Blind Localization of Early Room Reflections Using Phase Aligned Spatial Correlation

Tom Shlomo, *Student Member, IEEE*, and Boaz Rafaely, *Senior Member, IEEE*

***Abstract*—Blind estimation of the direction of arrival (DOA) and delay of room reflections from reverberant sound may be useful for a wide range of applications. However, due to the high temporal and spatial density of early room reflections and their low power compared to the direct sound, existing methods can only detect a small number of reflections. This paper presents PHALCOR (PHase ALigned CORrelation), a novel method for blind estimation of the DOA and delay of early reflections of a single source in a room that overcomes the limitations of existing solutions. PHALCOR is based on a signal model in which the reflection signals are explicitly modeled as delayed and scaled copies of the direct sound. A phase alignment transform of the spatial correlation matrices is proposed; this transform can separate reflections with different delays, enabling the detection and localization of reflections with similar DOAs. It is shown that the DOAs and delays of the early reflections can be estimated by separately analysing the left and right singular vectors of the transformed matrices using sparse recovery techniques. An extensive simulation study of a speaker in a reverberant room, recorded by a spherical array, demonstrates the effectiveness of the proposed method.**

***Index Terms*—Direction-of-arrival estimation, MUSIC, spherical array, room reflections, sparse recovery.**

I. Introduction

Estimation of the direction of arrival (DOA) and delay of room reflections is useful for many tasks in signal processing, such as speech enhancement and dereverberation [1], [2], source separation [3], optimal beamforming [4] and room geometry inference [5]. The early reflections have a key role in sound perception, as they can improve speech intelligibility and listener envelopment. They are also related to the impression of source width, loudness and distance [6], [7]. Therefore, exploitation of the early reflections can be beneficial in parametric spatial audio methods and spatial audio coding [8], [9].

Existing methods for the estimation of the parameters of early reflections can be categorized as blind and non-blind. Non-blind methods, operate on room impulse response signals, or, alternatively, assume that an anechoic recording of the sound source is available. Blind methods operate on microphone signals directly, and assume that no other information is available, as is often the case in many real world applications. This work focuses on blind estimation.

Spatial filtering, i.e., beamforming, can be utilized to blindly estimate the DOAs of the early reflections, as well as to separate reflection signals from the direct sound, which enables delay estimation using cross-correlation analysis [5]. However, when arrays of practical size are used, the spatial resolution achieved by beamformers is often insufficient, as the spatial density of early reflections can be very high [10]. Subspace methods, such as MUSIC or ESPRIT [11]–[14], can often provide higher resolution than beamformers. However, these methods require the sources to be uncorrelated, while in the case of early reflections, since all sources are delayed copies of the direct sound, reflected narrowband signals are highly correlated. Frequency smoothing is a common method to decorrelate source signals, enabling the application of subspace methods. However, frequency smoothing cannot decorrelate reflections that have similar delays. Furthermore, subspace methods typically require an estimation of the number of sources, which is a challenging task when the amplitudes of the sources greatly vary, as in the case of early reflections. Also, these methods require that the number of microphones is larger than the number of significant reflections, which can limit the number of detected reflections. By formulating the problem as an under-determined linear system, sparse recovery methods can also be utilized for the localization of early reflections; these methods attempt to find the smallest subset of sources that explains the measured signals [15]. In general, the performance of such methods depends on both the properties of the system (usually referred to as the dictionary), and the sparsity of the solution. Since for a given array geometry the dictionary remains fixed, the performance improves as the actual number of sources is reduced, and, in practice only the first few reflections are recoverable with practical arrays. Another type of methods that can localize correlated sources is based on modeling the source signals as deterministic unknowns. However, these methods often require non-linear optimization that is difficult to initialize, and model order selection, or, alternatively, peak selection in a steered response map. The latter, similarly to the beamforming approach, suffers from poor spatial resolution [11], [16]–[20]. Furthermore, since the early reflections signals are related to the direct sound, modeling each as a separate unknown signal does not fully exploit the problem structure, unnecessarily increasing the number of unknown parameters. A common limitation of the methods mentioned above is the use of a multiple source model that does not distinguish between sources with the same DOA. In summary, due to the

Manuscript received July 7, 2020; revised January 7, 2021; accepted February 1, 2021. Date of publication February 5, 2021; date of current version February 23, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Pia Addabbo. This work was supported by Facebook Reality Labs. *(Corresponding author: Tom Shlomo.)*

The authors are with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel (e-mail: tomshlomo@gmail.com; br@bgu.ac.il).

Digital Object Identifier 10.1109/TSP.2021.3057495

challenging nature of this task, and the limitations of current methods, no adequate solution seems to be available for blind estimation of the DOA and delay of early room reflections.

This paper presents PHALCOR (PHase ALigned CORrelation), a novel method for blind estimation of the DOA and delay of early reflections of a single source in a room. The proposed approach utilizes the inherent structure of early reflections - they are delayed and attenuated copies of the direct sound. More specifically, we use the property that the narrowband correlation between a source and its reflections has a phase that is linear in frequency to construct a transform that can separate reflections with different delays, which also enables the detection of multiple reflections from the same direction. Since the number of reflections with similar delays is usually small, the DOAs of reflections with similar delays are estimated using orthogonal matching pursuit (OMP), a sparse recovery technique. A simulation study demonstrates the performance of PHALCOR, in particular its ability to accurately detect a large number of reflections. Initial results of this work, with a simplified method and a much-reduced theoretical and performance analysis, is presented in [21], also submitted for publication as a conference paper.

The rest of this paper is organized as follows. Section II presents the necessary mathematical background on the plane wave amplitude density function and its spherical Fourier transform. Section III presents the system model. Section IV presents the theoretical foundations of the proposed method, while Section V describes the proposed algorithm. A simulation study and conclusions are presented in Sections VI and VII, respectively.

