Author(s): Sakari Tervo, Archontis Politis
Title: Direction of Arrival Estimation of Reflections from Room Impulse Responses Using a Spherical Microphone Array
Year: 2015
Version: Post print

Please cite the original version:
Sakari Tervo, Archontis Politis. Direction of Arrival Estimation of Reflections from Room Impulse Responses Using a Spherical Microphone Array. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23, 10, 1539-1551, October 2015. DOI: 10.1109/TASLP.2015.2439573

Rights: © 2015 IEEE. Reprinted with permission. In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Aalto University's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

This publication is included in the electronic version of the article dissertation: Politis, Archontis. Microphone array processing for parametric spatial audio techniques. Aalto University publication series DOCTORAL DISSERTATIONS, 195/2016.
Direction of Arrival Estimation of Reflections from Room Impulse Responses using a Spherical Microphone Array

Sakari Tervo and Archontis Politis.

Abstract—This paper studies the direction of arrival estimation of reflections in short time windows of room impulse responses measured with a spherical microphone array. Spectral-based methods, such as multiple signal classification (MUSIC) and beamforming, are commonly used in the analysis of spatial room impulse responses. However, the room acoustic reflections are highly correlated or even coherent in a single analysis window and this imposes limitations on the use of spectral-based methods. Here, we apply maximum likelihood (ML) methods, which are suitable for direction of arrival estimation of coherent reflections. These methods have been earlier developed in the linear space domain and here we present the ML methods in the context of spherical microphone array processing and room impulse responses. Experiments are conducted with simulated and real data using the em32 Eigenmike. The results show that direction estimation with ML methods is more robust against noise and less biased than MUSIC or beamforming.

Index Terms—Spatial Room Impulse Response, Room Acoustics, Spherical Microphone Arrays, Direction of Arrival

I. INTRODUCTION

DIRECTION of arrival (DOA) estimation of a sound wave arriving at a microphone array is an essential part of spatial room acoustic analysis and synthesis. The directional information is used together with pressure or energy to describe the sound field [1—4] or to reproduce a sound in spatial sound synthesis from a certain direction [5, 6]. Thus, it has a profound impact on how room acoustics are interpreted via the analysis or perceived through the spatial sound synthesis. The increasing number of publications and numerous array designs indicate the importance of the spherical microphone array processing [7—10]. Therefore, it can be considered as one of the most important approaches for spatial sound analysis nowadays. This paper studies the DOA estimation of reflections from a spatial room impulse response captured with a spherical microphone array. In particular, we focus on the case where more than one acoustic reflection is present in an analysis window.

In room impulse responses, the number of reflections arriving at the receiver in a short time window increases with the square of time. This effect is given as the echo density (11, p. 98). Therefore, after a relatively short time, one ends up in a situation where an analysis time window includes multiple reflections. Exact overlap of the reflections occurs if the path length from the source to the receiver is equal for two or more reflections. This is not an unusual case in room acoustics, but happens already for the first order reflections in symmetric source-receiver geometry. For example, such an overlap will occur if the microphone and the sound source are located somewhere on the same diagonal or central axis of a rectangular room.

The acoustic reflections are highly correlated or even coherent, especially in a narrow frequency band. The high correlation of the reflections causes problems for spectral-based DOA estimation methods [12], which are commonly applied in the spherical microphone array processing [2, 3, 13—17]. According to a classification given in [12], the spectral-based methods include multiple signal classification (MUSIC) method, the estimation of signal parameters via rotational invariant techniques (ESPRIT), and beamforming. These methods require that the source signals are independent, which is not true in the case of highly correlated reflections.

The estimation in the case of correlated signals has been enhanced by smoothing methods in the space domain in the previous decades [18, 19]. These techniques have been especially under research with uniform linear arrays [20—24]. Later on, several smoothing methods have also been developed in the context of spherical microphone array processing [2, 3, 16, 17]. These smoothing methods average the array covariance matrix over frequency [2], time [3], or space [16, 17] and require pre-processing before DOA estimation. In frequency smoothing the noise is whitened [2] and in time domain smoothing a stabilizing filter reduces the undesired amplification of noise. Spatial smoothing in the space domain is often implemented by averaging over subarrays formed from the original uniform linear array [12]. In spherical microphone array processing, the division into subarrays is obtained by transforming the spherical microphone array to a uniform linear array [17, 25]. Another approach to obtain spatial smoothing is to form the subarrays via eigenbeam space rotation [15, 16].

The smoothing methods have an apparent disadvantage when applied to room impulse responses. Namely, in each time step and frequency, the room impulse response may have a different response, i.e., the DOA, phase, and amplitude. Averaging over any domain reduces resolution in the respective domain. Averaging over time reduces temporal resolution of the estimate; averaging over frequency reduces frequency resolution; and averaging over space reduces spatial resolution.
since the number of microphones per subarray is lower than the number of microphones in the whole array.

Contrary to MUSIC and similar approaches, the maximum likelihood (ML) methods [12, 26] can handle coherent signals without any smoothing. This is possible due to the freedom of selection in the signal and noise model. Consequently, also a coherent signal model can be assumed. This signal model is $Q$-dimensional, where $Q$ is the number of reflections, since each reflection is modeled separately. If a spatial spectrum is evaluated with a grid of size $D$, where the appropriate sampling and size is defined by the user, the ML DOA estimation function has a dimensionality of $Q \times D$, whereas the spectral-based methods have a dimensionality of $1 \times D$. Therefore, as the number of reflections increases, the search space becomes quickly very large. For restricted computational time, the high dimensionality leads to the use of non-linear optimization algorithms for the ML methods [26]. As a general limitation in the ML methods, the number of reflections $Q$ that can be estimated must be smaller than the number of microphones $I$, i.e., $I > Q$ [12].

Analysis of room reflections is often based on the wideband assumption [2, 3, 6]. This assumption states that frequencies are delayed by the same amount of time. In theory, the wideband assumption holds if the surfaces are large and rigid. The wideband analysis of acoustic reflections can lead to a desired accuracy in the analysis [2, 3] or in the reproduction of the acoustics [6]. However, in the real world, the room impulse responses are always frequency and time dependent. On this basis, studies on acoustics benefit from the frequency band analysis of reflections since it enables a more accurate description of the room acoustic properties.

In this paper, we study the performance of ML methods and a large sample approximation called the weighted subspace fitting (WSF) in the analysis of spatial room impulse responses. Large sample approximation assumes that the number of available measurements is large. The cases where one or more reflections are present in an analysis window in a wide or narrow frequency band are of interest due to the above mentioned features of the room impulse responses. The contribution of this paper is the application of the ML methods to the analysis of room reflections with the spherical microphone array. Throughout the experimental section, the results of the ML methods are compared to MUSIC and beamforming.

II. MODELS

A. Room impulse response

A room impulse response is defined as the acoustic response, measured from a source to a microphone $i$ in an enclosed space. After the initial excitation, the sound wave propagates through the space and arrives at the receiver via multiple paths. On these paths, the sound wave is altered by several acoustic phenomena on the boundaries, such as, reflection, absorption, and diffraction [11, ch. 2]. These acoustic phenomena affect the amplitude and the phase of the traveling sound wave, possibly differently, in each frequency and at each incident angle.

Besides the boundary effects, the sound wave is affected by the propagation distance and attenuation by medium. Namely, the air absorption depends on the composition of the room air, the distance that the wave has traveled, and frequency [11, p. 147]. Thus, the air absorption alone suggests that the room impulse response is different in each time and frequency and therefore gives basis for the frequency dependent analysis. Moreover, if the sound source is assumed to be a point source the amplitude is attenuated due to the spherical spreading of the sound wave.

The number of reflections $Q$ per time interval $\Delta t$ is described by a quantity called the echo density, which is given asymptotically by [11, p. 98]

$$\frac{Q}{\Delta t} = \frac{4\pi c^3 t^2}{V}$$

where $V$ is the volume of the enclosure and $c$ is the speed of sound. According to Kuttruff [11, p. 98], this is valid for any geometry with a homogeneous medium. Shorter time interval $\Delta t$ reduces the number of reflections present in an analysis window.

When we are inspecting the impulse response in a single frequency, we limit the time window length as $\Delta t > 2\pi/\omega$, so that at least one period of the wave length is observed in the analysis window. Then, the time instant $t_Q$ where we have $Q$ or fewer reflections present in the analysis window in a single frequency is given as:

$$t_Q = t(\omega) \leq \sqrt{\frac{QV\omega}{8\pi^2 c^3}},$$

where $\omega$ is the angular frequency.

