.g., ego noise, and $*$ represents the convolution operator. If we decompose (1) as a sum of its direct-path component and its first few reflections, then the observed signal model can be written as:

$$y(n) = \sum_{q=1}^{R} g_q s(n - \tau_q) + v'(n) \tag{2}$$

where g_q is the attenuation of the q^{th} order sound reflection from the source to the microphone, and $v'(n)$ is a combined noise term constituted by the late reverberation (i.e., the $q > R$ components) and the additive background noise. This can be further decomposed as

$$y(n) = x_D(n) + x_R(n) + v'(n), \tag{3}$$

where $x_D(n) = g_1 s(n - \tau_1)$ is the direct path component, and $x_R(n) = \sum_{q=2}^{R} g_q s(n - \tau_q)$ is the early reflection components. The signal decomposition, can also be expressed using simple first order FIR filters, h_q, for $q = 1, \ldots, R$, as

$$y(n) = \sum_{q=1}^{R} h_q * s(n) + v'(n), \tag{4}$$

The transfer function of these filters are given by

$$H_q(z) = g_q z^{-\tau_q}, \tag{5}$$

for $q = 1, \ldots, R$. In many applications, the microphone and the sound source will be placed in fixed positions. In such cases the transfer function of h_1 can be either measured offline or computed analytically using the geometry, i.e., by computing g_1 and τ_1. In such cases, we can thus work with a modified signal model:

$$\overline{y}(n) = \sum_{q=2}^{R} h_q * s(n) + v'(n), \tag{6}$$

where $\overline{y}(n) = y(n) - x_D(n)$, and only the gains and delays of the early reflections are unknown. The estimation problem at hand, is thus to estimate these unknown quantities, τ_q and g_q for $q = 2, \ldots, R$, which are key components in acoustic SLAM and room geometry estimation methods.

III. Non-Linear Least Square (NLS) Estimator

If we take N samples of the observed signals $\mathbf{y}(n) = \begin{bmatrix} y(n) & y(n+1) & \cdots & y(n+N-1) \end{bmatrix}^T$ and assume that we know $s(n)$ we can formulate a nonlinear least squares (NLS) estimator, which is the maximum likelihood estimator when the noise is white Gaussian. Mathematically, this can be formulated as

$$\{\widehat{\mathbf{g}}, \widehat{\boldsymbol{\tau}}\} = \arg\min_{\mathbf{g}, \boldsymbol{\tau}} \|\overline{\mathbf{y}}(n) - \mathbf{x}(n)\|^2 \tag{7}$$

$$= \arg\min_{\mathbf{g}, \boldsymbol{\tau}} \left\| \overline{\mathbf{y}}(n) - \sum_{q=2}^{R} h_q * \mathbf{s}(n) \right\|^2, \tag{8}$$

where

$$\widehat{\boldsymbol{\tau}} = \begin{bmatrix} \widehat{\tau}_2 & \widehat{\tau}_3 & \cdots & \widehat{\tau} \end{bmatrix}^T, \tag{9}$$

$$\widehat{g} = \begin{bmatrix} \widehat{g}_2 & \widehat{g}_3 & \cdots & \widehat{g}_R \end{bmatrix}^T. \tag{10}$$

and $\overline{\mathbf{y}}(n)$, $\mathbf{x}_R(n)$ and $\mathbf{s}(n)$ are defined similarly to $\mathbf{y}(n)$. Moreover, the notation $a * \mathbf{b}$ denotes the convolution of each entry in the vector $\mathbf{b}$ with the scalar a, while $\widehat{c}$ denotes an estimate of the parameter c. Using Parseval's theorem, we can transfer (7) to the frequency domain, which yields

$$\{\widehat{\mathbf{g}}, \widehat{\boldsymbol{\tau}}\} = \arg\min_{\mathbf{g}, \boldsymbol{\tau}} \|\overline{\mathbf{Y}} - \mathbf{X}\|^2 \tag{11}$$

$$= \arg\min_{\mathbf{g}, \boldsymbol{\tau}} \left\| \overline{\mathbf{Y}} - \sum_{q=2}^{R} \mathbf{H}_q \odot \mathbf{S} \right\|^2, \tag{12}$$

where $\overline{\mathbf{Y}}$ and $\mathbf{X}$ are the length K DFT vectors of $\overline{\mathbf{y}}(n)$ and $\mathbf{x}(n)$, respectively. Moreover, $\mathbf{H}_q = g_q \mathbf{Z}(\tau_q)$ and

$$\mathbf{Z}(\tau) = \begin{bmatrix} 1 & e^{-j\tau 2\pi \frac{1}{K}} & \cdots & e^{-j\tau 2\pi \frac{K-1}{K}} \end{bmatrix}^T. \tag{13}$$

