

Estimation of acoustic echoes using expectation-maximization methods

Saqib, Usama; Gannot, Sharon; Jensen, Jesper Rindom

Published in:
Eurasip Journal on Audio, Speech, and Music Processing

DOI (link to publication from Publisher):
[10.1186/s13636-020-00179-z](https://doi.org/10.1186/s13636-020-00179-z)

Creative Commons License
CC BY 4.0

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

[Link to publication from Aalborg University](#)

Citation for published version (APA):
Saqib, U., Gannot, S., & Jensen, J. R. (2020). Estimation of acoustic echoes using expectation-maximization methods. *Eurasip Journal on Audio, Speech, and Music Processing*, 2020(1), Article 12.
<https://doi.org/10.1186/s13636-020-00179-z>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -


Take down policy

If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

RESEARCH

Open Access

Estimation of acoustic echoes using expectation-maximization methods

Usama Saqib^{1,2†}, Sharon Gannot^{1†} and Jesper Rindom Jensen^{1*†} 

Abstract

Estimation problems like room geometry estimation and localization of acoustic reflectors are of great interest and importance in robot and drone audition. Several methods for tackling these problems exist, but most of them rely on information about times-of-arrival (TOAs) of the acoustic echoes. These need to be estimated in practice, which is a difficult problem in itself, especially in robot applications which are characterized by high ego-noise. Moreover, even if TOAs are successfully extracted, the difficult problem of echolabeling needs to be solved. In this paper, we propose multiple expectation-maximization (EM) methods, for jointly estimating the TOAs and directions-of-arrival (DOA) of the echoes, with a uniform circular array (UCA) and a loudspeaker in its center for probing the environment. The different methods are derived to be optimal under different noise conditions. The experimental results show that the proposed methods outperform existing methods in terms of estimation accuracy in noisy conditions. For example, it can provide accurate estimates at SNR of 10 dB lower compared to TOA extraction from room impulse responses, which is often used. Furthermore, the results confirm that the proposed methods can account for scenarios with colored noise or faulty microphones. Finally, we show the applicability of the proposed methods in mapping of an indoor environment.

Keywords: TOA estimation, DOA estimation, Expectation-maximization, Active source localization, Robot/drone audition, Prewhitening

1 Introduction

During the past decade, there has been an increased research interest in robot and drone audition [1–3]. Hearing capabilities enable robots to understand and interact with humans [4]. Moreover, it has also been proven useful for sensing the physical environment. For example, it can be used for estimating the locations of acoustic sources, the position of a robot or drone, and the positions of acoustic reflectors and for inferring room geometry [5, 6]. Potentially, this can enable autonomous indoor operation of robots and drones.

Some different approaches for tackling the above estimation problems have already been considered. In a broad sense, these can be classified as being either passive or

active. The passive approach relies on using external sound sources in the environment to conduct the localization. Examples of such sources could be human speech, noise from machinery, or ego-noise from other robots or drones. This approach was, e.g., used for solving the acoustic simultaneous localization and mapping (aSLAM) problem [7–9]. With aSLAM, it is possible to estimate the robot location relative to a number of passive acoustic sources in its vicinity. One obvious advantage of such passive approaches is that they are non-intrusive since only already existing sounds are used in the estimation. This comes at a price, however, since many acoustic sources, such as human speech, contains periods of inactivity, which can lead to unreliable estimates. This is particularly true with moving objects such as robots and drones. Moreover, to facilitate autonomous indoor operation, it is of great importance to also estimate the location of acoustic reflectors, e.g., walls, which is difficult with the

*Correspondence: jrj@create.aau.dk

[†]Usama Saqib, Sharon Gannot and Jesper Rindom Jensen contributed equally to this work.

¹Audio Analysis Lab, CREATE, Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark

Full list of author information is available at the end of the article

passive approach, where only relative timing information is available.

The alternative, which we consider in this paper, is the active approach. In this approach, one or more loudspeakers are used to probe the environment using a known signal. Subsequently, a number of microphones are used to record the sound after it has propagated through the environment. Compared to the passive approach, this facilitates the estimation of the times-of-arrival (TOAs) of both the direct and reflected sound components. With this information, the localization accuracy can be increased significantly compared to the passive approach, and the task of acoustic reflector localization becomes less complex. In the following, we briefly outline some of the most recent and relevant work on active approaches. Some authors have considered the problem of estimating both room geometry and a robot's position with a setup consisting of a collocated microphone and speaker pair [10]. To achieve this, they utilize TOA estimates of the first order reflections. The TOAs are assumed known or estimated beforehand. To tackle the estimation problem with the considered single-channel setup (i.e., one microphone and one loudspeaker), they consider multiple observations from different time instances and locations, i.e., movement is assumed. Based on this, they then proposed two different methods: a method based on basic trigonometry, and another one based on Bayesian filtering. A similar approach also based on a priori RIR/TOA knowledge was considered using a multichannel setup in the context of robotics in [11]. Other authors considered an approach where the TOAs of the first order echoes are utilized for estimating the arbitrary convex room shapes [12]. As briefly mentioned, these as well as other active approaches do not consider the TOA estimation problem, which is an equally important and difficult problem in itself due to, e.g., spurious estimates [13]. Moreover, methods relying on first- and second-order reflections only suffer from the inevitable problem of echolabeling [14]. In addition to this, many methods are based on only one microphone and one loudspeaker, but this leads to ambiguity in the mapping of the TOA estimates of the first-order reflections unless more transducers are included or movement is exploited.

These issues will be addressed in this paper, where we consider a setup consisting of a microphone array which is collocated with a single loudspeaker. More specifically, we consider a uniform circular array that could be placed on the perimeter of, e.g., a drone or robot platform, with a loudspeaker located in its center. With this setup in mind, we propose a number of expectation-maximization (EM) methods for estimating both the TOAs and directions-of-arrival (DOA) of a number of the acoustic reflections. This has the benefit of not only yielding more accurate TOAs compared to a single-channel approach, but also of

reducing the ambiguity of the estimated reflections since the DOA is estimated simultaneously. In fact, this means that the estimates directly reveal the locations of mirror sources, which greatly simplifies the task of localizing the acoustic reflector positions. The proposed methods are derived in the time-domain, and, thus, estimates the parameters of interest directly from the recorded signals, i.e., not from estimated room impulse responses as in numerous state-of-the-art methods. While joint TOA and DOA estimation is a new topic in the context of robot and drone audition, it has been considered previously in multiuser and multipath communication systems [15–17]. However, it has not yet been considered for acoustic reflector localization to the best of our knowledge. The paper builds on the results reported in our earlier paper [18] and extends on this work in several ways. First, we relax our previous noise assumptions and derive the optimal estimators for these more realistic scenarios. The first scenario deals with spatially independent white Gaussian noise with different noise variances across the microphones, e.g., to simulate low quality or faulty microphones. The second scenario considered deals with spatio-temporally correlated noise, which we tackle using prewhitening. Here, we include different approaches for the prewhitening. Moreover, we have included a beamformer interpretation of one of the proposed multichannel estimators, which provides an intuitive understanding of the EM-based method. In addition to this, we included further experimental work to show case the merits of the different proposed estimators and how they compare with traditional methods.

The rest of the paper is organized as follows. In Section 2, we propose the signal model for the considered setup along with a problem formulation. Then, in Section 3, we briefly revisit the single-channel EM method for TOA estimation, which serves as our reference method. Inspired by this, we then proceed with the derivation of the different TOA and DOA estimators in Section 4. Finally, the paper closes with the experimental results and conclusions in Sections 5 and 6, respectively.

2 Problem formulation

We now proceed to lay the foundation for the derivation of EM-based methods for estimating the TOA and TDOA of the acoustic echoes. This is done by formulating the relevant temporal and spatial signal models.

2.1 Time-domain model

Consider a setup with a single loudspeaker and M microphones that are assumed to be collocated on some hardware platform, e.g., a mobile robot or a drone. The loudspeaker is used to probe the environment with a known sound while the microphones are used to record the sound emitted by the loudspeaker including its

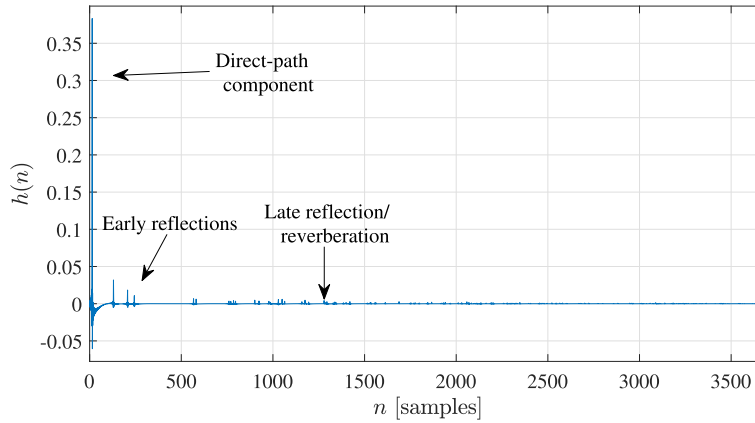


Fig. 1 An example of a synthetic room impulse response illustrating its different parts, i.e., the direct-path component, early reflections and late reflections/ reverberation

acoustic reflections from physical objects and boundaries, e.g., walls. Both the microphones and loudspeakers are assumed to be omnidirectional and ideal. While this assumption might not hold in practice, we do not consider the handling of non-ideal characteristics in this paper. As suggested in other work [5], this might be partly addressed by estimating and introducing another filter accounting for the hardware characteristics, which may also be included in the methods proposed later. Moreover, the non-ideal characteristic of the hardware, i.e., loudspeakers could be modeled as shown in [5], but this is not included when formulating the following estimator.