B. Space Domain

We present the location of microphone $i$ in the 3-D Cartesian coordinate system as

$$\mathbf{r}_i = r_i [\sin(\theta_i) \cos(\phi_i), \sin(\theta_i) \sin(\phi_i), \cos(\theta_i)]^T,$$

where $r_i$ and $\Phi_i = (\theta_i, \phi_i)$ are the standard spherical coordinates, radius $r_i \in [0, \infty)$, inclination $\theta_i \in [0, 180]^{\circ}$ and azimuth $\phi_i \in [-180, 180]^{\circ}$. Each acoustic event $q$ arriving to the microphone has traveled a path length $d_q$ from the source to the microphone and has a DOA $\Omega_q = (\theta_q, \phi_q)$ w.r.t to the array origin and it is described in Cartesian coordinates as:

$$\mathbf{r}_q = d_q [\sin(\theta_q) \cos(\phi_q), \sin(\theta_q) \sin(\phi_q), \cos(\phi_q)]^T.$$

An acoustic event is considered to be a sound wave which is altered by the acoustic phenomena listed above.

A measured wideband impulse response pressure signal $p_i$ in a microphone $i$ for a source signal $s$ is presented in the frequency domain as the sum of all acoustic events $q = 1, \ldots, Q$

$$p_i(k) = \sum_{q=1}^{Q} \hat{h}_{iq}(k) s_q(k) + n_i(k) \in \mathbb{C},$$

where $\hat{h}_{iq}(k)$ is the frequency response of an acoustic event, and $n_i(k)$ is a noise component, assumed independent and identically distributed for each microphone. Furthermore,
\[ k = \omega/c = 2\pi f/c \] is the wavenumber, where \( c \) is the speed of sound and \( f \) is frequency. In addition, \( \hat{s}_q(k) \) is the source signal that describes the frequency response of the source, originally emitted to some direction, and arriving to the microphone from the direction \( \mathbf{r}_q \). A typical source in room impulse response measurements is a loudspeaker. Note that this impulse response model is general and describes any room impulse response.

The frequency response of the acoustic event is dependent on the time delay \( \tau_q(k) \), that describes the time it has taken for the acoustic wave to travel to the microphone, and on the complex amplitude of each acoustic event \( \hat{c}_q \), i.e.,

\[
\hat{h}_q(k) = c_q(k) \exp(-i\tau_q(k)) \tag{4}
\]

where \( i = \sqrt{-1} \).

We further assume that the sources and reflections are in the far field with respect to the array, so that a plane wave model can be applied. Then, according to the plane wave model the complex amplitude is represented by

\[
c_q(k) = \hat{c}_q(k) \hat{s}_q(k) \quad \forall i,
\tag{5}
\]

where \( \hat{c}_q \) is the complex amplitude of plane wave \( q \) which includes all the acoustic phenomena that the wave has encountered before arriving to the microphone and \( \hat{s}_q \) is the response of the microphone \( i \) in the direction \( \Omega_q \) assumed to be known a priori. Furthermore, since a homogeneous medium and short time window analysis are assumed, the path length \( d_q \) is neglected in the directional analysis. Consequently, the plane wave time delay can be expressed w.r.t. to the origin, i.e., the center of the array, as

\[
\tau_q(k) = kr_q^T \mathbf{m}_i,
\tag{6}
\]

where \((\cdot)^T\) denotes the transpose of a vector or a matrix and \( r_q = [\sin(\theta_q) \cos(\phi_q), \sin(\theta_q) \sin(\phi_q), \cos(\phi_q)]^T \) is the normal to the wavefront.

We present the delay term and the microphone array response with

\[
h_q(k) = \hat{c}_q(k) \exp\{-ikr_q^T \mathbf{m}_i\},
\tag{7}
\]

and call this the steering vector. For example, for an ideal open microphone array \( \hat{c}_q = 1 \). The plane wave amplitude and the source response are presented as a product

\[
s_q(k) = \hat{c}_q(k) \hat{s}_q(k),
\tag{8}
\]

which we call the reflection signal. For the compactness of the rest of the paper, the measurements are described in a matrix form. The impulse responses in microphones \( i = 1, \ldots, I \), i.e., the array input, are described by the vector \( \mathbf{p} \)

\[
\mathbf{p}(k) = \left[ \begin{array}{c} p_1(k) \\ \vdots \\ p_I(k) \end{array} \right] = \left[ \begin{array}{c} h_1, \ldots, h_Q \end{array} \right] \left[ \begin{array}{c} \hat{s}_q(k) + n(k) \end{array} \right] = \mathbf{H}(\Omega) \mathbf{s}(k) + \mathbf{n}(k)
\tag{9}
\]

where \( \Omega = [\Omega_1, \ldots, \Omega_Q] \) is the set of all unknown direction of arrivals.

The noise and reflection signals are assumed to be zero mean complex Gaussian processes, i.e., \[26\]

\[
E\{s(k)\mathbf{s}(k)^H\} = R_s(k) \quad \text{and} \quad E\{n(k)\mathbf{n}(k)^H\} = \sigma^2 \mathbf{I},
\]

where \( E\{\cdot\} \) denotes expectation and \((\cdot)^H\) denotes a Hermitian transpose of a matrix. This leads to an array covariance matrix \[26\]:

\[
\mathbf{R}_{pp}(k) = E\{\mathbf{p}(k)\mathbf{p}(k)^H\} = \mathbf{H}(\Omega) R_{ss}(k) \mathbf{H}(\Omega)^H + \sigma^2 \mathbf{I},
\tag{10}
\]

The assumption on the Gaussian reflection signal is not necessarily true in the case of room impulse responses. That is, the reflection at a single frequency in a small time window may be of deterministic nature rather than random. A model for deterministic array covariance matrix is given in Section III.E.

C. Spherical harmonics domain

In order to apply spherical microphone array processing, the pressure and the array covariance matrix are described in the spherical harmonic (SH) domain. The formulation in this section follows the one given in \[3\].

The SH domain representation of the pressure \( p(k, r, \Omega) \) for an array with radius \( r \) and order \( N_h \) is given by the approximation of the spherical Fourier Transform and its inverse \[3\]:

\[
p_{nm}(k, r, \Omega) = \sum_{i=1}^I \alpha_i p(k, r, \Phi_i) |Y_n^m(\Phi_i)|^* \quad \text{and} \quad \tag{11}
\]

\[
p(k, r, \Omega) = \sum_{n=0}^{N_h} \sum_{m=-n}^{n} p_{nm}(k, r) Y_n^m(\Omega),
\tag{12}
\]

respectively, where \( Y_n^m(\cdot) \) is the spherical harmonic of order \( n \) and degree \( m \) \[9\]:

\[
Y_n^m(\Omega) = \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\cos \theta) e^{im\phi},
\tag{13}
\]

and \( P_n^m \) are the associated Legendre polynomials. Moreover, \((\cdot)^*\) denotes the complex conjugate, \( \alpha_i \) are the sampling weights to correct the orthonormality errors, \( \Phi_i \) are the microphone coordinates and \( \Omega = (\theta, \phi) \) is a steering direction where the inverse spherical Fourier transform is evaluated. The sampling weights are defined by the sampling scheme \[9\], and the order is defined by the number of microphones (see \[9\] for details). Throughout this paper we assume uniform sampling and hence the weights reduce to \( \alpha_i = 4\pi/I \). In addition, we use \((N_h + 1)^2 = 16\) harmonic coefficients, where the harmonic order of the array is \( N_h = 3 \), and the radius is \( r = 4.2 \text{ cm} \), due to the applied microphone array. Spatial aliasing for spherical arrays typically occurs when \( kr > N_h \) \[9\].