II. Mathematical Background

A. Notation

The notation used in the paper is presented briefly in this section. Lowercase boldface letters denote vectors, and uppercase boldface letters denote matrices. The k, l entry of a matrix $\mathbf{A}$ is denoted by $[\mathbf{A}]_{k,l}$. The complex conjugate, transpose, and conjugate transpose are denoted by $(\cdot)^*$, $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{H}}$, respectively. The Euclidean norm of a vector is denoted by $\|\cdot\|$. The outer-product of two vectors $\mathbf{a}$ and $\mathbf{b}$ is the matrix $\mathbf{a}\mathbf{b}^{\mathrm{H}}$. The imaginary unit is denoted by i.

$\mathbb{S}^2$ denotes the unit sphere in $\mathbb{R}^3$. The symbol $\Omega \in \mathbb{S}^2$ represents a direction in 3D-space, i.e., a pair of azimuth-elevation angles. $\angle(\Omega, \Omega') \triangleq \arccos(\Omega \cdot \Omega')$ is the angle between directions Ω and Ω'. "$\cdot$" is the dot product in $\mathbb{R}^3$.

B. Sound Field Representation Using Plane Wave Amplitude Density

Consider a sound field composed of M plane waves with amplitudes $a_1(f), \ldots, a_M(f)$ at frequency f, and directions $\Omega_1, \ldots, \Omega_M$. The sound pressure p at any point in space $\mathbf{x} \in \mathbb{R}^3$ can be formulated as follows:

$$p(f, \mathbf{x}) = \sum_{m=1}^{M} a_m(f) e^{ik\Omega_m \cdot \mathbf{x}}, \tag{1}$$

where $k = 2\pi f/c$ is the wave-number, and c is the speed of sound. When the sound field is composed of a continuum of plane waves, the summation is replaced by an integral over the entire sphere, and the amplitudes are replaced by the plane wave amplitude density (PWAD) $a(f, \Omega)$:

$$p(f, \mathbf{x}) = \int_{\mathbb{S}^2} a(f, \Omega) e^{ik\Omega \cdot \mathbf{x}} d\Omega.$$

For a fixed frequency, the PWAD is a function on the unit sphere. As such, it is possible to describe it by its spherical Fourier transform (SFT) coefficients [22]:

$$a_{n,m}(f) \triangleq \int_{\mathbb{S}^2} a(f, \Omega) \left[Y_n^m(\Omega)\right]^* d\Omega, \tag{2}$$

where Y_n^m is the order-n and degree-m spherical harmonic. The SFT of the PWAD can be used to represent the sound pressure as follows [22]:

$$p(f, \mathbf{x}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi i^n j_n(kr) Y_n^m(\Omega) a_{n,m}(f), \tag{3}$$

where $r = \|\mathbf{x}\|$, $\Omega = \mathbf{x}/r$, and j_n is the n'th order spherical Bessel function of the first kind.

Equation (3) can be well approximated by truncating the infinite sum to order $N = \lceil kr \rceil$ [22]. A microphone array can be used to estimate the coefficients of the SFT of the PWAD with order less than or equal to N, by inverting the (truncated) linear transformation (3), a process known as plane wave decomposition. The existence and stability of the inverse transform depend on the frequency f and physical properties of the (typically spherical) microphone array. Further details can be found in [22]. The resulting signals are stacked in a vector of length $(N+1)^2$ as follows:

$$\mathbf{a_{nm}} \triangleq \left[a_{0,0}, a_{1,-1}, a_{1,0}, a_{1,1}, \ldots, a_{N,N}\right]^{\mathrm{T}}.$$

The rest of this paper is described in terms of the SFT of the PWAD. Processing and analysis in this domain offer several advantages. First, the PWAD provides a description of the soundfield that is independent of the microphone array. Second, the steering vectors, i.e., the response to a plane-wave from a given direction, are frequency independent. The steering vector $\mathbf{y}(\Omega)$ is given by [22]:

$$\mathbf{y}(\Omega) \triangleq \frac{\sqrt{4\pi}}{N+1} \left[Y_0^0(\Omega), Y_1^{-1}(\Omega), \ldots, Y_N^N(\Omega)\right]^{\mathrm{H}}. \tag{4}$$

The constant $\frac{\sqrt{4\pi}}{N+1}$ was chosen for convenience such that $\|\mathbf{y}(\Omega)\| = 1$.

III. System Model

This section presents the system model used in the paper. Consider a sound field comprised of a single source in a room, with a frequency domain signal $s(f)$, and a DOA Ω_0, relative to a measurement point in the room. As the sound from the source propagates in the room, it is reflected from the room boundaries. The k'th reflection is modeled as a separate source with DOA Ω_k and signal $s_k(f)$, which is a delayed and scaled copy of the source signal [23]:

$$s_k(f) = \alpha_k e^{-i2\pi f \tau_k} s(f), \tag{5}$$

where τ_k is the delay relative to the direct sound, and α_k is the scaling factor. τ_0 and α_0 are accordingly normalized to 0 and 1, respectively. It is assumed that the delays are sorted such that $\tau_{k-1} \leq \tau_k$.

Let $\mathbf{a_{nm}}(f)$ denote the vector of the SFT coefficients of the PWAD, up to order N, as a function of frequency. Assuming the sources are in the far field, $\mathbf{a_{nm}}(f)$ is described by the following model:

$$\mathbf{a_{nm}}(f) = \mathbf{Y}\mathbf{s}(f) + \mathbf{n}(f), \tag{6}$$

where:

$$\mathbf{s}(f) \triangleq [s_0(f), \ldots, s_K(f)]^{\mathrm{T}} \tag{7}$$

$$\mathbf{Y} \triangleq [\mathbf{y}(\Omega_0), \ldots, \mathbf{y}(\Omega_K)], \tag{8}$$

$\mathbf{n}(f)$ includes noise and late reverberation terms, and K is the number of early reflections. Note that although both the model and the proposed method can be generalized to include near-field sources, this generalization may require near-field steering vectors and some information or estimation of source distances, which could relate to source delays. Nevertheless, the near-field steering vectors typically become useful for sources very close to the array (see, e.g. [24]), in which case room reflection may be negligible. Therefore, we proceed with the far-field assumption, valid for compact microphone arrays.