We can then formulate a general model for the signal recorded by microphone m , for $m = 1, \dots, M$, as

$$y_m(n) = h_m * s(n) + v_m(n) = x_m(n) + v_m(n), \quad (1)$$

where, $x_m(n) = h_m * s(n)$, h_m is the acoustic impulse response as measured from the loudspeaker to the m th microphone, and $s(n)$ is a known signal being played back by the loudspeaker. Finally, $v_m(n)$ is an additive noise term, which is supposed to model ego-noise from a robot/drone platform, interfering sound sources (e.g., human speakers), thermal sensor noise, etc., that is, the signal $s(n)$ is used to probe the environment to, eventually, facilitate the estimation of the parameters of the acoustic echoes, such as their TOA and TDOA. Thus, we proceed by rewriting the observation model as a sum of the individual reflections¹ in noise, i.e.,

$$y_m(n) = \sum_{r=1}^{\infty} g_{m,r} s(n - \tau_{\text{ref},r} - \eta_{m,r}) + v_m(n), \quad (2)$$

with $g_{m,r}$ being the attenuation of the r th reflection from the loudspeaker to the m th microphone, e.g., due to the inverse square law for sound propagation and sound

absorption in the acoustic reflectors. Furthermore, $\eta_{m,r} = \tau_{m,r} - \tau_{\text{ref},r}$ is the TDOA of the r th component measured between a reference point and microphone m , while $\tau_{m,r}$ and $\tau_{\text{ref},r}$ are the TOAs of the r th component on microphone m and the reference point, respectively.

Acoustic impulse responses often exhibit a certain structure, which can be characterized by two parts: the early part, which is sparse in time and contains the direct-path and early reflections, and the late part, which is a more stochastic, dense, and characterized by decaying tail of late reflections (Fig. 1). This suggests that we can split the model as [19]

$$y_m(n) = \sum_{r=1}^R g_{m,r} s(n - \tau_{\text{ref},r} - \eta_{m,r}) + d_m(n) + v_m(n), \quad (3)$$

where R is the number of early reflections, and $d_m(n)$ is the late reverberation. A common assumption is that the late reverberation can be modeled as a spatially homogeneous and isotropic sound field with time-varying power but known coherence function [20]. If we collect N samples from each microphone and assume stationarity within the corresponding time frame, the vector model for our observations becomes:

$$\mathbf{y}_m(n) = \sum_{r=1}^R g_{m,r} \mathbf{s}(n - \tau_{\text{ref},r} - \eta_{m,r}) + \mathbf{d}_m(n) + \mathbf{v}_m(n), \quad (4)$$

with $\mathbf{y}_m(n)$, $\mathbf{s}(n)$, $\mathbf{d}_m(n)$, and $\mathbf{v}_m(n)$ being vectors comprising N time samples of $y_m(n)$, $s(n)$, $d_m(n)$, and $v_m(n)$, respectively, e.g.,

$$\mathbf{y}_m(n) = [y_m(n) \quad \dots \quad y_m(n + N - 1)]^T,$$

This leaves us with the problem of estimating R unknown TOAs and MR TDOAs from the observations

¹In our definition, the direct-path component is one of the reflections, i.e., the 0th order reflection corresponding to $r = 1$.

$y_m(n)$, for $m = 1, \dots, M$. However, if we know the geometry of the loudspeaker and microphone array configuration, we can significantly reduce the dimensionality of this problem by further parametrizing the TDOAs in terms of the directions-of-arrival (DOAs).

2.2 Array model

While the array model can in principle be chosen arbitrarily, we choose to exemplify the TDOA modeling with a setup where the loudspeaker is placed in the center of a uniform circular array (UCA). Such a setup could be placed on, e.g., a robot or drone platform to enable the estimation of the angle of and distance to acoustic reflectors, e.g., to facilitate autonomous and sound-based navigation.

If we assume the reference point to be the center of the UCA, it can be shown that the TDOAs, for a setup like this, can be modeled as

$$\eta_{m,r} = d \sin \psi_r \cos(\theta_m - \phi_r) \frac{f_s}{c} \quad (5)$$

where d is the radius of the UCA, ψ_r and ϕ_r are the inclination and azimuth angles of the r th reflection, respectively, and θ_m is the angle of the m th microphone on the circle forming the UCA. These definitions are illustrated in the UCA example in Fig. 2. In addition to this, f_s is the sampling frequency, and c is the speed of sound.

The TDOA model in (5) can then be combined with the observation model in (4). By doing this, the estimation problem at hand is then simplified to the estimation of $2R$ angles, i.e., ψ_r and ϕ_r , for $r = 1, \dots, R$, rather than MR TDOAs. It should be noted here that the considered UCA configuration introduces ambiguities, e.g., an acoustic reflection impinging from an elevation of 0° will result in the same TDOAs as an acoustic reflection mirrored around the UCA plane, i.e., at an elevation angle of 180° . However, this ambiguity can easily be accounted for by applying the proposed methods on array structures

with microphones in all three dimensions, e.g., spherical microphone arrays [21].

3 Single-channel estimation

Before presenting the proposed TOA and TDOA estimators, we briefly revisit an EM-based method for single-channel TOA estimation, i.e., that is with a setup consisting of one loudspeaker and one microphone. The original version of this method was proposed in [22] under a white Gaussian noise assumption and serves as a reference for the proposed methods.

3.1 White Gaussian noise

In the following, we leave out the microphone index, i.e., subscript m , since only a single microphone is considered. We assume that the additive noise, i.e., both the late reverberation and the background noise is independent and identically distributed white Gaussian and zero-mean. Later, as part of the proposed multichannel methods, this assumption is substituted with a more realistic one, where the late reverberation is modeled as being spatio-temporally correlated. The signal model in (4) then reduces to

$$y(n) = \sum_{r=1}^R g_r s(n - \tau_r) + v(n), \quad (6)$$

where $v(n)$ is distributed as $\mathcal{N}(\mathbf{0}, \mathbf{C})$, with $\mathbf{0}$ being a vector of zeroes, $\mathbf{C} = E[v(n)v^T(n)] = \sigma_v^2 \mathbf{I}_N$ is the $N \times N$ covariance matrix of $v(n)$, σ_v^2 is its variance, \mathbf{I}_N denotes the $N \times N$ identity matrix, and $E[\cdot]$ is the mathematical expectation operator. The maximum likelihood (ML) estimator of the unknown parameters, i.e., the gains and the TOAs, is well known to be the nonlinear least squares (NLS) criterion in this case, i.e.,

$$\{\hat{\tau}, \hat{\mathbf{g}}\} = \underset{\tau, \mathbf{g}}{\operatorname{argmin}} \left\| y(n) - \sum_{r=1}^R g_r s(n - \tau_r) \right\|^2, \quad (7)$$

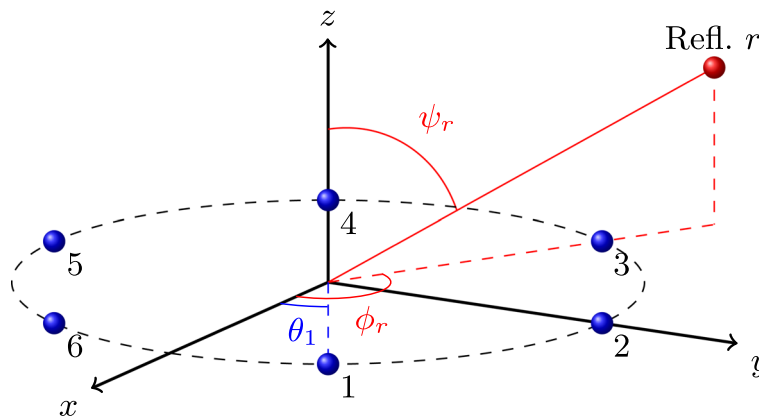


Fig. 2 Example of a uniform circular array with six microphones

where

$$\boldsymbol{\tau} = [\tau_1 \quad \cdots \quad \tau_R]^T,$$

$$\mathbf{g} = [g_1 \quad \cdots \quad g_R]^T.$$

While this estimator is statistically efficient, it also requires computationally costly search since the cost function is high-dimensional and non-convex with respect to the TOAs.

A computationally more efficient way of implementing this estimator could be to adopt the expectation-maximization (EM) approach for superimposed signals proposed in [22]. The concept behind this approach is to define the complete data as the observation of all individual signals, i.e., each of the individual early reflections in our case. According to the previously stated signal model in (4), the individual observations can be modeled as

$$\mathbf{x}_r(n) = g_r \mathbf{s}(n - \tau_r) + \mathbf{v}_r(n), \quad (8)$$

for $r = 1, \dots, R$, where $\mathbf{v}_r(n)$ is obtained by arbitrarily decomposing the combined noise term, $\mathbf{v}(n)$, into R different components adhering to

$$\sum_{r=1}^R \mathbf{v}_r(n) = \mathbf{v}(n). \quad (9)$$

Moreover, the observed signal can be written as the sum of individual observations such as:

$$\mathbf{y}(n) = \sum_{r=1}^R \mathbf{x}_r(n). \quad (10)$$

Following [22], we let the individual noise terms be independent, zero-mean, white Gaussian, and distributed as $\mathcal{N}(\mathbf{0}, \beta_r \mathbf{C})$. Furthermore, the scaling factors, β_r are non-negative, real-valued scalars that satisfy

$$\sum_{r=1}^R \beta_r = 1. \quad (11)$$

Under these assumptions, it can be shown that the EM algorithm for estimating the gains and the time-of-arrivals is given by [22]

E-step: for $r = 1, \dots, R$, compute

$$\hat{\mathbf{x}}_r^{(i)}(n) = \hat{g}_r^{(i)} \mathbf{s}(n - \hat{\tau}_r^{(i)}) + \beta_r \left[\mathbf{y}(n) - \sum_{k=1}^R \hat{g}_k^{(i)} \mathbf{s}(n - \hat{\tau}_k^{(i)}) \right]. \quad (12)$$

M-step:

$$\{\hat{g}_r, \hat{\tau}_r\}^{(i+1)} = \underset{g, \tau}{\operatorname{argmin}} \|\hat{\mathbf{x}}_r^{(i)}(n) - g \mathbf{s}(n - \tau)\|^2, \quad (13)$$

where $^{(i)}$ is denoting the iteration index. If the length, N , of the analysis window is long compared to the length of

the known signal, $\mathbf{s}(n)$, the M-step can be simplified as

$$\hat{\tau}_r = \underset{\tau}{\operatorname{argmax}} \hat{\mathbf{x}}_r^T(n) \mathbf{s}(n - \tau), \quad (14)$$

$$\hat{g}_r = \frac{\hat{\mathbf{x}}_r^T(n) \mathbf{s}(n - \hat{\tau}_r)}{\|\mathbf{s}(n)\|^2}. \quad (15)$$

We see that the estimation problem has been greatly simplified with this signals decomposition, since we now have $2R$ one-dimensional estimators rather than a $2R$ -dimensional estimator as in (7). From this simplified version of the M-step, we can make some interesting interpretations. First in (14), the individual observations are applied with a matched filter based on the known source signal. The TOA is estimated as the one maximizing the output power of the matched filter. Secondly, the estimated TOAs are used to obtain closed-form estimated of the gains in (15), which is based on a least squares fit between the known source signal and the estimated contribution of the r th component.