The noiseless pressure in the SH domain can be expressed in a matrix form by \[3\]

\[
p_{nm}(k, r) = Y_n^m(\Omega) \Gamma Y(\Phi) \mathbf{B}(kr) \mathbf{Y}(\Omega) s(k)
\tag{14}
\]
where \( \Phi = [\Phi_1, \ldots, \Phi_f] \) are the sensor positions in angular coordinates, \( \Gamma = \text{diag}\{\alpha_1, \alpha_2, \ldots, \alpha_f\} \) are the sampling weights, and \( Y^H(\Phi)IY(\Phi) = I \). The left hand side of Eq. (14) are the SH coefficients in a matrix form:

\[
P_{nm} = [p_{0,0}, p_{1,-1}, p_{1,0}, p_{1,1}, \ldots, p_{N_h,N_h}]^T, \tag{15}
\]

where \( k \) and \( r \) are not used for brevity and the spherical harmonics are expressed in a matrix form

\[
Y(\Omega) = \begin{bmatrix}
Y_0^0(\Omega_1), & Y_0^0(\Omega_2) & \cdots & Y_0^0(\Omega_q)
\end{bmatrix}^T
\]

\[
\begin{bmatrix}
Y_1^0(\Omega_1), & Y_1^{-1}(\Omega_2) & \cdots & Y_1^{-1}(\Omega_q)
\end{bmatrix}
\]

\[
\begin{bmatrix}
Y_1^{1}(\Omega_1), & Y_1^0(\Omega_2) & \cdots & Y_1^0(\Omega_q)
\end{bmatrix}
\]

\[
\begin{bmatrix}
\vdots & \vdots & \ddots & \vdots
\end{bmatrix}
\]

\[
\begin{bmatrix}
Y_{N_h}^{-N_h}(\Omega_1), & Y_{N_h}^{-N_h}(\Omega_2) & \cdots & Y_{N_h}^{-N_h}(\Omega_q)
\end{bmatrix}
\]

The array dependent coefficients are represented by a diagonal matrix:

\[
B(kr) = \text{diag}\{b_0(kr), b_1(kr), b_1(kr), b_1(kr), \ldots, b_N(kr)\} \tag{17}
\]

where the individual array dependent coefficients in the case of a rigid sphere are given as [3]:

\[
b_n(kr) = 4\pi i^n \left( j_n(kr) - \frac{j_n(kr)}{h_n'(kr)} \right), \tag{18}
\]

where \( j_n, h_n, j_n' \), and \( h_n' \) are the spherical Bessel and Hankel functions and their derivatives w.r.t. \( kr \), respectively.

### D. Spherical Harmonics with Real Arrays

Previous studies utilizing spherical microphone arrays have shown that by applying the theoretical discrete spherical harmonic transform on recordings of a real array as in Eq. (12) results in significant loss of accuracy in the transformed coefficients compared to the theoretical result for an ideal array with the same specifications as the real one [27, 28]. This is mainly due to deviations between the theoretical array model and the real one, such as positioning and calibration errors between microphones and microphone diaphragm effects. However the same studies indicate that by employing actual calibration measurements of the array in an anechoic chamber, the performance of the SHT approaches the theoretical one.

In essence the SHT of Eq. (12) can be represented by a set of encoding filters, and the equalization with the inverse modal weights \( B^{-1} \) can be obtained optimally in a least-squares sense. This is equivalent to the problem of obtaining a steering vector in the spherical domain that is optimal with respect to the measured array properties. Such measurement-based steering vector is employed in this work when the methods are applied to real data. The measurement procedure for the array at hand is detailed in Section IV.C.

Let us denote this measurement based steering vector as \( \hat{Y}(\Omega) \). Ideally this steering vector should be approximately equal to the theoretical steering vector to a plane wave in the transform domain

\[
\hat{Y}^H(\Omega) \approx Y^H(\Omega). \tag{19}
\]

Eq. (14) shows that the theoretical transformed array response, after equalization, can be factored into the following components

\[
Y^H(\Omega) = (B^{-1} Y^H(\Phi) \Gamma_\psi)(Y(\Phi)B Y^H(\Omega)) = W_{\text{enc}} h(\Omega), \tag{20}
\]

where \( W_{\text{enc}} \) expresses an encoding matrix that implements the SHT and equalization, and \( h(\Omega) \) is the array response vector to direction \( \Omega \). In the ideal case \( B^{-1} Y^H(\Phi) \Gamma_\psi Y(\Phi)B = I \) and Eq. (19) holds exactly. In practice both the encoding filters and the interpolated array response for an arbitrary \( \Omega \) can be computed from measurements of the response at \( L \) directions \( \Psi = [\psi_1, \ldots, \psi_L] \), uniformly distributed or e.g. arranged in a regular grid. The encoding filters can then be computed by a weighted regularized least-squares solution from the measurements as in [27, 29]

\[
W_{\text{enc}}(k) = Y^H(\Omega) \Gamma_\psi \hat{H}(k) \left( \hat{H}(k) \Gamma_\psi \hat{H}(k) + \beta(k)I \right)^{-1} \tag{21}
\]

where \( \hat{H}(k) = [\hat{h}_1, \ldots, \hat{h}_L] \) are the measured responses of the array to directions \( \Omega \) at frequency \( k \), \( \Gamma_\psi \) are appropriate weights for the measurement grid, and \( \beta \) is a regularization parameter set according to the indications in [27, 28]. In this work the weights of \( \Gamma_\psi \) were computed from the areas of the spherical Voronoi cells of the measurement grid.

To obtain an array steering vector \( \hat{h}(\Omega) \) at any direction with high accuracy, a spherical interpolation was performed by expanding the steering vector in terms of its measured SH coefficients, as

\[
\hat{h}(k, \Omega) = \left( Y^H(\Omega) \Gamma_\psi \hat{H}(k) \right)^T Y^H(\Omega) \tag{22}
\]

where the spherical harmonics matrices \( Y \) are evaluated up to an order determined by the number of measurements. For a regular grid we found that an appropriate truncation order \( N_t \) is approximately defined by \((N_t + 1)^2 = L/4\). By visually inspecting the energy of the SH coefficients, we concluded that coefficients beyond this truncation order are close to the noise level. Finally, by combining Eq. (21) and (22) we have the interpolated SH coefficients

\[
Y^H(k, \Omega) = W_{\text{enc}}(k) \hat{h}(k, \Omega). \tag{23}
\]

### III. Direction of Arrival Estimation

This section presents five DOA estimation methods. The first two, plane wave decomposition (PWD) and MUSIC, have been previously applied in the spherical microphone array processing [2, 3, 30], and the latter three, stochastic (S) and deterministic (D) ML methods, and WSF, are applied in the analysis of reflection via spherical microphone array processing for the first time in this paper. As mentioned earlier, WSF is a large sample approximation of the SML method.

The main difference between ML methods and WSF and the two previous methods, besides the fact that MUSIC and PWD are not maximum likelihood methods, is that in ML and WSF one can choose the signal and noise model freely. Therefore, whereas in PWD, MUSIC, and similar methods one searches for \( Q \) peaks from spatial spectrum of size \( D \), in ML
methods one searches for a single maximum (or equivalently a minimum) from an estimation function of size $Q \times D$. From this it follows that ML methods very quickly meet the curse of dimensionality when Q increases. Exhaustive search in this higher-dimensional space may be computationally prohibitive, but nonlinear optimization algorithms can often find a minimum in a reasonable amount of time. In section IV.A, non-linear optimization algorithms are presented for the ML methods and WSF.

Due to the nature of room impulse response and the applications in room acoustics, we are interested in two cases, the wideband analysis of reflections and the analysis of reflections in a single frequency. For the wideband analysis we apply the time domain smoothing algorithm, presented in [3], which is also reviewed briefly in this section. In the narrow band analysis, the smoothing algorithms will not provide any benefits, since the reflections are coherent, therefore the analysis is implemented for the frequency domain SH coefficients.

**A. Time domain smoothing**

We follow the formulations in [3], in the processing of the SH domain coefficients. The SH domain coefficients are normalized by multiplying Eq. (14) by $B^{-1}(kr)$ from the left side, which leads to

$$a_{nm}(k) = Y^H(\Omega)\hat{s}(k) + z(k),$$

where $z(k) = B^{-1}(kr)Y^H(\Phi)\Gamma n(k)$. The array covariance matrix in this case is given

$$R_{aa}(k) = Y^H(\Omega)R_{ss}(k)Y(\Omega) + R_{zz}(k)$$

where the noise covariance matrix

$$R_{zz}(k) = \sigma^2 I \|B^{-1}(kr)Y^H(\Phi)\Gamma\|^2;$$

is dependent on $k$ and $\|\cdot\|$ denotes the Frobenius norm.