That is, when the noise is white Gaussian, the maximum likelihood estimator can also be written as

$$\{\widehat{\mathbf{g}}, \widehat{\boldsymbol{\tau}}\} = \arg\min_{\mathbf{g}, \boldsymbol{\tau}} \left\| \overline{\mathbf{Y}} - \sum_{q=2}^{R} g_q \mathbf{Z}(\tau_q) \odot \mathbf{S}) \right\|^2 \tag{14}$$

$$= \arg\min_{\mathbf{g}, \boldsymbol{\tau}} J(\mathbf{g}, \boldsymbol{\tau}) \tag{15}$$

IV. RELAX NON-LINEAR LEAST SQUARE (RNLS) ESTIMATOR

The estimator in (14) can be shown to be statistically optimal when estimating $\mathbf{g}$ and $\boldsymbol{\tau}$ in the presence of additive white Gaussian noise. However, it is computationally expensive when estimating multiple TOA's as it will require a multi-dimensional search for different values of τ and $\mathbf{g}$, limiting its use in real-time, practical applications. Therefore, a RELAX procedure, originally proposed by [14] and later used in [13], will be adopted to iteratively calculate the value of $\boldsymbol{\tau}$ and $\mathbf{g}$. In order to implement the RELAX method, we will introduce a modified observed signal:

$$\mathbf{Y}_r = \overline{\mathbf{Y}} - \sum_{q=2, q\neq r}^{R} g_q \mathbf{Z}(\tau_q) \odot \mathbf{S} \tag{16}$$

where $\mathbf{Y}_r$ is a modified observation vector containing only the r'th early reflection and additive noise.With this we can then estimate the r'th gain and TOA as

$$\{\widehat{g}_r, \widehat{\tau}_r\} = \arg\min_{g,\tau} \|\mathbf{Y}_r - g_r \mathbf{Z}(\tau_r) \odot \mathbf{S})\|^2 \tag{17}$$

We can then solve for the linear gain parameter g_r by taking the derivative of the cost function and setting it equal to zero, yielding

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

where $\overline{\mathbf{Z}}(\tau_r) = \mathbf{Z}(\tau_r) \odot \mathbf{S}$. This can be inserted back into estimator in (17) to obtain the τ_r as

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

$$= \arg\max_{\tau} \mathbb{R}\left\{ \mathbf{Y}_r^H \overline{\mathbf{Z}}(\tau) \right\}. \tag{20}$$

That is, by solving the optimization problem in (20), we can calculate $\widehat{\tau}_r$ and its corresponding $\widehat{g}_r$ of the r'th reflection. This leads to the iterative RELAX-based procedure:

- Step 1: Assume that $R = 2$, i.e., that we have one first-order reflection of the sound. Estimate g_2 and τ_2 using (18) and (19) from $\mathbf{Y}_2 = \overline{\mathbf{Y}}$.
- Step 2: Assume $R = 3$. Estimate g_3 and τ_3 using (18) and (19) from $\mathbf{Y}_3$ computed with the current estimates of τ_2 and g_2. Then re-estimate g_2 and τ_2 from $\mathbf{Y}_2$ computed using the newly estimated values of g_3 and τ_3. Continue Step 2 until it converges (e.g., $\|J^i - J^{i+1}\|^2 < \epsilon$ where i is the iteration index and ϵ is a threshold value.
- Step 3: Assume $R = 4$. Estimate g_4 and τ_4 using (18) and (19) from $\mathbf{Y}_4$ computed with the current paramater estimates of the other reflections. Then re-estimate g_2 and τ_2 from $\mathbf{Y}_3$ computed using newly estimated reflection parameters. Then re-estimate g_3 and τ_3 from $\mathbf{Y}_3$ computed using the newly estimated reflection parameters. Continue until convergence.
- Remaining Steps: Continue until R is equal to the desired number of early reflections.