4 Multichannel estimation

We now proceed to consider the multichannel case, where we have one loudspeaker and multiple microphones. First, we consider a white Gaussian noise scenario similar to Section 3.1 where the noise is independent across the microphones, after which we turn to the more realistic scenarios with correlated noise.

4.1 Spatially independent white Gaussian noise

If we first assume that the noise is temporally white Gaussian and independent and the late reverberation is negligible, the signal model in (4) reduces to

$$\mathbf{y}_m(n) = \sum_{r=1}^R g_{m,r} \mathbf{s}(n - \tau_{\text{ref},r} - \eta_{m,r}) + \mathbf{v}_m(n), \quad (16)$$

for $m = 1, \dots, M$. Subsequently, we can aggregate the observations from all microphones in one model as

$$\mathbf{y}(n) = \sum_{r=1}^R \mathbf{H}(\eta_r, \mathbf{g}_r) \mathbf{s}(n - \tau_{\text{ref},r}) + \mathbf{v}(n) = [\mathbf{y}_1 \quad \mathbf{y}_2 \quad \cdots \quad \mathbf{y}_M]^T, \quad (17)$$

where $\mathbf{v}(n)$ is the stacked noise terms from each microphone defined similarly to $\mathbf{y}(n)$, and

$$\boldsymbol{\eta}_r = [\eta_{1,r} \quad \eta_{2,r} \quad \cdots \quad \eta_{M,r}]^T,$$

$$\mathbf{g}_r = [g_{1,r} \quad g_{2,r} \quad \cdots \quad g_{M,r}]^T.$$

In addition to this, we note that, under the assumptions of spatial independent white Gaussian noise, the covariance matrix, \mathbf{C} of the stacked noise, $\mathbf{v}(n)$ is diagonal and given by

$$\mathbf{C} = \operatorname{diag}(\sigma_{v_1}^2 \mathbf{I}_N, \sigma_{v_2}^2 \mathbf{I}_N, \dots, \sigma_{v_M}^2 \mathbf{I}_N), \quad (18)$$

where $\text{diag}(\cdot)$ is the operator constructing a diagonal matrix from the input of scalars/matrices) and \mathbf{C} is the $MN \times MN$ covariance matrix. Furthermore,

$$\mathbf{H}(\eta_r, \mathbf{g}_r) = \begin{bmatrix} g_{1,r} \mathbf{D}_{\eta_{1,r}}^T & \cdots & g_{M,r} \mathbf{D}_{\eta_{M,r}}^T \end{bmatrix}^T, \quad (19)$$

and \mathbf{D}_η is a circular shift matrix which delays a signal by $-\eta$ samples.

With these definitions, the ML estimator for the problem at hand becomes

$$\{\hat{\mathbf{g}}, \hat{\tau}, \hat{\eta}\} = \underset{\mathbf{g}, \tau, \eta}{\text{argmin}} J(\mathbf{g}, \tau, \eta), \quad (20)$$

where

$$J(\mathbf{g}, \tau, \eta) = \left\| \mathbf{y}(n) - \sum_{r=1}^R \mathbf{H}(\eta_r, \mathbf{g}_r) \mathbf{s}(n - \tau_{\text{ref},r}) \right\|_{\mathbf{C}^{-1}}^2 \quad (21)$$

such that $\|\mathbf{x}\|_{\mathbf{W}}^2 = \mathbf{x}^T \mathbf{W} \mathbf{x}$, where \mathbf{W} denotes the weighted 2-norm of \mathbf{x} . Moreover, \mathbf{g} , τ , and η are the parameter vectors containing all unknown gains, TOAs and TDOAs, respectively. In the single-channel case, the ML estimator ends up being high-dimensional and non-convex, resulting in a practically infeasible computational complexity if implemented directly. Therefore, we propose to adopt the EM framework also for the multichannel scenario.

Like in the single-channel approach, we consider the complete data to be all the individual observations of the reflections, but in this case from all the M microphones. Each of the observations can thus, for $r = 1, \dots, R$, be modeled as

$$\mathbf{x}_r = \mathbf{H}(\eta_r, \mathbf{g}_r) \mathbf{s}(n - \tau_{\text{ref},r}) + \mathbf{v}_r(n). \quad (22)$$

The decomposition is assumed to satisfy the conditions in (9)–(11). Then, it can be shown that the EM-algorithm for the multichannel estimation problem is given by

E-step: for $r = 1, \dots, R$, compute

$$\begin{aligned} \hat{\mathbf{x}}_r^{(i)}(n) &= \mathbf{H}(\hat{\eta}_r^{(i)}, \hat{\mathbf{g}}_r^{(i)}) \mathbf{s}(n - \hat{\tau}_{\text{ref},r}^{(i)}) \\ &+ \beta_r \left[\mathbf{y}(n) - \sum_{k=1}^R \mathbf{H}(\hat{\eta}_k^{(i)}, \hat{\mathbf{g}}_k^{(i)}) \mathbf{s}(n - \hat{\tau}_{\text{ref},k}^{(i)}) \right]. \end{aligned} \quad (23)$$

M-step: for $r = 1, \dots, R$,

$$\{\hat{\mathbf{g}}_r, \hat{\tau}_r, \hat{\eta}_r\}^{(i+1)} = \underset{\mathbf{g}, \tau, \eta}{\text{argmin}} J_r(\mathbf{g}, \tau, \eta), \quad (24)$$

with $J_r(\mathbf{g}, \tau, \eta)$ being a weighted least squares estimator defined as

$$J_r(\mathbf{g}, \tau, \eta) = \left\| \hat{\mathbf{x}}_r^{(i)}(n) - \mathbf{H}(\eta, \mathbf{g}) \mathbf{s}(n - \tau) \right\|_{\mathbf{C}^{-1}}^2. \quad (25)$$

If we explicitly write the cost function, we get

$$\begin{aligned} J_r(\mathbf{g}, \tau, \eta) &= \sum_{m=1}^M \frac{\|\hat{\mathbf{x}}_{m,r}(n)\|^2}{\sigma_{v_m}^2} \\ &+ \|\mathbf{s}(n - \tau)\|^2 \sum_{m=1}^M \frac{g_{m,r}^2}{\sigma_{v_m}^2} \\ &- 2 \sum_{m=1}^M \frac{g_{m,r} \hat{\mathbf{x}}_{m,r}^T(n) \mathbf{D}_{\eta_m}}{\sigma_{v_m}^2} \mathbf{s}(n - \tau), \end{aligned} \quad (26)$$

This can be used to simplify the M-step by making a few observations. Clearly, the first term in this expression does not depend on any parameter of interest. Moreover, if we assume that the analysis window is long compared to the length of the known source signal, $\mathbf{s}(n)$, we observe that the second term does not depend on either the TOAs or the TDOAs. That is, to estimate these time parameters, we only need to consider the maximization of the last term, i.e.,

$$\begin{aligned} \{\hat{\tau}_{\text{ref},r}, \hat{\eta}_r\} &= \underset{\tau, \eta}{\text{argmax}} \sum_{m=1}^M \frac{g_{m,r} \hat{\mathbf{x}}_{m,r}^T(n) \mathbf{D}_{\eta_m}}{\sigma_{v_m}^2} \\ &\times \mathbf{s}(n - \tau), \end{aligned} \quad (27)$$

The gains, $g_{m,r}$, and the noise statistics, $\sigma_{v_m}^2$, are unknown in practice. However, if the noise is assumed (quasi-)stationary, its variance can be estimated from microphone recordings acquired before emitting the known source signal, $\mathbf{s}(n)$. By taking the partial derivative of (26) with respect to $g_{m,r}$, we obtain the following closed-form estimate for $g_{m,r}$

$$\hat{g}_{m,r} = \frac{\hat{\mathbf{x}}_{m,r}^T(n) \mathbf{D}_{\hat{\eta}_m} \mathbf{s}(n - \hat{\tau}_{\text{ref},r})}{\|\mathbf{s}(n)\|^2}, \quad (28)$$

If the reflections are assumed to be in the far-field of the array, we can further simplify the estimators. In this case, the gains of reflection r will be the same across all microphones for $r = 1, \dots, R$. That is, we can instead estimate the TOAs and TDOAs as

$$\begin{aligned} \{\hat{\tau}_{\text{ref},r}, \hat{\eta}_r\} &\approx \underset{\tau, \eta}{\text{argmax}} \left(\sum_{m=1}^M \frac{\hat{\mathbf{x}}_{m,r}^T(n) \mathbf{D}_{\eta_m}}{\sigma_{v_m}^2} \right) \\ &\times \mathbf{s}(n - \tau). \end{aligned} \quad (29)$$

Subsequently, the gain estimator can then be reformulated as

$$\hat{g}_r = \left(\sum_{m=1}^M \frac{1}{\sigma_{v_m}^2} \right)^{-1} \sum_{m=1}^M \frac{\hat{\mathbf{x}}_{m,r}^T \mathbf{D}_{\hat{\eta}_m} \mathbf{s}(n - \hat{\tau}_{\text{ref},r})}{\sigma_{v_m}^2 \|\mathbf{s}(n)\|^2}, \quad (30)$$

If the geometry of the loudspeaker and microphone configuration is known, we further reduce the dimensionality of the estimation problem. This is achieved by parameterizing the TDOAs, $\eta_{m,r}$, for $r = 1, \dots, R$ and $m = 1, \dots, M$

using the array model, e.g., the one for a UCA configuration formulated in (5). Then, the TOA and TDOA estimator in the M-step can be written as

$$\{\hat{\tau}_{\text{ref},r}, \hat{\phi}_r, \hat{\psi}_r\} \approx \underset{\tau, \phi, \psi}{\text{argmax}} \left(\sum_{m=1}^M \frac{\hat{\mathbf{x}}_{m,r}^T(n) \mathbf{D}_{\eta_m}}{\sigma_{v_m}^2} \right) \times \mathbf{s}(n - \tau), \quad (31)$$

where η_m is replaced by the expression in (5). In this way, we only need to estimate two angles for each reflection, whereas the estimator in, e.g., (30) requires the estimation of M TDOAs (or $M - 1$ if one of the microphone positions is used as the reference point). That is, the computational benefits of using the array model increases as we increase the number of microphones. It can be shown that the resulting estimators in the M-step has an interesting interpretation as minimum variance distortionless response (MVDR) beamforming followed by a matched filter as we show in the following subsection.