Time domain version of $a_{nm}(k)$ is obtained via the inverse Fourier transform $F^{-1}\{\cdot\}$ as $a_{nm}(t) = F^{-1}\{a_{nm}(k)\}$. However, the multiplication with $B^{-1}$ leads to undesired noise amplification on certain frequencies [2, 3]. Therefore, in the time domain smoothing we use the same equalization as in [3]:

$$H_{eq}(k) = \left[\sigma^2 / \gamma \sum_{i=0}^{N_h} \sum_{m=-n}^{n} \|B^{-1}(kr)Y(\Phi)\Gamma\|^2\right]^{-1/2},$$

where the normalization factor $\sigma^2 / \gamma$ is constant for all frequencies. For the microphone array applied in this paper, the magnitude response of $H_{eq}(k)$ has the shape of a high-pass filter.

Because of the equalization, the array covariance matrix in the frequency domain becomes

$$\tilde{R}_{aa}(k) = Y^H(\Omega)\tilde{R}_{ss}(k)Y(\Omega) + \tilde{R}_{zz}(k)$$

and denote the equalized versions of the variables with \(\tilde{\cdot}\). In the estimation, we assume that the time domain version of noise matrix is independent of time, and spatially white, i.e., $R_{zz}(t) = \tilde{R}_{zz} = \sigma^2 I$, where $\sigma^2$ is the equalized variance.

The array covariance matrix estimate for $N$ time instants $t_1, \ldots, t_N$ is given as

$$\tilde{R}_{aa} = 1/N \sum_{i=1}^{N} \tilde{a}(t_i)\tilde{a}^H(t_i),$$

where $\tilde{a}_{nm}(t) = F^{-1}\{H_{eq}(k)\tilde{a}_{nm}(k)\}$ are the equalized SH coefficients in the time domain.

In this paper, we are also interested in the performance in single frequency bins. In the frequency domain estimation, we apply spatial whitening to the covariance matrix estimate and the steering vectors as in [2]. The following formulations in this section are presented for the time domain smoothing, but are equal to the frequency domain smoothing versions if the covariance matrix estimate and the steering vectors are replaced with their respective whitened versions.

**B. The steering matrix**

The steering vector or matrix, used in all the methods, is the Hermitean transpose of Eq. (16). For example, in ML methods and WSF, the matrix has the following form in the case of two reflections:

$$Y(\Omega) = \begin{bmatrix} Y^0(\Omega_1), Y_1^{-1}(\Omega_1), Y_1^0(\Omega_1), \ldots, Y_1^Q(\Omega_1) \\ V^0(\Omega_2), Y_2^{-1}(\Omega_2), Y_2^0(\Omega_2), \ldots, Y_2^Q(\Omega_2) \end{bmatrix}$$

since there are two possible reflections $\Omega_1$ and $\Omega_2$, and the number of harmonic components is $(N_h + 1)^2 = 16$. In contrast, in MUSIC and PWD, the steering is always 1-D and has the form

$$Y(\Omega) = \begin{bmatrix} Y^0(\Omega), Y_1^{-1}(\Omega), Y_1^0(\Omega), Y_1^1(\Omega), \ldots, Y_1^Q(\Omega) \end{bmatrix}.$$ 

These matrices illustrate the fundamental difference between 1-D and 2-D search space, applied for spectral-based methods and ML methods and WSF, respectively.

For ML methods and WSF the localization function $P(\Omega)$ is 2Q-dimensional since $\Omega$ has $Q$ dimensions, inclination and azimuth values. The minimum argument of the localization function for ML methods and WSF gives the DOA estimates:

$$\hat{\Omega} = \arg \min_{\Omega} \{P(\Omega)\}. $$

For spectral-based methods, the localization function $P(\Omega)$ is 2-dimensional, since $\Omega$ has 2 dimensions, inclination and azimuth. The DOA estimates for the spectral-based methods, are the $Q$ highest local maxima in $P(\Omega)$.

**C. Beamforming**

Beamforming has been a popular approach for direction estimation for several decades [12]. For spherical microphone arrays, beamforming is studied extensively [10, 13, 31, 32] and one of the most commonly applied beamformer is the plane wave decomposition (PWD) [13]. PWD has the characteristic
of maximum rejection of isotropic spatial noise and is given over a frequency range \( k = 1, \ldots, K \) as

\[
\hat{g}(\Omega) = \frac{1}{K} \sum_{k=1}^{K} W(k, \Omega) \hat{p}_{nn}(k). \tag{33}
\]

In the above, \( \hat{p}_{nn}(k) \) is the noisy version of Eq. (14) and the weighting is given by

\[
W(k, \Omega) = \frac{4\pi}{(N_h + 1)^2} Y(\Omega) B^{-1}(k \tau),
\]

where the first term ensures unity beamformer in the look direction. This form of PWD does not whiten the noise, and therefore will have a poor performance in the wideband case. Whitening of the noise can be implemented following [2].

With the time domain smoothing, PWD is implemented here as:

\[
\hat{g}(\Omega) = \frac{1}{N} \sum_{i=1}^{N} Y^H(\Omega) \hat{a}_{nn}(t_i), \tag{34}
\]

The output energy of the time-domain beamformer at \( \Omega \) is given as

\[
P_{\text{PWD}}(\Omega) = \| \hat{g}(\Omega) \|^2 = Y(\Omega) \hat{R}_{aa} Y^H(\Omega),
\]

where \( \hat{R}_{aa} \) is the estimated array covariance matrix, given in Eq. (31).

**D. MUSIC**

One of the most popular methods for direction estimation with spherical microphone arrays is MUSIC [2, 3]. The spatial spectrum of MUSIC is calculated as

\[
P_{\text{MUSIC}}(\Omega) = \frac{1}{\| E_n^H Y^H(\Omega) \|^2} \tag{35}
\]

where \( E_n \) is the array noise matrix from eigenvalue decomposition of the array covariance matrix estimate \( \hat{R}_{aa} \). This decomposition follows the form

\[
\hat{R}_{aa} = E \Lambda E^H = E_s \lambda_s E_s^H + E_n \lambda_n E_n^H \tag{36}
\]

where the superscripts \( s \) and \( n \) denote signal and noise subspaces, respectively, \( \lambda \) is the eigenvalue matrix, and \( E \) includes the right eigenvectors.

MUSIC requires a full rank reflection signal covariance matrix. Therefore, when MUSIC is applied to localize Q reflections, the estimated reflection signal covariance matrix should have Q eigenvalues deviating from the noise. This assumption is violated if we have too few snapshots of the array covariance matrix or coherent reflections. The snapshots here refer to time domain or frequency domain SH coefficients.

**E. Maximum likelihood methods**

Maximum likelihood methods are generally applied in several estimation tasks. For an overview on maximum likelihood estimation, the reader is referred to [33]. A requirement for the ML method is a signal and noise model. These models are dependent on the parameters that are estimated by the ML method. The problem in ML estimation is then to find the parameters of the model that most likely explain the observed data. In this paper, we apply two ML methods developed earlier in the space domain to the spherical microphone array processing.

1) **Stochastic Maximum Likelihood:** The first ML method is called stochastic (SML), due to the assumption that the reflection signals are stochastic processes. The array covariance matrix in the case of time domain smoothing takes the form

\[
\hat{R}_{aa} = Y^H(\Omega) \hat{R}_{ss} Y(\Omega) + \sigma_a^2 I.
\]

The probability density function of SML for time instant \( t_i = t_1, \ldots, t_N \) is given as

\[
p(\hat{A}_N|\Omega, \hat{R}_{aa}, \sigma_a^2) = \prod_{i=1}^{N} \frac{1}{\pi(N_h+1)^2 |\hat{R}_{aa}|} \exp \left( -\hat{a}^H(t_i) \hat{R}_{aa}^{-1} \hat{a}(t_i) \right), \tag{37}
\]

where \( \hat{A}_N = [\hat{a}(t_1), \ldots, \hat{a}(t_N)] \) are the equalized time domain SH coefficients, \( \hat{R}_{aa} \) is the modeled array covariance matrix and \( |\cdot| \) denotes the determinant of a matrix.