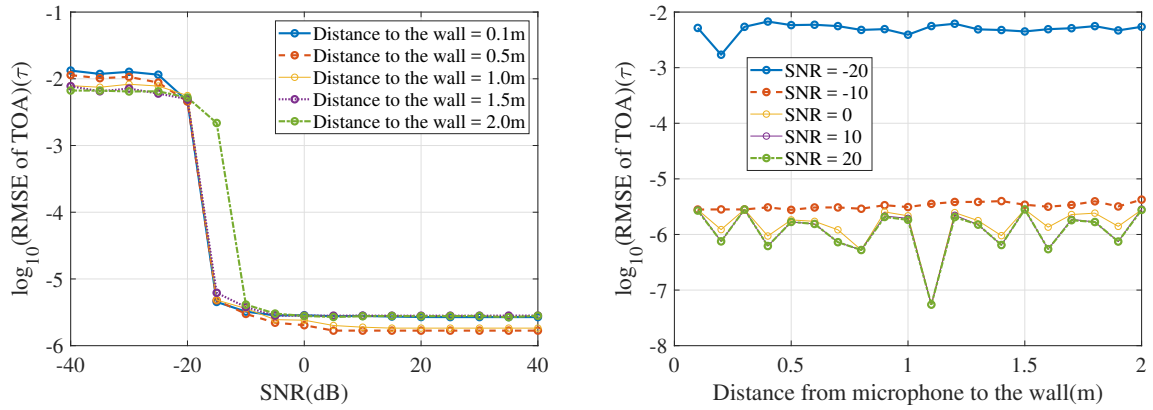
V. EXPERIMENTAL RESULTS AND EVALUATION

In this section, we will evaluate our proposed solution in a simulated room environment obtained with the Multichannel Room Acoustic Simulator (MCRoomSim) [15]. The performance was measured in terms of root mean squared error (RSME) with respect to the distance from the microphone and loudspeaker arrangement to the acoustic reflector, but also with respect to the noise level. Two experiments were conducted; one involving a random noise signal that is transmitted by the loudspeaker for different drone positions while the background noise is white Gaussian; and the other involved using more realistic drone ego noise (e.g., rotor noise) as the background noise. The drone sound was obtained from the DREGON dataset [16].

A room with a dimension of $10 \times 10 \times 6$ m was considered. To test the validity of our proposed solution, we use a collocated microphone-loudspeaker arrangement where the loudspeaker generate a known sound signal and a microphone is placed at a fixed distance of 0.1m directly underneath the loudspeaker. The microphone-loudspeaker arrangement was placed parallel to the x-axis of the room and was located at a position $r_s = [0.1, 5, 3]$ m while the microphone position is $r_m = [0.1, 5, 2.9]$ m. The position of the source and the microphone arrangement in relation to the wall is then varied from 0.1 m to 2 m in 0.2 m steps. Moreover, the sampling frequency was set to 44.1 kHz and the signal length was set to 2000 samples. As discussed in the previous section, we generate a known sound signal. For this particular experiment, we use a random noise signal as our sound source constituted by 2000 samples drawn from a Gaussian distribution. Furthermore, the speed of sound was fixed at 343 m/s. Then, additive white Gaussian noise was introduced at varying SNR levels ranging from -40 dB to 40 dB in 5dB steps. Similarly, the two evaluations (i.e., versus distance and SNR) was carried out with realistic drone ego noise as well. The ϵ value was set to 1×10^{-5}, which we found through experiments to be suitable for accurate estimation of the gains and TOAs with the RELAX procedure. Finally, 50 Monte Carlo simulation were conducted for each of the settings and the average results for each setting are shown.