4.2 Beamformer interpretation

Intuitively, if we were able to observe the reflections individually in noise and the noise is differently distributed across the microphones, then it would be natural to apply an MVDR beamformer to these to optimally account for the noise when estimating the TOAs and TDOAs. Let us consider the scenario where we have a filtering matrix, \mathbf{W} , which we use to process the individually observed reflections in (22):

$$\mathbf{z}(n) = \mathbf{W}^T \mathbf{x}_r(n). \quad (32)$$

Then, we define the residual noise power after this filtering as the normalized sum of the residual noise variances over the different time indices included in $\mathbf{z}(n)$, i.e., $n, n + 1, \dots, n + N - 1$. Mathematically, this is equivalent to

$$\begin{aligned} \sigma_{v_f}^2 &= \mathbb{E} \left[\frac{1}{N} \text{Tr} \left\{ \mathbf{W}^T \mathbf{v}_r(n) \mathbf{v}_r^T(n) \mathbf{W} \right\} \right] \\ &= \frac{\beta_r}{N} \text{Tr} \left\{ \mathbf{W}^T \mathbf{C} \mathbf{W} \right\}, \end{aligned} \quad (33)$$

where $\text{Tr}\{\cdot\}$ is the trace operator. Obviously, by inspection of the individual observation model in (22), we can see that the following expression needs to be satisfied for the filter to be distortionless with respect to the known source signal:

$$\mathbf{W}^T \mathbf{H}(\eta_r, \mathbf{g}_r) = \mathbf{I}_N. \quad (34)$$

That is, omitting the arguments of the steering matrix $\mathbf{H}(\eta_r, \mathbf{g}_r)$ for brevity, the problem of finding the MVDR solution for \mathbf{W} can be formulated as

$$\min_{\mathbf{W}} \text{Tr} \left\{ \mathbf{W}^T \mathbf{C} \mathbf{W} \right\} \quad \text{s.t.} \quad \mathbf{W}^T \mathbf{H} = \mathbf{I}_N. \quad (35)$$

It can be shown that the solution to the quadratic optimization problem with linear constraints is given by

$$\mathbf{W}_M = \mathbf{C}^{-1} \mathbf{H} \left(\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H} \right)^{-1}. \quad (36)$$

If we then apply the MVDR filtering matrix to the estimated observation of the r th reflection in noise, careful inspection reveals that

$$\mathbf{x}_r^T(n) \mathbf{W}_M = \frac{\sum_{m=1}^M \frac{\mathbf{g}_m \mathbf{x}_{m,r}^T(n) \mathbf{D}_{\eta_m}}{\sigma_{v_m}^2}}{\sum_{m=1}^M \frac{\mathbf{g}_m^2}{\sigma_{v_m}^2}}. \quad (37)$$

The denominator is clearly independent of either the TOA or the TDOAs of the r th reflection, so if the objective is to estimate these, we only need to consider the numerator. Interestingly, the numerator resembles the first part of the cost function in (28). This reveals the following interpretation of the M-step. First, the individual observations of the reflections are filtered by an MVDR filter, and the resulting output is then processed by a matched filter with the transmitted signal. The TOA and TDOAs that maximizes the output power of this operation are then the estimates for the r th reflection. This is in line with the findings in [23–25], where it was shown that the output of an MVDR/LCMV beamformer provide the sufficient statistics for estimating individual signals.

4.3 Spatio-temporally correlated noise

We now consider the scenario, where the noise is spatio-temporally correlated, a scenario practically encountered. For example, the late reverberation is often modeled as spatially homogeneous and isotropic sound field [19], resulting in a degree of spatial coherence which is dependent on the distance between the measurement points. Moreover, there might be interfering, quasi-periodic noise sources in the recording environment, like human talkers, ego-noise from a drone/robot, etc. For such scenarios, we can rewrite the model in (4) as

$$\mathbf{y}(n) = \sum_{r=1}^R \mathbf{H}(\eta_r, \mathbf{g}_r) \mathbf{s}(n - \tau_{\text{ref},r}) + \mathbf{d}(n), \quad (38)$$

where

$$\mathbf{d}(n) = \left[\mathbf{d}_1^T(n) \quad \mathbf{d}_2^T(n) \quad \cdots \quad \mathbf{d}_M^T(n) \right]^T. \quad (39)$$

To deal with scenarios like this, we can preprocess the observed signals, such that the white Gaussian noise assumptions of the EM method is satisfied.

One way to achieve this is to use spatio-temporal decorrelation technique. Let us consider the correlated noise terms of the model in (4), i.e., $\mathbf{d}_m(n)$, for $m = 1, \dots, M$. First, we define the spatio-temporal correlation matrix as

$$\mathbf{C}_d = \mathbb{E} \left[\mathbf{d}(n) \mathbf{d}^T(n) \right]. \quad (40)$$

If we assume that this matrix is Hermitian and positive definite, the Cholesky factorization of it is given by

$$\mathbf{C}_d = \mathbf{L}\mathbf{L}^T, \quad (41)$$

where \mathbf{L} is a lower triangular matrix with real and positive diagonal entries. That is, to whiten the noise term before estimating the unknown parameters, we can left-multiply the observation in (38) with \mathbf{L}^{-1} [26]. The prewhitened observations are thus given by

$$\begin{aligned} \bar{\mathbf{y}}(n) &= \mathbf{L}^{-1}\mathbf{y}(n) \\ &= \mathbf{L}^{-1} \sum_{r=1}^R \mathbf{H}(\eta_r, \mathbf{g}_r) \mathbf{s}(n - \tau_{\text{ref},r}) + \bar{\mathbf{d}}(n), \end{aligned} \quad (42)$$

where $\bar{\mathbf{d}}(n) = \mathbf{L}^{-1}\mathbf{d}(n)$. Based on this and [22], we end up with the following EM method for estimating the acoustic reflection parameters when the noise is correlated in time and space:

E-step: for $r = 1, \dots, R$, compute

$$\begin{aligned} \hat{\mathbf{x}}_r^{(i)}(n) &= \mathbf{H}(\hat{\eta}_r^{(i)}, \hat{\mathbf{g}}_r^{(i)}) \mathbf{s}(n - \hat{\tau}_{\text{ref},r}^{(i)}) \\ &+ \beta_r \left[\mathbf{y}(n) - \sum_{k=1}^R \mathbf{H}(\hat{\eta}_k^{(i)}, \hat{\mathbf{g}}_k^{(i)}) \mathbf{s}(n - \hat{\tau}_{\text{ref},k}^{(i)}) \right]. \end{aligned} \quad (43)$$

M-step: for $r = 1, \dots, R$,

$$\{\hat{\mathbf{g}}_r, \hat{\tau}_r, \hat{\eta}_r\}^{(i+1)} = \underset{\mathbf{g}, \tau, \eta}{\text{argmin}} \bar{J}_r(\mathbf{g}, \tau, \eta). \quad (44)$$

where

$$\bar{J}_r(\mathbf{g}, \tau, \eta) = \left\| \mathbf{L}^{-1} \left(\hat{\mathbf{x}}_r^{(i)}(n) - \mathbf{H}(\eta, \mathbf{g}) \mathbf{s}(n - \tau) \right) \right\|^2, \quad (45)$$

Eventually, we can explicitly write the cost function for the M-step as

$$\begin{aligned} \bar{J}_r(\mathbf{g}, \tau, \eta) &= \mathbf{x}_r^T(n) \mathbf{C}_d^{-1} \mathbf{x}_r(n) \\ &+ \mathbf{s}^T(n - \tau) \mathbf{H}^T(\eta, \mathbf{g}) \mathbf{C}_d^{-1} \mathbf{H}(\eta, \mathbf{g}) \mathbf{s}(n - \tau) \\ &- 2\mathbf{x}_r^T(n) \mathbf{C}_d^{-1} \mathbf{H}(\eta, \mathbf{g}) \mathbf{s}(n - \tau), \end{aligned} \quad (46)$$

Compared with the cost function in (26), the minimization of (46) is more challenging. For example, the second term in (46) will generally depend on the DOA/TDOAs. That is, if we assume the reflections to be in the far-field of the array, we can adopt an iterative estimation scheme, where we first estimate the TOA and TDOAs, then update the TDOAs, and, finally, estimate the gains, i.e., for $r = 1, \dots, R$:

Step 1: Obtain estimates of the TOA and TDOAs as

$$\{\hat{\tau}_r, \hat{\eta}_r\} = \underset{\tau, \eta}{\text{argmax}} \mathbf{x}_r^T(n) \mathbf{C}_d^{-1} \bar{\mathbf{H}}(\eta, \mathbf{g}) \mathbf{s}(n - \tau), \quad (47)$$

where

$$\bar{\mathbf{H}}(\eta) = \begin{bmatrix} \mathbf{D}_{\eta_1}^T & \cdots & \mathbf{D}_{\eta_M}^T \end{bmatrix}^T.$$

Step 2: Update the TDOA estimates as

$$\hat{\eta}_r = \arg \min_{\eta} \bar{J}_{2,r}(g_r, \eta) + \bar{J}_{3,r}(g_r, \eta), \quad (48)$$

where

$$\bar{J}_{2,r}(g_r, \eta) = g_r^2 \mathbf{s}(n - \hat{\tau}_r)^T \bar{\mathbf{H}}^T(\eta) \mathbf{C}_d^{-1} \bar{\mathbf{H}}(\eta) \mathbf{s}(n - \hat{\tau}_r) \quad (49)$$

$$\bar{J}_{3,r}(g_r, \eta) = -2g_r \mathbf{x}_r^T(n) \mathbf{C}_d^{-1} \bar{\mathbf{H}}(\eta) \mathbf{s}(n - \hat{\tau}_r). \quad (50)$$

Step 3: Estimate the unknown gain as

$$\hat{g}_r = \frac{\mathbf{x}_r^T(n) \mathbf{C}_d^{-1} \bar{\mathbf{H}}(\hat{\eta}_r) \mathbf{s}(n - \hat{\tau}_r)}{\mathbf{s}^T(n - \hat{\tau}_r) \bar{\mathbf{H}}^T(\hat{\eta}_r) \mathbf{C}_d^{-1} \bar{\mathbf{H}}(\hat{\eta}_r) \mathbf{s}(n - \hat{\tau}_r)}. \quad (51)$$

with the TOA and TDOA estimates from (47) and (48), respectively. If needed, these steps can then be repeated until convergence. It is also possible to simplify the M-step further by using particular signals as the known signal, $\mathbf{s}(n)$. By close inspection of the second term of the cost function in (48), we get

$$\begin{aligned} \bar{J}_{2,r}(g_r, \eta) &= g_r^2 \sum_{i=1}^M \sum_{j=1}^M c_{ij} \\ &\times \mathbf{s}^T(n - \tau - \eta_i) \mathbf{s}(n - \tau - \eta_j), \end{aligned} \quad (52)$$

where c_{ij} denotes the (i, j) th element of \mathbf{C}_d^{-1} . This reveals that, if the known probe signal is an uncorrelated noise sequence, it is reasonable to assume that this term is independent of both the TOA and the TDOAs, meaning that we can skip the update step in (48).