Let $\mathbf{R}(f)$ denote the spatial correlation matrix (SCM) at frequency f:

$$\mathbf{R}(f) \triangleq \mathbb{E}\left[\mathbf{a_{nm}}(f)\mathbf{a_{nm}}(f)^{\mathrm{H}}\right]. \tag{9}$$

Substituting Eq. (6) into Eq. (9), and assuming $\mathbf{n}(f)$ and $s(f)$ are uncorrelated, yields:

$$\mathbf{R}(f) = \mathbf{Y}\mathbf{M}(f)\mathbf{Y}^{\mathrm{H}} + \mathbf{N}(f), \tag{10}$$

where:

$$\mathbf{N}(f) \triangleq \mathbb{E}\left[\mathbf{n}(f)\mathbf{n}(f)^{\mathrm{H}}\right] \tag{11}$$

$$\mathbf{M}(f) \triangleq \mathbb{E}\left[\mathbf{s}(f)\mathbf{s}(f)^{\mathrm{H}}\right]. \tag{12}$$

IV. Phase Alignment of the Spatial Correlation Matrices

PHALCOR is based on a phase alignment transformation of the SCM. This section presents the definition and properties of this transformation.

A. Motivation

Before presenting the mathematical details, we first provide some motivation for this transformation.

Equation (10) can be rewritten as:

$$\mathbf{R}(f) = \sum_{k=0}^{K}\sum_{k'=0}^{K} [\mathbf{M}(f)]_{k,k'}\, \mathbf{y}(\Omega_k)\mathbf{y}(\Omega_{k'})^{\mathrm{H}} + \mathbf{N}(f). \tag{13}$$

It is apparent from equation (13) that neglecting $\mathbf{N}$, the matrix $\mathbf{R}$ is a mixture of the outer products of the steering vectors of the sources. The mixing coefficients are the entries of $\mathbf{M}$, and therefore it is henceforth referred to as the mixing matrix. Note that the mixing coefficients are frequency dependent, but the steering vectors are not.

Suppose that the k, k' entry of $\mathbf{M}(f)$ has a dominant magnitude, relative to all other entries. This leads to:

$$\mathbf{R}(f) \approx c\mathbf{y}(\Omega_k)\mathbf{y}(\Omega_{k'})^{\mathrm{H}} + \mathbf{N}(f) \tag{14}$$

for some $c \in \mathbb{C}$. Intuitively, estimating Ω_k and $\Omega_{k'}$ in such a case is easier than in the general case. However, according to (12), there may not be a dominant entry in $\mathbf{M}$. Not only is it Hermitian, but also the magnitudes of its entries are products of amplitudes between pairs of two sources. Assuming the amplitude of the direct sound may be dominant, the 0, 0'th entry that corresponds to the direct sound only may indeed be dominant. However, this is not helpful for localizing the early reflections. The processing presented below is designed to enhance specific entries in $\mathbf{M}$, so that specific reflections can be more easily localized.

B. Phase Aligned Spatial Correlation

We define the following matrix, which we call the phase aligned SCM:

$$\bar{\mathbf{R}}(\tau, f) \triangleq \sum_{j=0}^{J_f-1} w_j \mathbf{R}(f + j\Delta f) e^{i2\pi\tau j\Delta f}, \tag{15}$$

where $\tau \geq 0$, Δf is the frequency resolution, and J_f is an integer parameter representing the overall number of frequency points. $w_0, \ldots, w_{J_f-1}$ are non-negative weights. Note that when $\tau = 0$ and $w_j = 1$, $\bar{\mathbf{R}}$ is identical to the SCM obtained by frequency smoothing. The matrices $\bar{\mathbf{N}}(\tau, f)$ and $\bar{\mathbf{M}}(\tau, f)$ are similarly defined by replacing $\mathbf{R}$ in (15) with $\mathbf{N}$ and $\mathbf{M}$, respectively, such that:

$$\bar{\mathbf{R}}(\tau, f) = \sum_{k=0}^{K}\sum_{k'=0}^{K} \left[\bar{\mathbf{M}}(\tau, f)\right]_{k,k'} \mathbf{y}(\Omega_k)\mathbf{y}(\Omega_{k'})^{\mathrm{H}} + \bar{\mathbf{N}}(\tau, f). \tag{16}$$

Similarly to Eq. (13), Eq. (16) presents $\bar{\mathbf{R}}$ as a mixture of outer-products of pairs of sources' steering vectors. Next, it is proven that for a fixed f, the k, k' entry of $\bar{\mathbf{M}}(\tau, f)$ is maximized for $\tau = \tau_k - \tau_{k'}$. We begin the proof by deriving an explicit expression for the absolute value of the entries of $\bar{\mathbf{M}}$ for an arbitrary τ:

$$\begin{aligned}
&\left|\left[\bar{\mathbf{M}}(\tau, f)\right]_{k,k'}\right| \\
&= \left|\sum_{j=0}^{J_f-1} [\mathbf{M}(f_j)]_{k,k'}\, w_j e^{i2\pi\tau j\Delta f}\right| \\
&= \left|\sum_{j=0}^{J_f-1} \mathbb{E}\left[s_k(f_j)s_{k'}(f_j)\right] w_j e^{i2\pi\tau j\Delta f}\right| \\
&= \left|\sum_{j=0}^{J_f-1} \alpha_k\alpha_{k'}^{*}\sigma_s^2(f_j) w_j e^{i2\pi j\Delta f(\tau-(\tau_k-\tau_{k'}))}\right|,
\end{aligned} \tag{17}$$

TABLE I
PARAMETERS FOR THE NUMERICAL EXAMPLE

k	τ_k (ms)	α_k
0	0	1
1	2	0.7
2	6	0.5

where $f_j = f + j\Delta f$ and

$$\sigma_s^2(f_j) \triangleq \mathbb{E}\left[|s(f_j)|^2\right]. \tag{18}$$

The second equality in Eq. (17) is due to the definition of $\mathbf{M}(f)$ in Eq. (12), while the third equality is due to Eq. (5). Now, by a simple use of the triangle-inequality:

$$\begin{aligned}\left|\left[\bar{\mathbf{M}}(\tau, f)\right]_{k,k'}\right| &\leq \sum_{j=0}^{J_f-1} \left|\alpha_k \alpha_{k'}^* \sigma_s^2(f_j) w_j\right| \\ &= \left|\left[\bar{\mathbf{M}}(\tau_k - \tau_{k'}, f)\right]_{k,k'}\right|. \end{aligned} \tag{19}$$

The last equality is true since σ_s^2 and w_j are non-negative. Along with (16), this result implies that among all possible delays τ, it is the delay between two sources that maximizes the contribution of the outer product of their steering vectors to $\bar{\mathbf{R}}(\tau, f)$. This observation is at the core of our method. To better understand its implications, consider the following special case.