As usually in ML methods, the solution is found from the negative log-likelihood which is given for SML as:

\[
\log(p(\hat{A}_N|\Omega, \hat{R}_{aa}, \sigma_a^2)) = \log |\hat{R}_{aa}| + \text{Tr}\{ \hat{R}_{aa}^{-1} \hat{R}_{aa} \}. \tag{38}
\]

Solving the negative log-likelihood w.r.t. \( S, \sigma^2 \) (see [26] for details), leads to

\[
\hat{R}_{ss}(\Omega) = Y^H(\hat{R}_{aa} - \sigma_a^2 I) Y^H(\Omega) \tag{39}
\]

\[
\sigma_a^2 = \frac{1}{(N_h+1)^2 - Q} \text{Tr}\{ \hat{P}_Y(\Omega) \hat{R}_{aa} \}, \tag{40}
\]

respectively. In the above

\[
Y^H(\Omega) = (Y(\Omega) Y^H(\Omega))^{-1} Y(\Omega) \tag{41}
\]

\[
\hat{P}_Y(\Omega) = I - Y^H(\Omega) Y(\Omega) \tag{42}
\]

are the pseudo-inverse of \( Y(\Omega) \) and the orthogonal projector onto the null space of \( Y^H(\Omega) \), respectively. The localization function for SML is given as [26]:

\[
P_{\text{SML}}(\Omega) = \log \left\{ \frac{|Y(\Omega) \hat{R}_{ss} Y^H(\Omega) + \sigma_a^2 I|}{|Y(\Omega)|^2} \right\}. \tag{43}
\]

2) **Deterministic Maximum Likelihood:** The deterministic model makes no assumptions on the signal waveforms. That is, the average signal waveform has the form \( E[\hat{a}(t_i)] = Y^H(\Omega) \tilde{s}(t_i) \). When the average signal waveform is deduced from the signals, the array covariance matrix is only dependent on the noise term

\[
\hat{R} = E\{ (\hat{a}(t_i) - E[\hat{a}(t)]) (\hat{a}(t_j) - E[\hat{a}(t)])^H \} = \sigma_a^2 I
\]

The likelihood function for DML when several snapshots are available is given as [26]:

\[
p(\hat{A}_N|\Omega, \hat{S}_N, \sigma_a^2) = \prod_{i=1}^{N} \frac{1}{|\sigma_a^2 I|} \times \exp \left( -\frac{(\hat{a}(t_i) - Y^H(\Omega) \tilde{s}(t_i))^H (\hat{a}(t_i) - Y^H(\Omega) \tilde{s}(t_i))}{\sigma_a^2} \right), \tag{44}
\]
where \( \hat{S}_N = [\hat{s}(t_1), \ldots, \hat{s}(t_N)] \) are the reflection signals. From the negative log-likelihood, setting \( \Omega \) and \( \hat{S}_N \) constant, the variance can be estimated as [26]:

\[
\hat{\sigma}_a^2 = \frac{1}{(N_n + 1)^2} \text{Tr}(P_Y^\perp(\Omega)\hat{\mathbf{R}}_{aa}).
\]  

(45)

Using this in the negative log-likelihood leads to a non-linear least-squares problem, from where the localization function and reflection signal estimates can be written as [26]:

\[
P_{\text{DML}}(\Omega) = \text{Tr}\left\{P_Y^\perp(\Omega)\hat{\mathbf{R}}_{aa}\right\}
\]

and

\[\hat{S}_N = Y^\dagger(\Omega)A_N,\]

(46)

(47)

respectively.

In the incoherent case, i.e., for uncorrelated reflection signals and large sample size, MUSIC is asymptotically equivalent to DML [26].

3) Weighted Subspace Fitting: Subspace fitting methods are suboptimal approximations of the above maximum likelihood methods. They are of interest since they have a lower computational complexity than the ML methods. In addition, for large sample sizes and if some conditions are fulfilled they are asymptotically equivalent to ML methods. [26]

The WSF method is a large sample approximation of the SML method, and its localization function is given as [26]:

\[
P_{\text{WSF}}(\Omega) = \text{Tr}\left\{P_Y^\perp(\Omega)\hat{\mathbf{E}}_{a}W\hat{\mathbf{E}}^H_{a}\right\},
\]

\[
W = \lambda^2\lambda_s^{-1}
\]

(48)

where \( \lambda = (\lambda_s - \hat{\sigma}_a^2) \) and \( \lambda_s \) is the signal eigenvalue matrix, as previously, and \( \hat{\sigma}_a^2 = \frac{1}{(N_n + 1)^2 - Q} \sum \text{diag}\{\lambda_n\} \) is the average of \( (N_n + 1)^2 - Q \) smallest eigenvalues. This weighting gives the lowest asymptotic error variance, as shown in [26].

F. Cramér-Rao Lower Bound on the estimation accuracy

Estimation theory studies the performance of the methods by comparing their covariance of the estimation error against a theoretical lower bound. A commonly used bound on the estimation covariance matrix error is the Cramér-Rao lower bound (CRLB). [26, 33]

For the cases studied in this article, the error covariance of an unbiased estimate \( \Omega \), i.e., \( E\{\Omega\} = \Omega \) is bound by

\[
E\{(\Omega - \hat{\Omega})(\Omega - \hat{\Omega})^T\} = \left[-E\left\{\frac{\partial^2 \log(p(\hat{\mathbf{A}}_{n}\mid \mathbf{Y}))}{\partial\mathbf{\Omega}\partial\mathbf{\Omega}^T}\right\}\right]^{-1}.
\]

(49)

In the following \( \Omega \) and \( k \) are omitted from the notation for compactness. Substituting the SML probability density function in Eq. (37) to the above gives the stochastic CRLB [26]:

\[
\{\text{CRLB}_{\text{stochastic}}\}_{ij} = \frac{\hat{\sigma}_a^2}{2} \left[\Re\left\{\text{Tr}\left\{Y_i^\dagger P_YY_i^H\mathbf{R}_{ss}Y_i^\dagger P_{\text{Y}}\mathbf{R}_{ss}Y_i\right\}\right]\right]^{-1},
\]

(50)

where \( \Re\{\cdot\} \) is the real part of a complex number, and \( Y_i \) is

\[
Y_i = \frac{\partial Y}{\partial \Omega_i},
\]

(51)

the partial derivative of \( Y \) w.r.t. \( i \)th element in \( \Omega \). The elements of \( Y_i \) are the partial derivatives of spherical harmonics w.r.t. inclination \( \theta \) and azimuth \( \phi \), which are given as:

\[
\frac{\partial Y^m_n(\Omega)}{\partial \theta} = -\frac{\sqrt{(n+1)(n-m)!}}{4\pi (n+m)!} e^{im\phi} \csc(\theta) \times \]

\[
\left[(n+1) \cos(\theta) P_n^m(\cos(\theta)) + (m-n-1) P_{n+1}^m(\cos(\theta))\right]
\]

(52)

and

\[
\frac{\partial Y^m_n(\Omega)}{\partial \phi} = \text{im} Y^m_n(\Omega),
\]

(53)

respectively, where \( \csc(\cdot) \) is the cosecant function. These partial derivatives are available in the literature and, for example, in MATHEMATICA. In [34], the CRLB for a single stochastic source in the SH domain is presented.

Substituting the DML probability density function in Eq. (44) to Eq. (49) leads to the deterministic CRLB [26]:

\[
\{\text{CRLB}_{\text{det}}\}_{ij} = \frac{\hat{\sigma}_a^2}{2} \left[\Re\left\{\text{Tr}\left\{Y_i^\dagger P_YY_i^H\mathbf{R}_{ss}\right\}\right]\right]^{-1}.
\]

(54)

G. Detection of reflections

DOA estimation requires knowledge on the number of reflections \( Q \). Numerous methods for resolving the number of reflections, i.e., detection, have been proposed, and the reader is referred to [12] for an overview. Some of the detection methods estimate the subspace dimensions from the eigenvalue matrix, for example, by statistically testing how many of the eigenvalues belong to the noise space [26]. For detection of partially correlated source signals, the best approaches are the model-based approaches, such as generalized likelihood ratio test and WSF detection [26]. These approaches simultaneously detect the number of \( Q \) and estimate the DOA. It is therefore expected that these methods should also perform well for the detection of reflection signal, which are correlated or coherent. In this paper, we do not investigate the detection, but assume that \( Q \) already exists as a prior knowledge, as in previous research on this topic [2, 3].

IV. EXPERIMENTS

This section describes simulation and real data experiments with a spherical microphone array. In all cases, we use em32 Eigenmike®, microphone array which has 32 capsules on the surface of a rigid sphere. A technical description, microphone positions etc. of the Eigenmike, is given for example in [35].

The results of the experiments are investigated with the root mean squared error (RMSE),

\[
\text{RMSE} = \sqrt{E\{(\Omega - \hat{\Omega})(\Omega - \hat{\Omega})^T\}},
\]

which is compared against the square root of the CRLB. In all the simulation experiments the RMSE is averaged over 100 Monte-Carlo Samples.