4.4 Kronecker decomposition

Another challenge with the prewhitening based estimator is the inversion of the noise covariance matrix, \mathbf{C}_d , which has a high dimension of $NM \times NM$. However, if we assume that the covariance matrix is separable, we can approximate it with two smaller matrices [27], i.e.,

$$\mathbf{C}_d \approx \mathbf{C}_s \otimes \mathbf{C}_t. \quad (53)$$

where \mathbf{C}_s and \mathbf{C}_t represents the spatial and temporal correlation matrices of dimensions $M \times M$ and $N \times N$, respectively, and \otimes denotes the Kronecker product operator. Since $(\mathbf{C}_s \otimes \mathbf{C}_t)^{-1} = \mathbf{C}_s^{-1} \otimes \mathbf{C}_t^{-1}$, we now only need to invert these smaller matrices, which is both numerically and computationally preferable. Moreover, we can now conduct the prewhitening using the Cholesky factorization of these smaller matrices due to the mixed-product property, yielding

$$\mathbf{C}_s \otimes \mathbf{C}_t = \mathbf{L}_s \mathbf{L}_s^T \otimes \mathbf{L}_t \mathbf{L}_t^T = (\mathbf{L}_s \otimes \mathbf{L}_t) (\mathbf{L}_s^T \otimes \mathbf{L}_t^T). \quad (54)$$

In other words, by assuming separability, we can approximate \mathbf{L} in (41) by $\mathbf{L}_s \otimes \mathbf{L}_t$. Eventually, it can be shown that, for uncorrelated probe signals, the Kronecker product decomposition allows us to rewrite the first step of the M-step in (44) as

Step 1:

$$\{\hat{\tau}_r, \hat{\eta}_r\} = \underset{\tau, \eta}{\operatorname{argmax}} \mathbf{x}_r^T(n) \left(\mathbf{C}_s^{-1} \otimes \mathbf{C}_t^{-1} \right) \bar{\mathbf{H}}(\eta, \mathbf{g}) \mathbf{s}(n - \tau),$$

$$= \underset{\tau, \eta}{\operatorname{argmax}} \operatorname{tr} \left(\mathbf{X}_r^T(n) \mathbf{C}_t^{-1} \mathbf{S}_{\tau, \eta}(n) \mathbf{C}_s^{-1} \right) \quad (55)$$

$$= \underset{\tau, \eta}{\operatorname{argmax}} \sum_{m=1}^M \tilde{\mathbf{x}}_{m,r}^T(n) \tilde{\mathbf{s}}(n - \tau - \eta_m) \quad (56)$$

where

$$\mathbf{X}_r(n) = [\mathbf{x}_{1,r}(n) \quad \cdots \quad \mathbf{x}_{M,r}(n)], \quad (57)$$

$$\mathbf{S}_{\tau, \eta}(n) = [\mathbf{D}_{\eta_1} \mathbf{s}(n - \tau) \quad \cdots \quad \mathbf{D}_{\eta_M} \mathbf{s}(n - \tau)],$$

$$= [\mathbf{s}(n - \tau - \eta_1) \quad \cdots \quad \mathbf{s}(n - \tau - \eta_M)], \quad (58)$$

and the vectors $\tilde{\mathbf{x}}_{m,r}(n)$ and $\tilde{\mathbf{s}}(n - \tau - \eta_m)$ are the prewhitened observation and probe signals for microphone m , respectively, defined as the m th columns of the following matrices:

$$\tilde{\mathbf{X}}_r(n) = \mathbf{L}_t^{-1} \mathbf{X}_r(n) \mathbf{L}_s^{-T} \quad (59)$$

$$\tilde{\mathbf{S}}_{\tau, \eta}(n) = \mathbf{L}_t^{-1} \mathbf{S}_{\tau, \eta}(n) \mathbf{L}_s^{-T}. \quad (60)$$

These expressions can be interpreted in the following way. The left hand multiplication with \mathbf{L}_t^{-1} corresponds to temporal prewhitening of all the microphone signals, whereas the right hand multiplication with \mathbf{L}_s^{-T} corresponds to spatial prewhitening of all time snapshots.

Step 2: With the Kronecker decomposition, the second term of the cost function in (49) becomes

$$\bar{J}_{2,r}(g_r, \eta) = g_r^2 \operatorname{tr}(\tilde{\mathbf{S}}_{\tau, \eta}^T(n) \tilde{\mathbf{S}}_{\tau, \eta}(n)). \quad (61)$$

This does not depend on the TOAs and TDOAs, so the Kronecker decompositions allow us to skip the intermediate step of updating the TDOAs as in (48). We can therefore directly proceed to conducting the closed form estimate of the gains as

$$\hat{g}_r = \frac{\sum_{m=1}^M \tilde{\mathbf{x}}_{m,r}^T(n) \tilde{\mathbf{s}}(n - \tau - \eta_m)}{M \|\tilde{\mathbf{s}}(n)\|^2}. \quad (62)$$

Even after all the presented simplifications and assumptions, the computational complexity of the proposed methods might still be considered relatively high due to their iterative and multidimensional nature. However, although not considered in this paper, we expect that further reductions in the computational complexity can be obtained by employing, e.g., the space alternating generalized expectation (SAGE) algorithm rather than the EM algorithm [28], or through a recursive EM procedure as suggested in [29], where the number of iterations per time instance can be reduced by instead tracking the parameters of interest over time.

4.5 Temporal prewhitening with filter

One issue with this prewhitening approach still is that the number samples in time might be relatively high in practice. The consequence of this is that, even with the Kronecker decomposition of the noise correlation matrix, the inversion of \mathbf{L}_t might be intractable in practice since its dimensions equal the number of time samples. An alternative approach could be to use a lower order filter for the prewhitening instead [30]. If we assume that the noise follows an autoregressive model, we can approximate it as:

$$d(n) \approx \sum_{p=1}^P a_p d(n - p). \quad (63)$$

Given the noise correlation matrix, \mathbf{C}_t , we can obtain the AR coefficients of the noise using the Levinson-Durbin recursion. The prewhitening filter is then formed using the AR coefficients as the coefficients of a P th order FIR filter, $h_{pw}(p) = a_p$. Subsequently, the prewhitened signals are obtained as

$$\tilde{x}_{m,r}(n) = \sum_{p=0}^P h_{pw}(p) x_{m,r}(n - p), \quad (64)$$

$$\tilde{s}(n) = \sum_{p=0}^P h_{pw}(p) s(n - p), \quad (65)$$

where $h_{pw}(0) = 1$.

4.6 Covariance estimation

In the previous subsections, we have considered the covariance matrices as known quantities. However, we need to estimate these from the observed data in practice. If no particular structure is assumed for the covariance matrix, a common approach is to use the following estimator [31]

$$\hat{\mathbf{C}}_d = \frac{1}{N - K + 1} \sum_{n=0}^{N-K} \mathbf{d}(n) \mathbf{d}(n)^T, \quad (66)$$

where

$$\mathbf{d}(n) = [\mathbf{d}_1(n) \quad \cdots \quad \mathbf{d}_M(n)]^T, \quad (67)$$

$$\mathbf{d}_m(n) = [d_m(n) \quad \cdots \quad d_m(n + K - 1)]^T. \quad (68)$$

As evident from, e.g., (47), the estimated covariance needs to be invertible. This requires that

$$K \leq \frac{N + 1}{M + 1}. \quad (69)$$

where K is the number of snapshots, N is the number of samples of the signal, and M is the number of microphones. Consequently, we can only use relatively short temporal subvectors, $\mathbf{d}_m(n)$ in the estimation of the covariance matrix when the number of microphones is increased.

Algorithm 1: Flip-flop algorithm [32].

Result: Estimates of temporal and spatial covariance matrices, $\hat{\mathbf{C}}_t$ and $\hat{\mathbf{C}}_s$.

$$\mathbf{D}(n) = [\mathbf{d}_1(n) \quad \cdots \quad \mathbf{d}_M(n)];$$

$$\hat{\mathbf{C}}_s = \mathbf{I};$$

$$\hat{\mathbf{C}}_t = \frac{1}{M(N-K+1)} \sum_{n=0}^{N-K} \mathbf{D}(n) \hat{\mathbf{C}}_s^{-1} \mathbf{D}^T(n);$$

repeat

$$\left| \begin{array}{l} \hat{\mathbf{C}}_s = \frac{1}{K(N-K+1)} \sum_{n=0}^{N-K} \mathbf{D}^T(n) \hat{\mathbf{C}}_t^{-1} \mathbf{D}(n); \\ \hat{\mathbf{C}}_t = \frac{1}{M(N-K+1)} \sum_{n=0}^{N-K} \mathbf{D}(n) \hat{\mathbf{C}}_s^{-1} \mathbf{D}^T(n); \end{array} \right.$$

until convergence;

If it is assumed that the multichannel noise samples in $\mathbf{d}(n)$ follows a multichannel matrix normal distribution, the maximum likelihood (ML) estimator for the noise covariance matrix can be derived [32]. Unfortunately, the resulting estimator is not closed form, but it can be implemented using the iterative flip-flop algorithm in Algorithm 1. In some cases, e.g., if one of the covariance matrices are close to being rank deficient, this iterative procedure can be problematic, since their inverses are required. Different approaches for dealing with this and the computational complexity of the iterative procedure have been considered [31, 33]. Alternatively, a non-iterative estimator can be used such as [31]

$$\hat{\mathbf{C}}_s = \frac{1}{(N-K+1)\text{tr}(\hat{\mathbf{C}}_t)} \sum_{n=0}^{N-K} \mathbf{D}^T(n) \mathbf{D}(n), \quad (70)$$

$$\hat{\mathbf{C}}_t = \frac{1}{(N-K+1)\text{tr}(\hat{\mathbf{C}}_s)} \sum_{n=0}^{N-K} \mathbf{D}(n) \mathbf{D}^T(n), \quad (71)$$

where

$$\mathbf{D}(n) = [\mathbf{d}_1(n) \quad \mathbf{d}_2(n) \quad \cdots \quad \mathbf{d}_M(n)]. \quad (72)$$