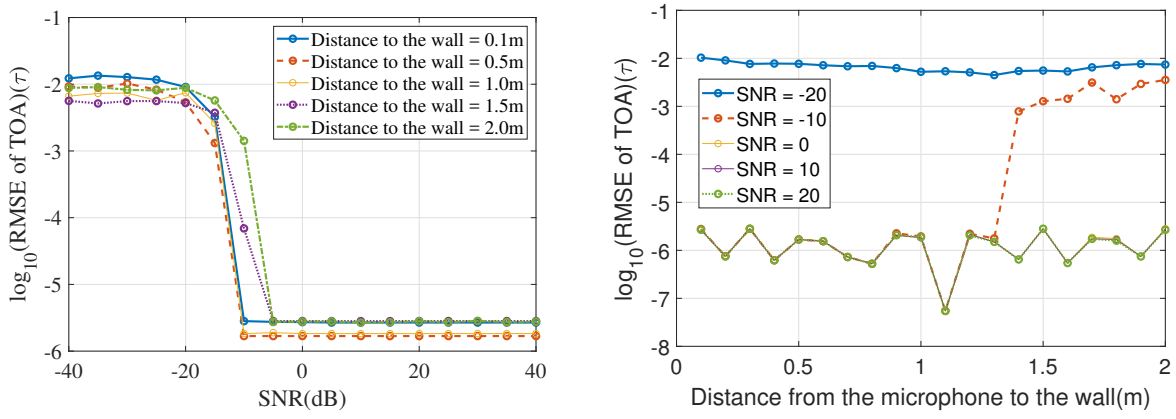
A. Algorithm testing with additive white Gaussian noise as the sensor noise

In the first experiment, we tested the performance of our proposed method with white Gaussian background noise. As seen in Fig.1(a), the proposed method give low estimation errors for SNRs above -15 dB for distances between 0.1 m and 1.0 m, whereas for the higher distances, this is the case for SNRs above -10 dB. Moreover, as seen in 1(b), the proposed method could estimate reflector's distance up to 2m when the background noise level is above -20 dB. Furthermore, the



(a) RMSE vs SNR (b) RMSE vs distance

Fig. 1: Performance metrics of proposed method using a Gaussian noise as the background noise. RMSE of TOA were measured against varying (a) SNR and (b) distance of collocated microphone-loudspeaker from one of the wall



(a) RMSE vs SNR (b) RMSE vs distance

Fig. 2: Performance metrics of propose method using a drone sound as the background noise for a large room. RMSE of TOA were measured against varying (a) SNR and (b) distance of collocated microphone-loudspeaker from one of the wall

algorithm was tested on a standard desktop computer using MATLAB as the simulation environment running on Microsoft Windows 10 operating system with a an Intel Core i7 CPU with 3.40 GHz processing speed and 16 GB of Random Access Memory (RAM). The average time for the algorithm for estimating first-order early reflection is around 1.71 seconds which we believe would be suitable for any drone application. The average computation time could be further reduced when estimating the distances over time and reducing the grid size τ in (19). This is possible if we estimate distances at time instances zero and then at time instance one, the algorithm could use previous estimates of distance to search for TOAs using a reduced grid size.

B. Algorithm testing with drone noise as a background noise

In this experiment, we tested the performance of the proposed method in the presence of drone ego noise as the background noise. As seen in 2(b), the performance is comparable to 1(b). Moreover, it show the TOAs estimator starts to break down at -10 dB when increasing the distance above 1 m. These observations are expected, because the local SNR decreases as the distance of the proposed microphone- loudspeaker setup is increased against the wall. Moreover, similar behaviour will be expected across the remaining SNR values if we evaluate the estimator beyond $2m$.

C. Detecting multiple peaks using RELAX procedure

In a real-world situation, drones could be placed in near multiple acoustic reflectors, in which case we want to estimate multiple TOAs. This can be done with the RELAX procedure, we can estimate all the reflections associated with the reflecting surfaces. This was evaluated with a room of dimensions $6 \times 6 \times 2.4$ m that was simulated in MCRoomSim and the collocated loudspeaker-microphone pair was placed at a loca-

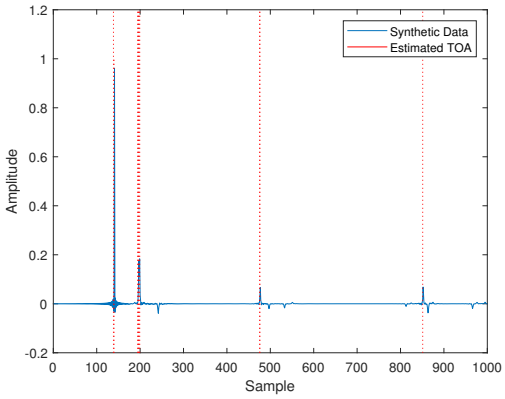


Fig. 3: Detection of multiple reflections using the proposed iterative procedure.

tion of $r_s = [0.1, 1, 3]$m and $r_m = [0.1, 1, 2.9]$m, respectively. As seen in Fig. 3, multiple reflections are recorded by the microphone, each associated with a wall inside a room. The estimated TOAs are close to strongest of the true TOAs of the walls.