In the simulations of this paper, the signal-to-noise ratio (SNR) is reported as the space-domain SNR, i.e., \( \text{SNR} = (1/\sigma^2) = (\sigma^{-2}) \). However, perhaps a more meaningful SNR value is the effective SNR, which can be calculated as the relation between equalized reflection signal variance and the
equalized noise variance, i.e., \( \text{SNR}_c = \| \hat{s}(k) \|^2 / \hat{\sigma}_s^2 \). The noise is simulated as i.i.d. complex Gaussian random variable in the space domain with a variance \( \sigma^2 \) in all the simulated cases of this paper.

A. Search of the global minimum/maximum via non-linear optimization methods

As discussed previously, the localization functions are highly non-linear and therefore non-linear optimization methods must be applied if computational time is a requirement. In this paper we use Newton-type search algorithm to find the minimum of SML, DML, and WSF localization functions, similarly as in [26], where Levenberg-Marquardt (LM) technique is used. The LM technique requires the true gradient and Hessian matrices of the localization function. Instead of LM, here we use the Quasi-Newton (QN) method with Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, which estimates the Hessian via finite differences. We chose QN method since it is readily available in MATLAB in the fminunc-function. The iteration in the QN method is stopped if the function value or the estimated parameters do not change more than \( 1 \times 10^{-10} \) between two consecutive iterations. Moreover, the maximum number of iterations, allowed for the QN method, is set to 10000.

The QN method requires a reasonable initial guess for the estimated parameters. In the simulations of this paper, the parameters are initialized around the true values in the QN method, by adding a random number to the true values. The random numbers are drawn from a uniform distribution with a variance of 2 degrees and mean 0. In real situations, the true values are not necessarily available and therefore we follow the suggestion in [26], and search the initial value for SML, DML, and WSF via a non-linear optimization algorithm called the alternating projection (AP) algorithm [36]. AP has a property of always converging to a local minimum, which is not necessarily global. However, here, the local minimum found by AP is assumed to be also the global minimum of the applied search grid. We use \( Q + 2 \) iterations in the AP algorithm including the initialization and a grid which has a 5 degree resolution w.r.t. to both azimuth and inclination.

In simulated situations, we use the QN method with BFGS algorithm to find the maxima of MUSIC and PWD localization functions and initialize the parameters around the true values. In real situations, we initialize MUSIC and PWD by finding \( Q \) highest local maxima from their spatial spectra. These spatial spectra are calculated for the same grid that is used above in the AP algorithm.

B. Simulation

1) A single wideband reflection: We first simulate a case where one reflection is present in a time window. The sampling frequency is set to \( f_s = 48 \text{ kHz} \) and the length of the window is \( N = 2048 \). In addition, the reflection is simulated as a wide band plane wave reflection \( s(t) = \delta(t-1024/f_s) \) arriving from \( \Omega = (90,0)° \) and with a harmonic order \( N_h = 3 \). Moreover, the time domain SH coefficients are calculated as shown in section III.A. via the discrete Fourier Transform of size \( N \). For a single plane wave, the reflection signal covariance matrix is

\[
\mathbf{R}_{ss} = \sum_{k=1}^{N} \| H_{eq}(k) \|^2 = \mathbf{E}_{eq},
\]

where \( E_{eq} \approx 0.48 \) is the total energy of the equalized reflection signal.

Figure 1 shows the RMS error for each method over 100 Monte-Carlo samples at different noise variances \( \sigma^2 \). Also theoretical performance in the simulated case, i.e., the CRLB is shown for the stochastic and deterministic case in Fig. 1. It can be seen that the methods have about equal performance in the case of one reflection. The RMSE of the methods follows the CRLB when \( \text{SNR} \leq -10 \text{ [dB]} \). The estimation of inclination angle has a lower RMSE, as predicted also by the CRLB.

2) Two wideband reflections: In the second example, we simulate four cases where two reflections arrive during the same time window. Similarly as in [3, Section IV.B], two wideband reflections are simulated of which the second one is delayed by \( d \) samples. That is, the signals for the first and second reflection are \( s_1(t) = \delta(t-1024/f_s) \) and \( s_2(t) = \delta(t-1024+d)/f_s \), respectively. The four simulated delays are \( d = 0, d = 0.5, d = 1, \) and \( d = 8 \) samples. When the delay approaches 0, the reflection signal covariance matrix approaches coherence, and when \( d \) increases the signal covariance matrix approaches incoherence. For the four cases studied here, the reflection signal covariance matrices are

\[
\mathbf{R}_{ss} = E_{eq} \begin{bmatrix} 1.00 & 1.00 \\ 1.00 & 1.00 \end{bmatrix}, \quad \mathbf{R}_{ss} = E_{eq} \begin{bmatrix} 1.00 & 0.85 \\ 0.85 & 1.00 \end{bmatrix},
\]
\[
\mathbf{R}_{ss} = E_{eq} \begin{bmatrix} 1.00 & 0.50 \\ 0.50 & 1.00 \end{bmatrix}, \quad \text{and} \quad \mathbf{R}_{ss} = E_{eq} \begin{bmatrix} 1.00 & -0.04 \\ -0.04 & 0.97 \end{bmatrix}
\]

for \( d = 0, d = 0.5, d = 1, \) and \( d = 8 \), respectively.

Both of the reflections are simulated as plane waves with harmonic order \( N_h = 3 \) and DOAs \( \Omega_1 = (90,0)° \) and \( \Omega_2 = (90,\phi_2)° \), where the direction of the second plane...
wave is varied as $2^{1+n_w \times 0.2}$, $n_w = 0, 0.5, 1, \ldots 7$ [°]. In these experiments the signal-to-noise ratio is set to 60 dB.

The time domain SH coefficients are evaluated as previously and they are windowed using a $N_w = 25$ samples long rectangular window. This window is centered around the first reflection at $1024/f_s$. The array covariance matrix estimate is then averaged over $N_w = 25$ samples in the time domain as shown in Eq. (31).

Figure 2 show the results for the cases. The first observation from the results is that the performance of PWD is poor in the DOA estimation of multiple reflections as its RMSE increases as $\phi_2$ increases. This is explained by a bias in the estimation. Namely, the global maximum in PWD is found in between the true values and the second maximum is $180^\circ$ apart from the first one. The two reflections produce a peak in the spatial spectrum of PWD in between the true values of $\phi_1$ and $\phi_2$. Because of the high amount of energy that this maximum has, the RMSE is lower than CRLB, when $\phi_2 \leq 3^\circ$ with $d = 0$. The second maximum in the PWD spectrum is produced by the conjugate value of the steering vector that produced the global maximum. When $d = 0$, also MUSIC shows similar behavior as PWD and its estimation is similarly biased. When $\phi_2 > 60^\circ$ the performance of PWD and MUSIC slightly improves. This is due to the fact that separation is more than the Rayleigh resolution $180^\circ / N_b = 60^\circ$, and the methods are able to separate the two reflections. However, also when $\phi_2 > 60^\circ$, both MUSIC and PWD estimates are biased.

ML methods and WSF have a lower value than CRLB when $\phi_2 \leq 3^\circ$ in the coherent case, i.e., when $d = 0$. Also this is caused by a bias in the estimation. All of the methods have the global minimum in between the true values and they get strong evidence for the biased estimate, similarly as PWD above. Thus, the information contained in the second order moments of the array covariance matrix is too coherent for unbiased estimation when $\phi_2 \leq 3^\circ$ and $d = 0$. First, the reflection signals are coherent. Second, the steering vectors are very similar when the reflections are arriving from directions that are close to each other. To obtain a higher performance when $\phi_2 \leq 3^\circ$, more microphones would be required. For $d = 0$, when $\phi_2 > 3^\circ$ the ML methods and WSF follow the CRLB and perform clearly better than PWD or MUSIC.

In the partially correlated cases $d = 0.5$ and $d = 1$, ML methods and WSF have the same performance. MUSIC has a slightly higher RMSE than the ML methods and WSF. That is, the ML methods and WSF outperform MUSIC in the partially correlated case. In the almost incoherent case with $d = 8$, MUSIC, the ML methods and WSF have the same performance. PWD performance when $d = 0.5$, $d = 1$ and $d = 8$ is similar as for the case when $d = 0$.