As indicated in (70), the trace of the temporal covariance is assumed to be known. This might not be the case in practice; however, in most situations, we can simply replace it by an arbitrary value, since its main purpose is to resolve the ambiguity

$$\mathbf{C}_d = \mathbf{C}_s \otimes \mathbf{C}_t = \left(\frac{1}{\alpha} \mathbf{C}_s \right) \otimes (\alpha \mathbf{C}_t). \quad (73)$$

4.7 Non-stationary noise

While the stationarity assumption may not hold in practice, there are a number of ways to address this problem. For example, we may reduce the length, N , of the probe signal and the analysis window, which would naturally increase the validity of the assumption. Alternatively, we may decouple the prewhitening and estimation parts, as suggested in Section 4.5. In this way, we may first prewhiten our signal using a filter, and

then apply the proposed estimators with a white Gaussian noise assumption on the prewhitened signals. This approach can be exploited to take the non-stationarity of the noise into account by updating the prewhitening filters over time, according to the changing AR coefficients of the noise. Estimating non-stationary noise parameters, however, is more difficult, since the statistics need to be tracked during the presence of the desired signal, i.e., the probe signal and its reflections in our case. This problem has been well-investigated in other audio signal processing problems, such as speech enhancement [34–37].

5 Results and discussion

In this section, we investigate the performance of the different variants of the proposed EM method. More specifically, we consider the variant assuming spatially independent white Gaussian in Section 4.1 resulting in noise variance weighting (EM-UCA-NW), and its special case where the noise variance is assumed equal (EM-UCA) [18]. Moreover, we consider the setup with correlated noise proposed in Section 4.3 resulting in the prewhitening-based approach (EM-UCA-PW). The experiments were carried out using signals that were generated using the room impulse response generator [38]. The dimensions of the simulated room were set to $8 \times 6 \times 5$ m, the reverberation time (T_{60}) was set to 0.6 s while the speed of sound is fixed at 343 m/s. The loudspeaker was positioned at the center of an UCA at $(1 \times 1.5 \times 2.5)$ m while the UCA has $M = 4$ microphones with a radius of $d = 0.2$ m. Although, any type of known broadband signal could be used to probe the environment, such as a chirp signal or maximum length sequences (MLS) [39], we decided to use a white Gaussian noise sequence as the known sound source, $s(n)$, consisting of 1,500 samples from a Gaussian distribution. This sequence was subsequently zero-padded to get a total signal length of 20,000 samples. The objective of the zero-padding was to get a longer analysis window to ensure that the first few reflections are present in the observation. Moreover, as discussed in Section 4.3, the reason for using a WGN sequence is that the EM estimator can be simplified if the probe signal is an uncorrelated signal. In addition to this, using such a broadband sequence minimizes the effects of spatial aliasing [40]. The sampling frequency f_s was set to 22,050 Hz. We assumed that the direct component is subtracted from the observed signal given that we know the arrangement of the loudspeaker and the microphones. Knowing the array geometry enables either offline measurement of the impulse response of the direct-path component offline or analytical computation of the impulse response of the direct-path component based on the geometry. The background noise comprises of two components: one being

diffuse spherical noise and the other being thermal sensor noise. The diffuse spherical noise was generated using the method described in [41] using the rotor noise of a drone from the DREGON database [3]. The drone audio file used to generate the diffuse spherical noise corresponds to rotors running at 70 revolutions per second (RPS). The thermal sensor noise was simulated as spatially independent white Gaussian noise. Both these noises were added to the observed signal before estimating the parameters. The evaluation was then conducted for different signal-to-diffuse noise ratios (SDNRs) and signal-to-sensor noise ratios (SSNRs). In the following subsections, we evaluate the performance of our proposed method in various conditions.

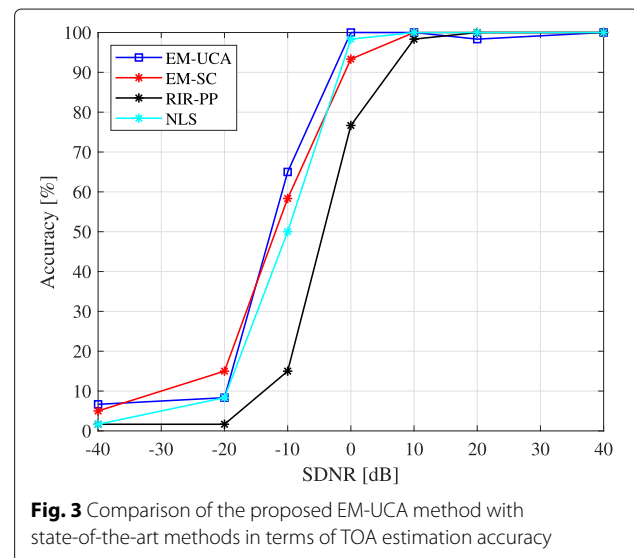
5.1 Comparison of with state-of-the-art

The aim of the first experiment was to compare the proposed method with existing state-of-the-art methods. The EM algorithm was set to estimate $R = 3$ reflections with 40 iterations and β was set to $\frac{1}{R}$. The main application for this manuscript is acoustic reflector mapping for robot audition. For this application, the mapping should be possible in unknown, complex environments, and we therefore do not rely on trivial room geometry models as opposed to many of the traditional methods for room geometry estimation [10–12]. Therefore, we chose to use a small number of reflections in the estimation (i.e., $R = 3$), to mainly estimate the TOAs/DOAs of first-order reflections impinging from nearby acoustic reflectors. These can be directly mapped to acoustic reflector positions based on the estimated time and angle of arrival. While this will not facilitate the localization of all acoustic reflectors at any given time instance, we can carry out such estimation over time and space, to generate a map of an arbitrary room geometry (see Section 5.4). An alternative to choosing a fixed reflection order would be to combine the proposed method with order estimation methods [42, 43]. To initialize the method, the gain estimates, $\hat{g}_{m,r}$, were sampled from a uniform distribution over the interval $[0; 1]$, the TOAs, $\hat{\tau}_{1,r}$, were sampled from a uniform discrete distribution over the time indices corresponding to the analysis window, and the DOAs, $\hat{\phi}_r$, were sampled from a uniform distribution over the interval $[0^\circ; 360^\circ]$. After emitting and recording the known source signal, an analysis window of each recording was considered starting from τ_{\min} samples to τ_{\max} samples after the source signal was emitted. In this experiment, the analysis window was set such that the search is made between 0.5 to 2 m. This was done to primarily capture the first order reflections. The lower bound was chosen because we can only search for reflectors that are outside the geometry of the array, which, in our experiments, had a radius of 0.2 m. After 2 m, the performance of the proposed method degrades because the energy of

the reflected signals decrease quadratically over distance, which motivated the choice of the upper limit.

The proposed EM method (EM-UCA) was compared to the single-channel EM method (EM-SC) in [22] in terms of TOA accuracy, applied to the first microphone. Moreover, these were compared with a common approach to extracting TOAs from estimated RIR through peak-picking (RIR-PP). Finally, the performance was also compared with our previous work [44] termed the non-linear least squares estimator (NLS). The results for the TOA estimation are shown in Fig. 3, where the accuracy was defined as the percentage of TOA estimates that were within $\pm 2\%$ tolerance of one of the true parameters of the first-order reflections computed using the image-source method. This was measured for different SDNRs while the SSNR was fixed to 10 dB, and for each SDNR, the accuracy was measured over 100 Monte-Carlo simulations. As seen in Fig. 3, the proposed method clearly outperforms the existing method by providing higher accuracy at lower SDNRs.

Furthermore, the computation time of the RIR-PP and the proposed method, EM-UCA, were measured. This test was performed in MATLAB using the built-in function *timeit* on a standard desktop computer running a Microsoft Windows 10 operating system with an Intel Core i7 CPU with 3.40 GHz processing speed and 16 GB of RAM. A Monte Carlo simulation with 100 trials was performed on each method and an average time was calculated. The measured computation times of the RIR-PP and the EM-UCA were 0.0063 s and 25.74 s, respectively, for $R = 1$ and an SDNR of 40 dB. This shows that the improved estimation accuracy with the proposed method comes at the cost of a higher computational complexity. It is important to stress, however, that in applications such as acoustic reflector localization with a drone, it is



common to have negative SNR conditions [45], where the RIR-PP method may fail to provide accurate estimates as opposed to the proposed method (see, e.g., Fig. 3). Moreover, the computational cost could be reduced further by, e.g., employing the recursive EM approach [29, 46]. If the TOA/DOA estimation is carried out continuously over time and space, the EM algorithm may be initialized using previous estimates, which may significantly reduce the number of iterations needed for convergence. Another potential computational saving may be obtained by deriving the proposed methods in the frequency domain.

5.2 Evaluation for different diffuse noise conditions

In the second experiment, we evaluated the effect of the proposed prewhitening approach under different diffuse noise conditions. To test the performance of the EM algorithm under such realistic scenarios, we test our estimator for different SDNRs in the interval $[-40; 10]$ dB while setting the SSNR to 40 dB. Here, we are comparing the EM algorithm with and without the prewhitening in terms of both TOA and DOA estimation accuracy as seen in Figs. 4 and 5, respectively. The diffuse rotor noise is indeed correlated with strong periodic components, but the results show that the proposed prewhitening approach can successfully account for this and can retain a high estimation accuracy at SDNRs levels 20 dB lower than those needed for the EM-UCA approach.