C. Special Case: White Source Signal

In this subsection the source signal is assumed to be white, such that $\sigma_s^2(f)$ is constant in f. The weights w_j are all set to 1. Equation (17) can thus be further simplified:

$$\begin{aligned}\left|\left[\bar{\mathbf{M}}(\tau, f)\right]_{k,k'}\right| &= \sigma_s^2 \left|\alpha_k \alpha_{k'}^* \sum_{j=0}^{J_f-1} e^{i2\pi j \Delta f(\tau - (\tau_k - \tau_{k'}))}\right| \\ &= \sigma_s^2 \left|\alpha_k \alpha_{k'}^* D_{J_f}\left(\Delta f(\tau - (\tau_k - \tau_{k'}))\right)\right|, \end{aligned} \tag{20}$$

where:

$$D_n(x) \triangleq \begin{cases} n & x \in \mathbb{Z} \\ \frac{\sin(\pi n x)}{\sin(\pi x)} & x \notin \mathbb{Z}. \end{cases} \tag{21}$$

$D_n(x)$ often arises in Fourier analysis, and is related to the Dirichlet kernel. It has a sinc-like behavior, with a main lobe centered around $x = 0$, and a null-to-null width of $2/n$. Correspondingly, $|[\bar{\mathbf{M}}(\tau, f)]_{k,k'}|$ has a main lobe centered at $\tau = \tau_k - \tau_{k'}$, and a width of $2(J_f \Delta f)^{-1}$. Therefore, J_f determines the temporal resolution, which affects the ability to separate reflections with different delays. This result can be used as a guideline for choosing J_f. Note that the same analysis is valid for non white signals, if the weights satisfy $w_j \propto 1/\sigma_s^2(f_j)$.

Next we consider a numerical example where there are $K = 2$ reflections, with parameters summarized in Table I. Fig. 1(a) presents the entries of the untransformed mixing matrix $\mathbf{M}(f)$. Since the signal is white, the k, k' entry is a complex sinusoidal of the form $\sigma_s^2 \alpha_k \alpha_{k'}^* e^{i2\pi f(\tau_k - \tau_k')}$ (see Eqs. (5) and (12)). Its real and imaginary parts are added to the k and k' axes, respectively, for the purpose of illustration of the complex function. This illustration demonstrates that as the delay between two sources is decreased, the period of the corresponding entry, as a function of f, increases. At the extreme, the delay between a source and itself is zero, and so the diagonal entries are constant in frequency. Fig. 1(c) presents the absolute value of the entries of $\mathbf{M}(f)$. Note that the absolute value does not depend on f in this case. It is evident that $\mathbf{M}(f)$ is not sparse.

Fig. 1(b) presents the entries of the phase aligned mixing matrix $\bar{\mathbf{M}}(\tau, 0)$, where $\Delta f J_f = 2000$ Hz. In this plot, only the absolute value of the matrix entries is shown, and is added to the k-axis for the purpose of illustration. Figs. 1(d) to 1(g) show several cross sections of Fig. 1(b), i.e., the absolute values of the entries of $\bar{\mathbf{M}}(\tau, 0)$ for selected values of τ. When τ is equal to a delay between two sources, which is the case in Figs. 1(e)–(g), $\bar{\mathbf{M}}(\tau, 0)$ is approximately sparse, and the most dominant entry is the one corresponding to the two sources. For other values of τ, all entries of $\bar{\mathbf{M}}(\tau, 0)$ are relatively small. For $\tau = 0$, $\bar{\mathbf{M}}(\tau, 0)$ is approximately a diagonal matrix. Since in that case the off diagonal entries can be interpreted as correlations between different sources, this demonstrates the fact that frequency smoothing ($\tau = 0$) performs decorrelation of the sources. If the reflections had the same delay, $\bar{\mathbf{M}}(0, 0)$ would contain dominant off diagonal entries, and frequency smoothing would have failed to decorrelate the sources. Furthermore, while for $\tau = 0$ all 3 sources are dominant simultaneously, for values of τ that correspond to delays between sources, only a subset of the sources are dominant.