The CRLB does not predict the RMSE very well when RMSE is high [37]. That is, CRLB assumes a low error variance. According to [26], CRLB and RMSE are approximately in agreement when the theoretical standard deviation of the DOA estimation of the reflections is less than half the angle separation. This is also approximately true in the above studied cases. When $d = 0$ and $\phi_2 = 3^\circ$, CRLB is $2.1^\circ$ for the first reflection and $2.3^\circ$ for the second reflection which is more than the half of the angle separation $3/2 = 1.5^\circ$. 

![Fig. 2. RMS error for all the methods over 100 Monte-Carlo Samples and CRLB against $\phi_2$, when the second reflection is delayed by $d$ samples and SNR = 60 dB.](image-url)
C. Two reflections in a single frequency

Next we study the performance of the methods in a single frequency. As explained above, since the processing is applied in the frequency domain, the noise is whitened as in [2]. That is, the whitened frequency domain covariance matrix and steering vectors are used instead of the time domain versions in Eqs. (16) and (31). When analyzing DOA in a single frequency with a microphone array, one should consider the spatial aliasing. As mentioned, spatial aliasing typically occurs in high frequencies, when \(k r > N_h\) [9]. For the applied array this limit is about 4.7 kHz. However, when analyzing single frequencies, the ML methods and WSF are unaffected by the spatial aliasing, since \(B(kr)\) is continuous, i.e., it does not contain any zeros, and \(\mathbf{Y}(\Phi)\) and \(\mathbf{Y}(\Omega)\) are equal for all frequencies.

When the analysis is performed in a single frequency the reflection signal, the covariance matrix is inevitably coherent, i.e.:

\[
\tilde{R}_{ss}(k) = \|H_{eq}(k)\|^2 \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},
\]

(55)

and the array covariance matrix estimate becomes

\[
\tilde{R}_{aa}(k) = \hat{a}(k)\hat{a}^H(k).
\]

The selected frequencies for this experiment are the center frequencies of the octave bands \(f = 62.5 \times 2^n\), where \(n = 0, 1, \ldots, 8\). It should be noted that the array covariance matrices are not averaged over the octave bands, but the performance is examined in single frequencies as shown in Eq. (56).

The reflections are simulated as wideband reflections with \(d = 0\), \(N = 2048\), \(N_h = 3\), and SNR is set to 60 dB. Frequency domain SH coefficients are evaluated with Eq. (24) via the discrete Fourier Transform, and the frequency bins that are the closest to Fourier frequencies \(f\) are used in the analysis. The frequency resolution is \(f_s/N \approx 23\) Hz and the -3 dB bandwidth for each analyzed frequency is about 21 Hz for each frequency due to the aliasing or “spectral leakage” from the windowing.

The results of the experiment for different frequencies are shown in Fig. 3. The results show that the methods obtain a similar performance as in the wideband case when \(f = 4\) kHz and \(f = 8\) kHz, which is expected since \(f = 4\) kHz gives the best performance. The reason why RMSE < CRLB when \(f = 4\) kHz, \(f = 8\) kHz and \(f = 16\) kHz is the same as in the wideband case, explained above. That is, the difference between the steering vectors for reflections arriving from almost the same angle is small, and the reflection signal covariance matrix is coherent. PWD and MUSIC suffer from the same problems in the case of single frequencies as in the wideband case when \(d = 0\).

When \(f = 62.5\) Hz, \(f = 125\) Hz, and \(f = 250\) Hz, the estimation with all methods is biased as the RMSE is on average about 2 degrees for all values of \(\phi_2\). We can see from Fig. 3, that for frequencies the \(f = 62.5\) Hz, \(f = 125\) Hz, and \(f = 250\) Hz, the CRLB is higher than the half of the angle separation with all the studied separation angles. Thus, it is expected that CRLB and RMSE do not meet when the error variance is as high. The poor performance on these frequencies is a consequence of a poor effective SNR. In low frequencies, due to the array geometry, the effective SNR is lower than in higher frequencies.

D. Real data

1) Measurement setup: To test the methods in real situations, measurements were made in a semi-anechoic room with dimensions 12m ×11m ×7m. The octave band reverberation times as well as theoretical absorption coefficients, calculated with Sabine’s equation are shown in Table I. The room has highly absorptive walls and ceiling, which are treated with 5 cm of mineral wool, placed 50 cm in front of a concrete wall to generate an air gap, as shown in Fig. 4. The material of highly reflective floor is linoleum on concrete. Seven reflections were introduced to the measurement by building a reflective corner to the room, shown also in Fig. 4. This corner was built of two reflective projection silver screens of size 3 ×3 m, parallel to the walls of the room. The reflections include the first, second, and third order reflections from the inserted corner building and floor in addition to the direct sound from the source to the array.

A Genelec 1029A was placed in the source position and em32 Eigenmike to the receiver position, as shown in Fig. 4. The loudspeaker was facing the em32 microphone array and the acoustic center of the loudspeaker as well as the center of the microphone array was at 1.3 m height. The applied loudspeaker is flat in the direction of the reflective surface up to 1 kHz, and is attenuated about 6 dB from 1 kHz to 10 kHz in the directions ±45° w.r.t. the central horizontal plane. Based on tabulated values, it is estimated that the absorption coefficient of the projection silver screens is less than 0.1 for frequencies above 500 Hz and less than 0.1 for the floor for all frequencies. Due to the size of the reflective surface, material, and the geometry of the source-array setup, it is estimated that reflections from the reflective surface are obtained for frequencies above 250 Hz.

An impulse response was measured with a 10 s long exponential sine-sweep from 1 Hz to 24 kHz at a sampling rate of \(f_s = 48\) kHz [38]. The impulse responses of the 32 em32 Eigenmikes as well as the SH time domain coefficients are shown in Fig. 5, which are normalized in the visualization such that the maximum absolute value is 1. The impulse responses were truncated to 7.2 ms after the direct sound, resulting in a total length of 1024 samples. The truncation was done to include only the reflections that are under study, although, as shown in Fig. 5, there are no visible reflections in the impulse response after the 7th reflection.
The steering vectors were estimated from a set of anechoic measurements of the em32 Eigenmike. The impulse responses were measured at 48 kHz from 1 Hz to 24 kHz with a 5 s long exponential sine sweep signal. The measurement grid had 5 and 10 degree resolutions for azimuth and inclination, respectively, leading to a total of 1225 points. Moreover, the distance from the source, a Genelec 8020B loudspeaker, to the em32 microphone array was 2 m. The initial time delay of the direct sound and the internal delay due to buffers etc. was first removed from the responses. Then, an additional direction-dependent sinusoidally varying delay at the measured responses, due to the acoustic center of the array not being exactly fixed while the array was rotated, was compensated in a similar way as described in [28, 39]. Finally, the measurements were truncated to 2048 samples and processed as shown in Section III.D. to obtain SH coefficients in the frequency domain for harmonic order $N_h = 3$.

3) Processing and results: We study the performance of the methods with two experiments. The first experiment is a wideband analysis using the time smoothing algorithm, as suggested in [3]. The second experiment is a single frequency analysis in the frequency domain. In the wideband evaluation, time domain SH coefficients are windowed using a 48 sample long (1 ms) Hanning window around the time of arrival of each reflection and the direct sound, as shown in Fig. 5(b). The array covariance matrix is then averaged from 48 time domain SH coefficients for each of the eight time windows. A single direction of arrival is estimated for each of windows. Thus, we set the number of reflections to $Q = 1$ for each time window. In the single frequency analysis, we window the time domain impulse responses with two rectangular windows of

Fig. 3. RMSE of the ML methods over 100 Monte-Carlo Samples and CRLB for selected frequencies when SNR = 60 dB. Best performance is achieved when $f = 4$ kHz.

Fig. 4. The measurement setup in the semi-anecho room.

Fig. 5. (a) The 32 room impulse responses $\hat{p}(t)$ measured with em32 Eigenmike and the (b) corresponding real part of the $(N_h + 1)^2 = 16$ SH coefficients $\hat{a}_{nm}$ in time domain, as well as the windowing (---) applied in the experiments. Note that in both (a) and (b) the impulse responses overlap in the visualization heavily.