5.3 Evaluation for faulty/noisy microphone conditions

In this experiment, we consider a scenario where one microphone is excessively noisy compared to the other microphones. An example of this could be a robot platform, where one microphone is placed closer to an ego-noise source such as a fan, leading to TOA and DOA estimation errors. To simulate this effect, we set thermal

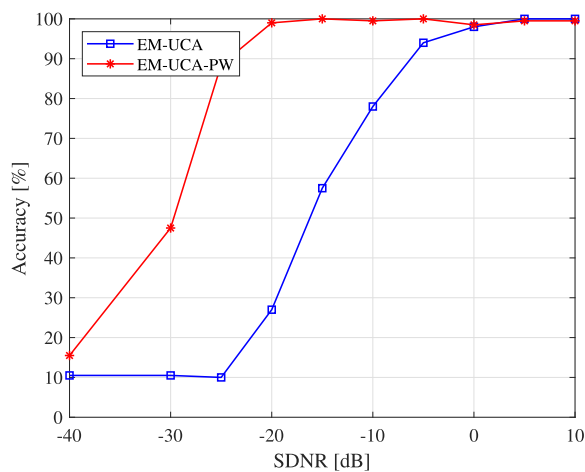


Fig. 4 TOA estimation accuracy of the proposed EM method with and without prewhitening

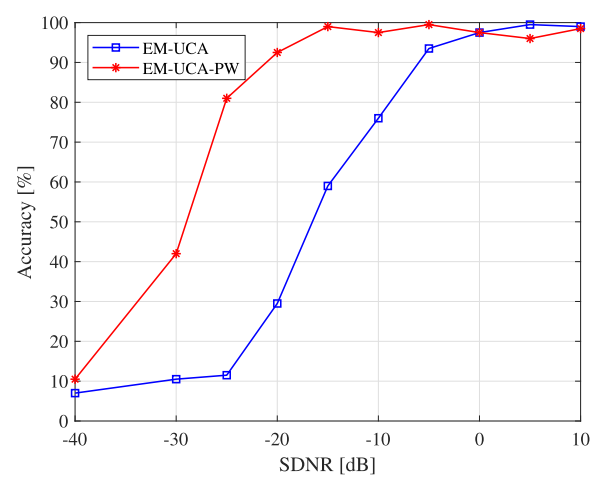


Fig. 5 DOA estimation accuracy of the proposed EM method with and without prewhitening

noise of a single microphone to an SSNR level of -10 dB, while the thermal noise of the remaining microphones are set to an SSNR level of 40 dB. As seen in Figs. 6 and 7, the performance of the EM algorithm with noise variance weighting is less affected by the high thermal sensor noise in terms of both TOA and DOA estimation accuracy. Moreover, we conducted an experiment without diffuse noise, where the SSNR level of the faulty microphone was changed from -40 to 0 dB. These results are shown in Figs. 8 and 9, and show that the estimation accuracy is already degrading from 0 dB SSNR and downwards when using the EM-UCA approach, whereas the proposed EM-UCA-NW approach retains a high accuracy.

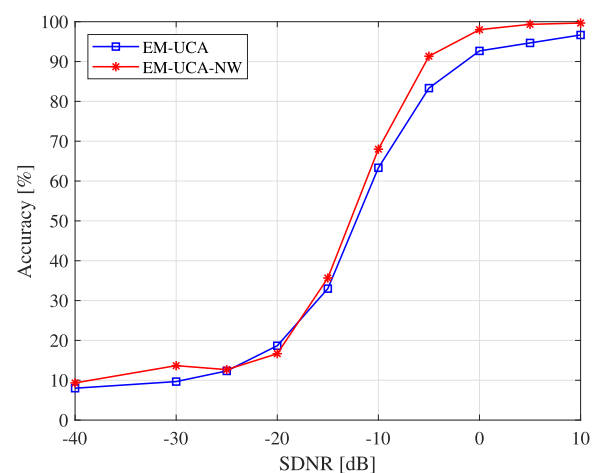
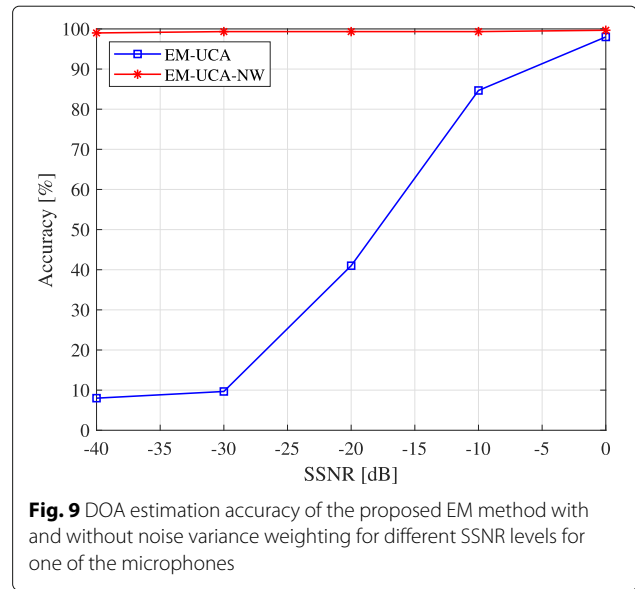
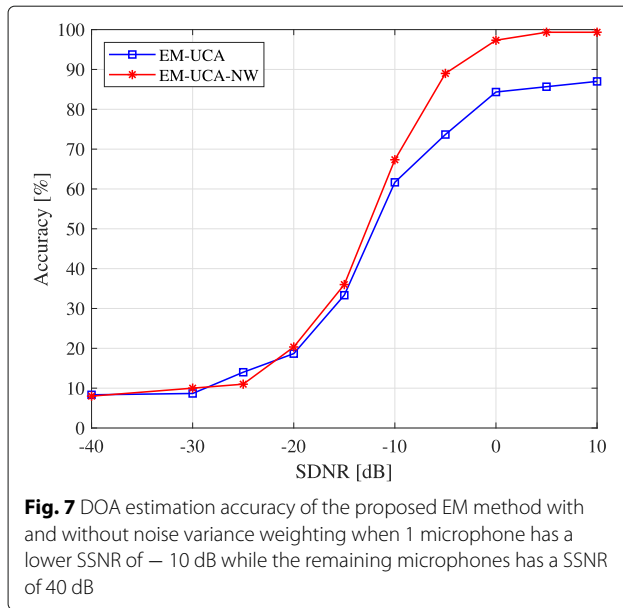


Fig. 6 TOA estimation accuracy of the proposed EM method with and without noise variance weighting when 1 microphone has a lower SSNR of -10 dB while the remaining microphones has a SSNR of 40 dB



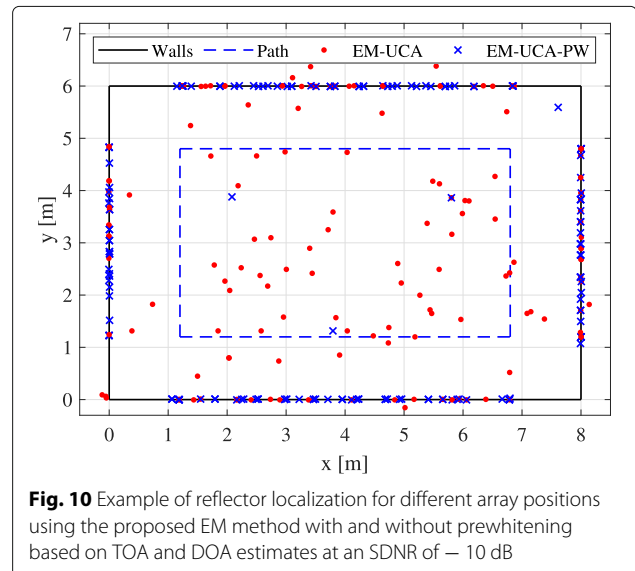
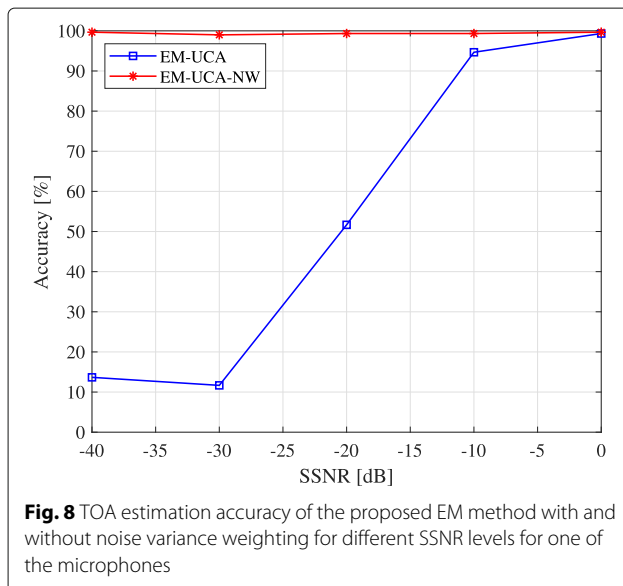
5.4 Application example of the proposed method

We consider an application example where the localization of the acoustic reflectors is done using the proposed EM method with and without prewhitening. More specifically, we have used filter-based prewhitening approach as discussed in Section 4.5. This experiment thus shows how the proposed method can be used to map an environment using a moving robot platform. The room parameters were kept the same as the earlier experiment. Furthermore, the SDNR was set to -10 dB corresponding to a strong ego-noise. The loudspeaker-microphone arrangement was similar to the previous experiments and follows the a predefined path as shown in Fig. 10 indicated by the

blue dashed line. As depicted in the figure, the EM algorithm with prewhitening performs better at estimating acoustic reflector using the estimated TOAs and DOAs, compared to EM algorithm without prewhitening.

6 Conclusion

In this paper, we consider the problem of estimating the time- and direction-of-arrivals of acoustic echoes using a loudspeaker emitting a known source signal and multiple microphones. Among other examples, this is an important problem in robot and drone audition, where these parameters can reveal the positions of nearby acoustic reflectors and thus facilitate mapping and navigation of a physical environment. Some methods exist for solving



the problems of acoustic reflector localization and room geometry estimation; however, most of these rely on a priori information, e.g., of the TOAs or DOAs of the acoustic echoes. However, estimating these is a difficult problem on its own, which is dealt with by the methods proposed herein. Moreover, even when the TOAs are estimated for some of the traditional approaches, the difficult problem of echolabeling needs to be solved, since the order of the corresponding reflection is generally unknown. We therefore propose different methods for estimating, not only the TOAs, but also the DOAs of acoustic echoes. By estimating the DOAs also, it is possible to resolve some of the ambiguity introduced by knowing only the TOAs. The proposed method is based on the expectation-maximization framework and are derived to be optimal under different conditions ranging from the simple white Gaussian noise scenario to scenarios with correlated and colored noise. In the experiments, we show that proposed methods are able to estimate the TOAs and DOAs with higher accuracy and noise robustness compared to existing methods. Moreover, we show that some of the proposed variants can account for colored noise and scenarios where a microphone is faulty or more noisy than the other microphones of the array. Finally, we conducted a more applied experiment, where it is illustrated how a room can be mapped from the estimated parameters, which is relevant to, e.g., autonomous robot and drone applications. While the proposed method has a higher computation time than traditional methods, this can be reduced significantly by adopting the recursive EM scheme and deriving the proposed methods in the frequency domain.