D. Signal-Informed Weights Selection

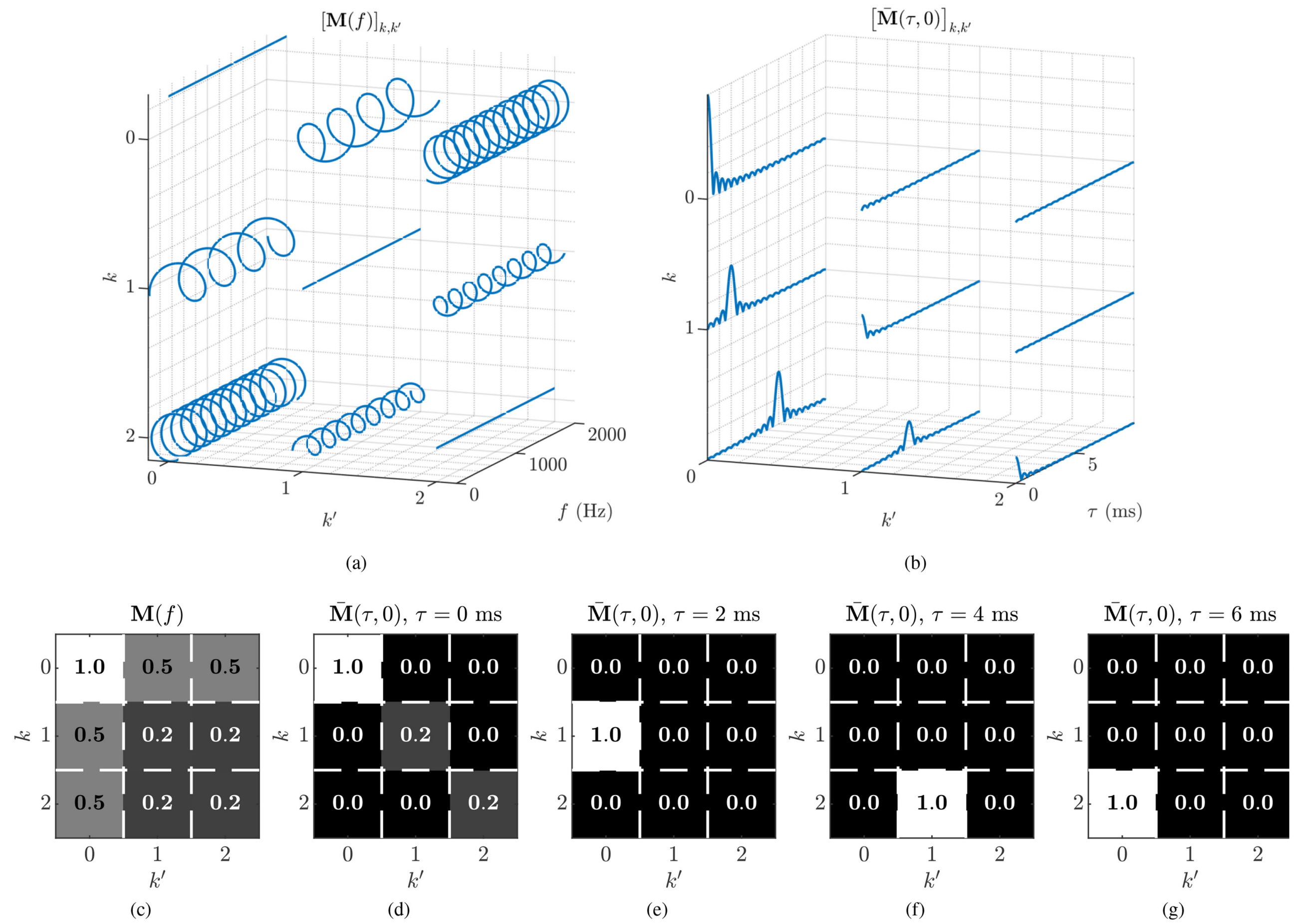
The analysis presented in the previous subsection shows that if the weights $\{w_j\}_j$ are inversely proportional to $\sigma_s^2(f_j)$, the phase alignment transform can effectively separate reflections with different delays. As $\sigma_s^2(f)$ is usually unknown, it must be estimated from the data. A very coarse, yet simple, estimate is given by the trace of $\mathbf{R}(f)$. By neglecting $\mathbf{N}$ in Eq. (13) and substituting Eqs. (12), (5) and (18), we get:

$$\begin{aligned} tr\left(\mathbf{R}(f)\right) &\approx \sum_{k,k'} [\mathbf{M}(f)]_{k,k'} \operatorname{tr}\left(\mathbf{y}(\Omega_k)\mathbf{y}(\Omega_{k'})^{\mathrm{H}}\right) \\ &= \sum_{k,k'} \mathbb{E}\left[s_k(f) s_{k'}^*(f)\right] \mathbf{y}(\Omega_{k'})^{\mathrm{H}} \mathbf{y}(\Omega_k) \\ &= \sigma_s^2(f) \sum_{k,k'} e^{-i2\pi f(\tau_k - \tau_{k'})} \alpha_k \alpha_{k'}^* \mathbf{y}(\Omega_{k'})^{\mathrm{H}} \mathbf{y}(\Omega_k) \\ &= \sigma_s^2(f) \sum_k |\alpha_k|^2 \left(1 + b_k(f)\right), \end{aligned} \tag{22}$$

where:

$$b_k(f) \triangleq 2\Re\left(\sum_{k'>k} \frac{\alpha_{k'}}{\alpha_k} e^{i2\pi f(\tau_k - \tau_{k'})} \mathbf{y}(\Omega_k)^{\mathrm{H}} \mathbf{y}(\Omega_{k'})\right) \tag{23}$$

and $\Re$ returns the real part of a complex scalar. We argue that b_k is typically small in comparison to 1, since usually the amplitudes decay rapidly. Furthermore, when two reflections have similar



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Fig. 1. (a) Entries of the unprocessed mixing matrix $[\mathbf{M}(f)]_{k,k'}$ for different values of f, with the real and imaginary part of the entries added to the k and k' axes for the purpose of illustrating the complex function. (b) Entries of the phase aligned mixing matrix $[\bar{\mathbf{M}}(\tau,0)]_{k,k'}$ for different values of τ, with the absolute value of the entries added to the k axis for the purpose of illustration. (c) The absolute values of the entries of the mixing matrix $[\mathbf{M}(f)]_{k,k'}$. (d)–(g) The absolute values of the entries of the phase aligned mixing matrix $[\bar{\mathbf{M}}(\tau,0)]_{k,k'}$ for different values of τ. In (c)–(g) the entries are normalized such that the maximum value is 1.

amplitudes, it is usually the case that they have very different DOAs, which implies that the inner product of their steering vectors is small [22].

The weights could have another role. Eq. (20) suggests that even if the weights are inversely proportional to σ_s^2, reflections with delays other than τ may still be dominant in $\bar{\mathbf{M}}(\tau,f)$, as D_n have strong side lobes. The side lobe levels can be reduced by introducing a window function, at the expense of increasing the width of the main lobe [25].