2) Measurement of the steering vectors: The steering vectors were estimated from a set of anechoic measurements of the em32 Eigenmike. The impulse responses were measured at 48 kHz from 1 Hz to 24 kHz with a 5 s long exponential sine sweep signal. The measurement grid had 5 and 10 degree resolutions for azimuth and inclination, respectively, leading to a total of 1225 points. Moreover, the distance from the source, a Genelec 8020B loudspeaker, to the em32 microphone array was 2 m. The initial time delay of the direct sound and the internal delay due to buffers etc. was first removed from the responses. Then, an additional direction-dependent sinusoidally varying delay at the measured responses, due to the acoustic center of the array not being exactly fixed while the array was rotated, was compensated in a similar way as described in [28, 39]. Finally, the measurements were truncated to 2048 samples and processed as shown in Section III.D. to obtain SH coefficients in the frequency domain for harmonic order $N_h = 3$.

3) Processing and results: We study the performance of the methods with two experiments. The first experiment is a wideband analysis using the time smoothing algorithm, as suggested in [3]. The second experiment is a single frequency analysis in the frequency domain. In the wideband evaluation, time domain SH coefficients are windowed using a 48 sample long (1 ms) Hanning window around the time of arrival of each reflection and the direct sound, as shown in Fig. 5(b). The array covariance matrix is then averaged from 48 time domain SH coefficients for each of the eight time windows. A single direction of arrival is estimated for each of windows. Thus, we set the number of reflections to $Q = 1$ for each time window. In the single frequency analysis, we window the time domain impulse responses with two rectangular windows of
length 512 samples, tapered in the beginning and end. The windows are selected so that both of the windows include four acoustic events, as shown in Fig. 5(a). The first window includes the direct sound and three reflections and the second one includes four reflections. That is, in the estimation, we use $Q = 4$ for both windows. The array covariance matrix is calculated using a single frequency SH coefficients in the frequency domain, as above in the simulations in Section III.C.

The results of the real data experiments are shown for each method in Table II as an average RMSE over the eight DOA estimates. As we can see from the results, there is not much difference in the performance of the methods in the wideband case. This result was also predictable from the first simulation result. That is, when $Q = 1$, all the methods perform equally.

The results of the single frequency analysis in Table II show that SML performs clearly worse than DML and WSF. This degradation in the performance is caused by the fact that SML does not converge to the global minimum but to a local minimum. This result is already shown in [26, p. 61], where SML has a worse probability to converge to a global minimum than DML or WSF. PWD and MUSIC perform worse than the other methods, as expected from the simulations. As a conclusion of the experiments, DML and WSF have a similar performance and should thus be preferred in the analysis of multiple reflections, especially in the single frequency analysis. In addition, WSF has an advantage over DML since it is computationally lighter, especially if $Q$ is large, which is typical in room acoustics.

4) Discussion: The results of this paper show that if a single reflection is present in the analysis window, all the studied methods perform well. However, if there are two reflections that are highly correlated, the estimation of PWD and MUSIC is biased and the ML methods are by far a better choice. As shown in Fig. 2, all the methods benefit if smoothing is applied. Similarly as with the time-domain smoothing, the performance of all the methods would increase if frequency smoothing would be applied.

In the real experiments the reflective wall does not reflect low frequencies well and the estimation for them is very erroneous, as shown in Table II. Actually the time window in the analysis of the lowest frequencies ($f < 125$ Hz) should be longer, since the low frequencies travel a longer path. They go through the projection screen, the mineral wool, and are eventually reflected from the concrete wall. Namely, as the air gap in the wall is 0.45 m, frequencies above about 200 Hz are attenuated in the air gap, but lower frequencies are reflected back. This can also be observed from the theoretical $\alpha_i$ in Table I. In addition, as shown in the simulations, the performance of DOA estimation is poor on lower frequencies ($f < 500$ Hz). However, even if the windowing would be correct for the lowest frequencies, the DOA estimation would still suffer from the same problems as in the simulated case with low frequencies. For the analysis of multiple coherent reflections in low frequencies ($f < 250$ Hz), a microphone array with a larger radius should be used.

As shown here, the performance of DOA estimation of several reflections with a rigid sphere depends on the analyzed frequency. The radius of the sphere affects on the accuracy of the estimation in the specific frequency as shown by the CRLB. The larger the radius, the better the performance in lower frequencies. With the current array of radius 4.2 cm the best performance was achieved around 4 kHz. If the analysis would require similar performance, say around 250 Hz, the radius should be about 67.2 cm.

V. CONCLUSIONS AND FUTURE WORK

This paper studied the direction estimation of reflections from spatial room impulse responses. The analysis case of multiple reflections in a small time window was of special interest. This case is known to be difficult to methods such as MUSIC or ESPRIT since the reflections are highly correlated or even coherent and these methods rely on the independence of the reflections. The problem of estimating the DOA of coherent reflections has not been solved previously. However, the case of partially correlated or incoherent reflections has been solved previously by smoothing, i.e., averaging, the array covariance matrix over time [3], frequency [2] or space [16].

A room impulse response may have different characteristic at every time instant and in every frequency. Therefore, smoothing results in a lowered resolution with respect to time, frequency, or space, which impairs the detailed analysis of the room acoustics. This paper proposed to use ML methods in the DOA estimation of reflections, where smoothing operations can be avoided and which can cope with coherent signals.

Simulations and real data experiments were conducted to study the performance of the direction estimation with PWD beamformer, MUSIC, two ML methods, a deterministic and a stochastic one, and a subspace technique, WSF, which is a large sample approximation of the stochastic ML. The results

<table>
<thead>
<tr>
<th>Fr. [Hz]</th>
<th>Method</th>
<th>RMSE of inclination $\theta$ [°]</th>
</tr>
</thead>
<tbody>
<tr>
<td>62.5</td>
<td>PWD</td>
<td>13.51</td>
</tr>
<tr>
<td></td>
<td>MUSIC</td>
<td>29.82</td>
</tr>
<tr>
<td></td>
<td>DML</td>
<td>45.86</td>
</tr>
<tr>
<td></td>
<td>SML</td>
<td>31.40</td>
</tr>
<tr>
<td></td>
<td>WSF</td>
<td>45.86</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fr. [Hz]</th>
<th>Method</th>
<th>RMSE of azimuth $\phi$ [°]</th>
</tr>
</thead>
<tbody>
<tr>
<td>62.5</td>
<td>PWD</td>
<td>112.06</td>
</tr>
<tr>
<td></td>
<td>MUSIC</td>
<td>108.38</td>
</tr>
<tr>
<td></td>
<td>DML</td>
<td>85.54</td>
</tr>
<tr>
<td></td>
<td>SML</td>
<td>90.35</td>
</tr>
<tr>
<td></td>
<td>WSF</td>
<td>85.53</td>
</tr>
</tbody>
</table>

This is the author’s version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication. The final version of record is available at http://dx.doi.org/10.1109/TASLP.2015.2439573
show that the ML methods and WSF have better performance than MUSIC or PWD in the studied case of partially correlated or coherent reflections. They are able to correctly estimate the direction of several reflections with a much higher resolution than MUSIC or PWD. Moreover, an example analysis of a real situation showed that the ML methods and WSF allow a much more accurate analysis of the room acoustics than MUSIC or PWD, since frequency analysis is possible. Moreover, real data experiments showed that the stochastic ML performs worse than the deterministic one or WSF, due to a lower convergence probability. Therefore, DML or WSF should be preferred in the analysis of reflections. In addition, WSF is computationally more light, which makes it the best approach for the analysis of reflections out of the five methods tested in this paper.

The results of this paper show that ML methods can be applied to investigate acoustics of rooms, such as, concert halls [4], in more detail than ever before. In addition, the studied ML methods can be applied to obtain a more detailed synthesis of room acoustics with spatial sound reproduction methods.

REFERENCES


Copyright (c) 2015 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.


Sakari Tervo was born in Kuopio, Finland, in 1983. He received a M.Sc. degree in audio signal processing from Tampere University of Technology in 2006 and a D.Sc. degree in the field of acoustic signal processing from Aalto University in 2012. He has been a visiting researcher in Philips Research, in the Netherlands, in 2007, and in the University of York in 2010. Currently, he works as a post-doctoral researcher in Aalto University.

Archontis Politis obtained his M.Eng. degree in Civil Engineering at Aristotle’s University of Thessaloniki, Greece, and his M.Sc. degree in Sound & Vibration Studies at ISVR, University of Southampton, UK, in 2006 and 2008 respectively. From 2008 to 2010 he worked as a graduate acoustic consultant at Arup Acoustics, Glasgow, UK, and as a researcher in a joint collaboration between Arup Acoustics and the Glasgow School of Arts, on interactive auralization of architectural spaces using 3D sound techniques. Currently, he is pursuing a doctoral degree in the field of parametric spatial sound recording and reproduction.