Abbreviations

TOA: Time-of-arrival; EM: Expectation-maximization; UCA: Uniform circular array; SNR: Signal-to-noise ratio; DOA: Direction-of-arrival; aSLAM: Acoustic simultaneous localization and mapping; RIR: Room impulse response; TDOA: Time difference-of-arrival; ML: Maximum likelihood; MVDR: Minimum variance distortionless response; LCMV: Linearly constrained minimum variance; SAGE: Space alternating generalized expectation; FIR: Finite impulse response; AR: Autoregressive; EM-UCA: Proposed method without prewhitening or noise weighting; EM-UCA-NW: Proposed method with only noise weighting; EM-UCA-PW: Proposed method with only prewhitening; T_{60} : Reverberation time (60 dB); RPM: Revolutions per minute; DREGON: Database of drone audio recordings; SDNR: Signal-to-diffuse-noise ratio; SSNR: Signal-to-sensor-noise ratio; EM-SC: Single channel EM method; RIR-PP: RIR-based method with peak picking; NLS: nonlinear least squares

Acknowledgements

Not applicable.

Authors' contributions

JRJ and SG designed the idea for the manuscript. JRJ and US conducted the experiments. All the authors contributed to the writing of this work. Moreover, all author(s) read and approved the final manuscript.

Funding

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

Availability of data and materials

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Audio Analysis Lab, CREATE, Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark. ²Bar-Ilan University, 5290002 Ramat-Gan, Israel.

Received: 13 April 2020 Accepted: 15 July 2020

Published online: 08 August 2020

References

1. C. Rascon, I. Meza, Localization of sound sources in robotics: a review. *Robot. Auton. Syst.* **96**, 184–210 (2017)
2. H. W. Lollmann, A. Moore, P. A. Naylor, B. Rafaely, R. Horaud, A. Mazel, W. Kellermann, in *Hands-free Speech Comm. and Microphone Arrays*. Microphone array signal processing for robot audition, (2017), pp. 51–55
3. M. Strauss, P. Mordel, V. Miquet, A. Deleforge, in *IEEE/RJS Int. Conf. Intelligent Robots and Systems*. DREGON: dataset and methods for UAV-embedded sound source localization, (2018), pp. 5735–5742
4. F. Badeig, Q. Pelorson, S. Arias, V. Drouard, I. D. Gebru, X. Li, G. Evangelidis, R. Horaud, in *Int. Conf. Multimodal Interaction*. A distributed architecture for interacting with NAO, (2015)
5. F. Antonacci, J. Filos, M. R. P. Thomas, E. A. P. Habets, A. Sarti, P. A. Naylor, S. Tubaro, Inference of room geometry from acoustic impulse responses. *IEEE Trans. Audio Speech Lang. Process.* **20**(10), 2683–2695 (2012)
6. M. Coutino, M. B. Møller, J. K. Nielsen, R. Heusdens, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Greedy alternative for room geometry estimation from acoustic echoes: a subspace-based method, (2017), pp. 366–370
7. J.-S. Hu, C.-Y. Chan, C.-K. Wang, M.-T. Lee, C.-Y. Kuo, Simultaneous localization of a mobile robot and multiple sound sources using a microphone array. *Adv. Robot.* **25**(1–2), 135–152 (2011)
8. S. Ogiso, T. Kawagishi, K. Mizutani, N. Wakatsuki, K. Zempo, Self-localization method for mobile robot using acoustic beacons. *ROBOMECH J.* **2**(1), 12 (2015)
9. C. Evers, P. A. Naylor, Acoustic SLAM. *IEEE/ACM Trans. Audio Speech Lang. Process.* **26**, 1484–1498 (2018)
10. M. Kreković, I. Dokmanić, M. Vetterli, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* EchoSLAM: simultaneous localization and mapping with acoustic echoes, (2016), pp. 11–15
11. L. Nguyen, J. V. Miro, X. Qiu, in *IEEE/RSJ Int. Conf. Intell. Robots and Syst.* Can a robot hear the shape and dimensions of a room? (2019), pp. 5346–5351
12. T. Wang, F. Peng, B. Chen, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* First order echo based room shape recovery using a single mobile device, (2016), pp. 5346–5351
13. I. J. Kelly, F. M. Boland, Detecting arrivals in room impulse responses with dynamic time warping. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(7), 1139–1147 (2014)
14. M. D. Plumbley, Hearing the shape of a room. *Proc. Natl. Acad. Sci. U S A.* **110**(30), 12162–12163 (2013)
15. L. B. Nelson, H. V. Poor, Iterative multiuser receivers for CDMA channels: an EM-based approach. *IEEE Trans. Commun.* **44**(12), 1700–1710 (1996)
16. M. C. Vanderveen, C. B. Papadakis, A. Paulraj, Joint angle and delay estimation (JADE) for multipath signals arriving at an antenna array. *IEEE Commun. Lett.* **1**(1), 12–14 (1997)
17. J. Verhaevert, E. V. Lil, A. V. de Capelle, Direction of arrival (DOA) parameter estimation with the SAGE algorithm. *Signal Process.* **84**(3), 619–629 (2004)
18. J. R. Jensen, U. Saqib, S. Gannot, in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.* An EM method for multichannel TOA and DOA estimation of acoustic echoes, (2019)
19. S. Braun, A. Kuklasinski, O. Schwart, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, J. Jensen, Evaluation and comparison of late reverberation power spectral density estimators. *IEEE/ACM Trans. Audio Speech Lang. Process.* **26**(6), 1056–1071 (2018)
20. B. F. Cron, C. H. Sherman, Spatial-correlation functions for various noise models. *J. Acoust. Soc. Am.* **34**(11), 1732–1736 (1962)
21. H. Sun, T. D. Abhayapala, P. N. Samarasinghe, in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.* Active noise control over 3D space with multiple circular arrays, (2019), pp. 135–139
22. M. Feder, E. Weinstein, Parameter estimation of superimposed signals using the EM algorithm. *IEEE Trans. Acoust. Speech Signal Process.* **36**(4), 477–489 (1988)

23. O. Schwartz, S. Gannot, E. A. P. Habets, Multispeaker LCMV beamformer and postfilter for source separation and noise reduction. *IEEE/ACM Trans. Audio Speech Lang. Process.* **25**(5), 940–951 (2017)
24. R. Balan, J. Rosca, in *Proc. IEEE Workshop Sensor Array and Multichannel Signal Process.* Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase, (2002), pp. 209–213
25. L. L. Scharf, *Statistical signal processing: detection, estimation, and time series analysis*. (Addison-Wesley Publishing Company, Michigan, 1991)
26. P. C. Hansen, S. H. Jensen, Prewhitening for rank-deficient noise in subspace methods for noise reduction. *IEEE Trans. Signal Process.* **53**(10), 3718–3726 (2005)
27. G. Reinsel, Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *J. Am. Statist. Assoc.* **77**(377), 190–195 (1982)
28. J. A. Fessler, A. O. Hero, Space-alternating generalized expectation-maximization algorithm. *IEEE Trans. Signal Process.* **42**(10), 2664–2677 (1994)
29. O. Schwartz, S. Gannot, Speaker tracking using recursive EM algorithms. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(2), 392–402 (2014)
30. S. M. Nørholm, J. R. Jensen, M. G. Christensen, Instantaneous fundamental frequency estimation with optimal segmentation for nonstationary voiced speech. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(12), 2354–2367 (2016)
31. M. H. Castaneda, J. A. Nossék, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Estimation of rank deficient covariance matrices with Kronecker structure, (2014), pp. 394–398
32. P. Dutilleul, The MLE algorithm for the matrix normal distribution. *J. Statist. Comput. Simul.* **64**(2), 105–123 (1999)
33. K. Werner, M. Jansson, P. Stoica, On estimation of covariance matrices with Kronecker product structure. *IEEE Trans. Signal Process.* **56**(2), 478–491 (2008)
34. R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.* **9**(5), 504–512 (2001)
35. T. Gerkmann, R. C. Hendriks, Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. *IEEE Trans. Audio Speech Lang. Process.* **20**(4), 1383–1393 (2012)
36. R. C. Hendriks, R. Heusdens, J. Jensen, in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* MMSE based noise PSD tracking with low complexity, (2010), pp. 4266–4269
37. J. K. Nielsen, M. S. Kavalekalam, M. G. Christensen, J. Boldt, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Model-based noise PSD estimation from speech in non-stationary noise, (2018), pp. 5424–5428
38. E. A. P. Habets, Room impulse response generator. Technical report, Technische Universiteit Eindhoven (2010). Ver. 2.0.20100920. <https://github.com/ehabets/RIR-Generator>
39. D. Florencio, Z. Zhang, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Maximum a posteriori estimation of room impulse responses, (2015), pp. 728–732
40. J. Dmochowski, J. Benesty, S. Affes, On spatial aliasing in microphone arrays. *IEEE Trans. Signal Process.* **57**(4), 1383–1395 (2009)
41. E. A. P. Habets, I. Cohen, S. Gannot, Generating nonstationary multisensor signals under a spatial coherence constraint. *J. Acoust. Soc. Am.* **124**(5), 2911–2917 (2008)
42. K. Han, A. Nehorai, Improved source number detection and direction estimation with nested arrays and ULAs using jackknifing. *IEEE Trans. Signal Process.* **61**(23), 6118–6128 (2013)
43. P. Stoica, Y. Selen, Model-order selection: a review of information criterion rules. *IEEE Signal Process. Mag.* **21**(4), 36–47 (2004)
44. U. Saqib, J. R. Jensen, in *Proc. European Signal Processing Conf.* Sound-based distance estimation for indoor navigation in the presence of ego noise, (2019), pp. 1–5
45. A. Deleforge, D. Di Carlo, M. Strauss, R. Serizel, L. Marcenaro, Audio-based search and rescue with a drone: highlights from the IEEE signal processing cup 2019 student competition [SP competitions]. *IEEE Signal Process. Mag.* **36**(5), 138–144 (2019)
46. K. Weisberg, S. Gannot, O. Schwartz, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* An online multiple-speaker DOA tracking using the Cappé-Moulines recursive expectation-maximization algorithm, (2019), pp. 656–660

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)