E. Rank 1 Approximation of Phase Aligned SCM

In the following, the dependence on the frequency f is omitted for brevity. It is important to note that there is no direct access to $\bar{\mathbf{M}}(\tau)$; it is only indirectly observed through $\bar{\mathbf{R}}(\tau)$ as given by Eq. (16). When the proposed transformation succeeds in enhancing a single entry in $\bar{\mathbf{M}}(\tau)$, $\bar{\mathbf{R}}(\tau)$ will be dominated by a single outer product of steering vectors $\mathbf{y}(\Omega)\mathbf{y}(\Omega')^{\mathrm{H}}$. As an outer product is a rank 1 matrix, it is expected that the rank 1 approximation of $\bar{\mathbf{R}}(\tau)$ would perform denoising, i.e., reduce the contribution of $\bar{\mathbf{N}}(\tau)$. The optimal rank 1 approximation (in the least squares sense) of $\bar{\mathbf{R}}(\tau)$ is denoted by $\bar{\mathbf{R}}_1(\tau)$, and is given by truncating its singular value decomposition (SVD):

$$\bar{\mathbf{R}}_1(\tau) = \sigma_\tau \mathbf{u}_\tau \mathbf{v}_\tau^{\mathrm{H}},$$

where σ_τ denotes the first (largest) singular value of $\bar{\mathbf{R}}(\tau)$, and $\mathbf{u}_\tau$ and $\mathbf{v}_\tau$ denote corresponding left and right singular vectors, respectively. Besides performing denoising, the SVD also separates the two steering vectors, as $\mathbf{u}_\tau$ and $\mathbf{v}_\tau$ are approximately equal (up to phase) to $\mathbf{y}(\Omega)$ and $\mathbf{y}(\Omega')$, respectively. If there are several reflections with the same delay τ, $\bar{\mathbf{R}}(\tau)$ is still approximately of rank 1 since the dominant entries in $\bar{\mathbf{M}}(\tau)$ all appear at the same column. However, in that case $\mathbf{u}_\tau$ is not a single steering vector, but rather a linear combination of the steering vectors of the reflections with delay τ.

V. Algorithm Description

This section describes the PHALCOR algorithm for estimating the DOAs and delays of the early reflections. The algorithm is based on the analysis of the first singular vectors of the phase aligned SCM $\bar{\mathbf{R}}(\tau,f)$ that was presented in Section IV. The

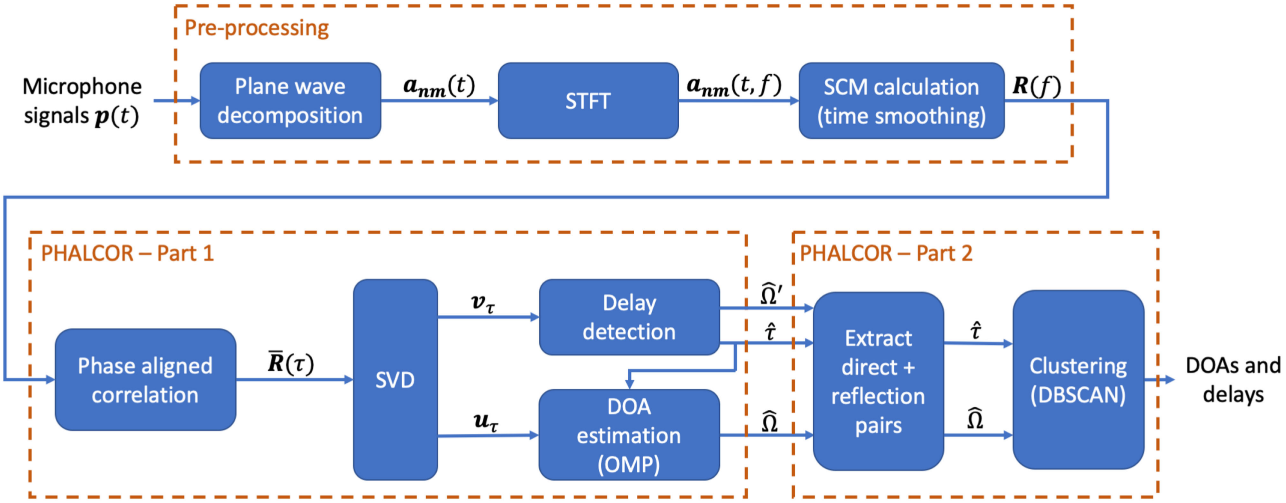


Fig. 2. Block diagram of the proposed method.

analysis in Section IV requires the plane wave decomposition signals to be in the frequency domain. In practice, these are approximated using the short time Fourier transform (STFT) which enables localized analysis in both time and frequency. It is assumed that the window length of the STFT is sufficiently larger than τ_K, such that the multiplicative transfer function (MTF) approximation in the STFT is applicable for Eq. (5) [26]. Note that in the following, τ is the parameter of the phase alignment transform as in Eq. (15), and should not be confused with the time index of the STFT.

The algorithm is performed in two parts. In the first and main part, local time frequency estimates of reflection delays and DOAs are computed; this is performed separately on different regions in the time frequency domain. In the second part, cluster analysis is performed on the local estimates to obtain global estimates that are more robust and accurate. Fig. 2 presents a block diagram of the proposed method.

A. Part 1: Local Time Frequency Estimations

The first, and main, part of the algorithm consists of three steps as described below.

1) Phase Alignment Transform: For each time-frequency bin, $\mathbf{R}$ is estimated by replacing the expectation in Eq. (9) with averaging across J_t adjacent bins in time. Then, $\bar{\mathbf{R}}(\tau, f)$ is calculated for $\tau = 0, \Delta\tau, \ldots, (J_\tau - 1)\Delta\tau$ using Eq. (15). $\Delta\tau$ dictates the delay estimation resolution, while J_τ dictates the maximal detectable delay of a reflection. The selection of these parameters is discussed in Section V-C, as well as a method to efficiently calculate $\bar{\mathbf{R}}$ using the fast Fourier transform (FFT). The weights are set using the method described in Section IV-D:

$$w_j = \frac{W_j}{\mathrm{tr}\left(\mathbf{R}(f_j)\right)}, \tag{24}$$

where W_j is the j'th sample of a Kaiser window of length J_f, with the β parameter set to 3 [25].