VI. Discussion and Future work

In this paper, we proposed an active approach to estimate TOAs using a collocated loudspeaker-microphone arrangement. Our iterative and model-based approach to TOA estimation could, e.g, be implemented on a UAV as part of a collision-avoidance system. The proposed method, is based on a model of early reflections leading to a statistically optimal NLS estimator. To handle the computationally complex problem of estimating multiple TOAs of multiple reflectors in this way, also proposed and iterative implementation of the estimator. In the experiments, we evaluated the method in different noisy scenarios, showing that our proposed method is robust and accurate up to at least a distance of 2 m with negative SNRs, both with additive white Gaussian noise and more realistic ego noise from the rotors of a drone. This indicate that the propose probing approach would not be too intrusive, as the TOAs can be estimated even when the ego noise is louder than the probing sound. In the future iteration of this research, we will test the performance of our proposed method on an actual UAV. Moreover, we aim at extending the proposed method to use an array of microphones so we can estimate both the distance and the direction of the early reflections.

References

[1] M. A. Al-Ammar et al., "Comparative Survey of Indoor Positioning Technologies, Techniques, and Algorithms," 2014 International Conference on Cyberworlds, Santander, 2014, pp. 245-252.

[2] E. Eade and T. Drummond, "Edge landmarks in monocular SLAM," in In Proc. British Machine Vision Conf, 2006.

[3] I. Dokmanic, L. Daudet, and M. Vetterli, From acoustic room reconstruction to SLAM, in Proc. IEEE Int. Conf. Acoust. Speech Signal Process.,Shanghai, China, 2016, pp. 63456349.

[4] Y. E. Baba, A. Walther and E. A. P. Habets, 3D Room Geometry Inference Based on Room Impulse Response Stacks, in IEEE/ACMTransactions on Audio, Speech, and Language Processing, vol. 26, no.5, pp. 857-872, May 2018.

[5] T. Wang, F. Peng and B. Chen, "First order echo based room shape recovery using a single mobile device," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp. 21-25.

[6] M. Krekovi, I. Dokmani and M. Vetterli, EchoSLAM: Simultaneous localization and mapping with acoustic echoes, 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),Shanghai, 2016, pp. 11-15.

[7] S. Tervo, J. Patynen, and T. Lokki, Acoustic reflection localization from room impulse responses, ACTA Acustica United Acustica, vol. 98, pp. 418440, 2012.

[8] G. Defrance, L. Daudet, J.-D. Polack: Detecting arrivals within room impulse responses using matching pursuit. 11th International Conference on Digital Audio Effects (DAFx-08), 2008, 14.

[9] G. Defrance, L. Daudet, J.-D. Polack: Using matching pursuit for estimating mixing time within room impulse responses. Acta Acustica united with Acustica 95 (2009) 10821092.

[10] C. Falsi, D. Dardari, L. Mucchi, M. Z. Win: Time of arrival estimation for uwb localizers in realistic environments. EURASIP Journal on Applied Signal Processing (2006) 152152.

[11] I. Kelly and F. Boland, Detecting arrivals in room impulse responses with dynamic time warping, IEEE/ACMTrans. Audio, Speech, Lang. Process., vol. 22, no. 7, pp. 11391147, Jul. 2014.

[12] J. R. Jensen, J. K. Nielsen, M. G. Christensen and S. Holdt Jensen, "On frequency domain models for TDOA estimation," IEEE International Conference on Acoustics, Speech and Signal Processing, 2015, pp. 11–15.

[13] J. R. Jensen, J. K. Nielsen, R. Heusdens and M. G. Christensen, DOA estimation of audio sources in reverberant environments, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp. 176-180.

[14] J. Li and P. Stoica, "Efficient mixed-spectrum estimation with applications to target feature extraction," in IEEE Transactions on Signal Processing, vol. 44, no. 2, pp. 281-295, Feb. 1996.

[15] A. Wabnitz, N. Epain, C. Jin and A. Van Schaik, "Room acoustics simulation for multichannel microphone arrays," in Proceedings of the International Symposium on Room Acoustics, pp. 1-6, 2010.

[16] M. Strauss, P. Mordel, V. Miguet and A. Deleforge, "DREGON: Dataset and Methods for UAV-Embedded Sound Source Localization," in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1-8, 2018.