2) Delay Detection: The next step is to detect values of τ that are equal to a reflection's delay. Based on the analysis presented in Section IV, we suggest the detection of such values of τ by thresholding the following signal:

$$\rho(\tau) = \max_{\Omega' \in \mathbb{S}^2} \left| \mathbf{y}(\Omega')^{\mathrm{H}} \mathbf{v}_\tau \right|, \tag{25}$$

where $\mathbf{v}_\tau$ is a first right singular vector of $\bar{\mathbf{R}}(\tau)$. By the Cauchy-Schwartz inequality, since both $\mathbf{v}_\tau$ and $\mathbf{y}(\Omega)$ are unit vectors, $\rho(\tau) \in [0, 1]$ and is equal to 1 if and only if $\mathbf{v}_\tau$ is equal (up to phase) to a steering vector. The threshold is set empirically to $\rho_{\min} = 0.9$.

We denote by $\hat{\Omega}'(\tau)$ the direction that attains the maximum in Eq. (25):

$$\hat{\Omega}'(\tau) = \arg\max_{\Omega' \in \mathbb{S}^2} \left| \mathbf{y}(\Omega')^{\mathrm{H}} \mathbf{v}_\tau \right|. \tag{26}$$

When τ is equal to a reflection's delay, $\hat{\Omega}'(\tau)$ is an estimate of Ω_0 (the DOA of the direct sound). Note that when τ is equal to a delay between two reflections (and not a delay between a reflection and the direct sound), $\rho(\tau)$ may be high as well, leading to false detections. However, such detections are distinguishable from valid ones, as $\hat{\Omega}'(\tau)$ will be different from Ω_0; as detailed in Section V-B, the first step of part 2 discards such detections.

3) DOA Estimation: The next step is estimating the DOAs of the reflections. This step is performed separately for every τ selected in the previous step. Let $\mathbf{u}_\tau$ denote a first left singular vector of $\bar{\mathbf{R}}(\tau)$. According to the analysis presented in Section IV, $\mathbf{u}_\tau$ is approximately equal to a linear combination of the steering vectors of the reflections with a delay close to τ. If there is only a single such reflection, then its DOA can be estimated using a similar method to that of the direct sound (Eq. (26)). However, in practice there might be several reflections with similar delays. Since their number is expected to be quite small, we utilize sparse recovery to estimate the DOAs. Specifically, we aim to solve the following problem: Find the smallest set of directions $\hat{\Omega}_1, \ldots \hat{\Omega}_S$ and coefficients $x_1, \ldots x_S$, such that

$$\left\| \sum_{s=1}^{S} x_s \mathbf{y}(\hat{\Omega}_s) - \mathbf{u}_\tau \right\|^2 \leq \epsilon_{\mathbf{u}}, \tag{27}$$

where $\epsilon_{\mathbf{u}} \in (0, 1)$ is a predefined threshold, set experimentally to 0.4. In the context of sparse recovery, the set of vectors $\{\mathbf{y}(\Omega) : \Omega \in \mathbb{S}^2\}$ is known as the dictionary, and its elements are known as atoms. The optimization problem is computationally intractable and cannot be exactly solved in practice. Nevertheless, there has been extensive research on algorithms that find approximate solutions. In this paper we chose to apply the orthogonal matching pursuit (OMP) algorithm [27]. Although there are more sophisticated sparse recovery algorithms, we chose OMP for several reasons. First, it is simple, and has a low computational cost. Second, although originally designed for finite dictionaries, it is easily extended for our infinite dictionary case. Finally, it is especially suited for our problem by the following argument. Early reflections with similar delays usually have very different DOAs as they typically originate from different walls. If the angle between the DOAs is larger than π/N, the corresponding steering vectors are approximately orthogonal [22] and the projection step in OMP only removes the contribution of steering vectors of DOAs the have already been found, without affecting the rest.

The OMP algorithm is applied on $\mathbf{u}_\tau$ for every detected τ. Values of τ where the resulting S is larger than $S_{\max}$ are discarded. The value of $S_{\max}$ was empirically set to 3.

B. Part 2: Cluster Analysis

The input for this part is a list of the local estimates obtained from part 1. Each estimate is a triplet of the form $(\hat{\tau}, \hat{\Omega}, \hat{\Omega}')$, corresponding to the delay of a reflection, its DOA, and the DOA of the direct sound. The goals of this part are: first, to remove outliers, and second, to obtain global estimates for the DOAs and delays of the early reflections.

The first step, denoted in Fig. 2 by "Extract direct + reflection pairs," discards estimates where the angle between $\hat{\Omega}'$ and Ω_0, is larger than some predefined threshold, set empirically to 10 degrees. In general Ω_0 is not known; however, it can be estimated by selecting the peak in the histogram of $\hat{\Omega}'$.

Next, a clustering algorithm is applied to the remaining estimates. We chose the DBSCAN algorithm [28], as it does not require an initial estimate of the number of clusters, as the number of reflections is automatically estimated within the algorithm. Furthermore, DBSCAN can automatically detect outliers by not assigning them to any cluster. We note, however, that other clustering algorithms may provide improved performance, especially from a computational viewpoint, as the complexity of our implementation of DBSCAN is quadratic in the number of data points. Nevertheless, a comprehensive investigation of the clustering method is beyond the scope of this work. DBSCAN has two positive parameters ϵ and MINPTS, and operates as follows. Two points are defined as neighbors if the distance between them is less then ϵ. A core point is defined as a point that has MINPTS or more neighbors. A noise point is a non core point, for which none of its neighbors are core points. The algorithm iterates over all the points in the dataset, and assigns two points to the same cluster if one of them is a core point. Noise points are not assigned to any cluster.

The metric we used is the following:

$$d\left((\tau_a, \Omega_a), (\tau_b, \Omega_b)\right) = \sqrt{\left(\frac{\angle(\Omega_a, \Omega_b)}{\gamma_\Omega}\right)^2 + \left(\frac{\tau_a - \tau_b}{\gamma_\tau}\right)^2}, \tag{28}$$

where γ_Ω and γ_τ are normalization constants, set empirically to 15 degrees and 500 microseconds, respectively. As the metric is normalized, the parameter ϵ is simply set to 1. MINPTS is empirically set to 10 percent of the number of neighbors of the point that has the largest number of neighbors.

Finally, the global estimates are calculated for each cluster by averaging the local estimates of the points assigned to it. Note that while the local delay estimates are confined to a grid of resolution $\Delta\tau$, the global delay estimates, being averages of local estimates, are not. The fact that each DOA estimate has an associated delay, is a major advantage of our method, as it enables the separation of clusters even if they have similar DOAs.

C. Practical Considerations

1) Avoiding Redundant Processing: The information captured in $\bar{\mathbf{R}}$ contains contributions from a rectangular region in the time-frequency domain of size $J_t \times J_f$. Therefore, it is expected that the results of part 1 of the algorithm would be similar when applied to regions with a large overlap. To reduce computation time, the overlap between the time-frequency regions of part 1 can be reduced. For example, in the simulation study discussed below, an overlap of 87.5% in frequency was selected, such that each frequency is processed 8 times.

2) Acceleration Using the FFT: Part 1 of the algorithm requires the calculation of $\bar{\mathbf{R}}(\tau)$ for a grid of values of τ. Note that if $(\Delta\tau \cdot \Delta f)^{-1} \in \mathbb{N}$, then the sequence

$$\left(\bar{\mathbf{R}}\left(j\Delta\tau\right)\right)_{j=0}^{J_\tau - 1} \tag{29}$$

is equal (up to scaling and appropriate zero-padding) to the first J_τ terms of the inverse discrete Fourier transform (taken entry wise) of the sequence

$$\left(w_j \mathbf{R}\left(f + j\Delta f\right)\right)_{j=0}^{J_f - 1}, \tag{30}$$

so $\bar{\mathbf{R}}$ can be calculated efficiently for the grid of delays using the FFT. A further reduction in the computation time can be achieved using the following identity, obtained directly from Eq. (15) and from the fact that $\mathbf{R}$ is Hermitian:

$$\bar{\mathbf{R}}\left(\tau\right) = \bar{\mathbf{R}}\left(\Delta f^{-1} - \tau\right)^{\mathrm{H}}. \tag{31}$$

Thus, it is sufficient to perform the FFT on only the upper-triangular entries of $\mathbf{R}$.

3) Periodicity of the Phase Aligned SCM: It is apparent from Eq. (15) that $\bar{\mathbf{R}}(\tau)$ is periodic with respect to τ, with period Δf^{-1}. This periodicity does not introduce ambiguity in the delay estimation for the following reason. When the STFT window size is T, the frequency resolution of the STFT Δf satisfies $\Delta f \leq 1/T$. Therefore, reflections with delay τ larger than Δf^{-1} necessarily do not satisfy the MTF approximation criteria, since $\tau > T$. This analysis also shows that to avoid unnecessary calculations, J_τ should be chosen such that $J_\tau \Delta\tau < \Delta f^{-1}/2$.

4) Selecting the Parameters of the Phase Alignment Transform: The calculation of $\bar{\mathbf{R}}$ requires the selection of three parameters: J_f, $\Delta\tau$ and J_τ (see Section V-A1). The number of frequency bins J_f, should be chosen to be high enough such that the temporal resolution (given by $(\Delta f J_f)^{-1}$, see Section IV-C) is sufficient. For example, if $J_f \Delta f = 2000\,\mathrm{Hz}$, then the phase alignment transform can separate two reflections if their delays are spaced by more than $0.5\,\mathrm{ms}$. However, J_f cannot be set arbitrarily high. First, the frequency independence of steering vectors (see Section II-B) is in practice limited to a given band, depending on the geometry of the microphone array. Second, some of our model assumptions may only be valid for bands of limited width. For example, the linear phase assumption in Eq. (5) may, in practice, only hold within a local frequency region.

Once J_f has been determined, a convenient way to set $\Delta\tau$, the delay estimation resolution, is:

$$\Delta\tau = \frac{1}{M\Delta f}, \tag{32}$$

where M is an integer that satisfies $M \geq J_f$. This choice guarantees that $\Delta\tau \leq (J_f \Delta f)^{-1}$, and also that $(\Delta\tau \cdot \Delta f)^{-1} \in \mathbb{N}$, so the FFT can be used to calculate $\bar{\mathbf{R}}$. Increasing M beyond J_f would increase the resolution of delay estimation; however, it would also increase the computation time of the algorithm.

Finally, J_τ, the number of grid points over τ, should be chosen such that $(J_\tau - 1)\Delta\tau$, the maximal detectable delay, is sufficiently small relative to T, the window length of the STFT, such that the MTF criteria for Eq. (5) holds. From our experience, $(J_\tau - 1)\Delta\tau \approx T/10$ is sufficient.

5) Maximizing Over the Sphere: Both Eq. (25) and the OMP algorithm require maximizing functions of the form $f(\Omega) = |\mathbf{y}(\Omega)^{\mathrm{H}}\mathbf{x}|$ over the sphere. Note that this is equivalent to finding the maximum of a signal on the sphere whose SFT is given by $\mathbf{x}$. We use Newton's method to perform this maximization, with initialization obtained by sampling the sphere with a nearly uniform grid of 900 directions [29].

6) Computational Complexity: We focus in the analysis on part 1 of PHALCOR since the pre-processing steps (including plane wave decomposition, STFT, and time smoothing) and part 2 (cluster analysis) are standard and shared among many methods. Part 1 of PHALCOR is calculated independently for every selected time-frequency region, and therefore the total computation time grows linearly with the duration of the input signal. As the phase alignment transform can be calculated very efficiently using the FFT, the main bottlenecks here are the SVD, delay detection and DOA estimation.

While calculating the SVD of an SCM is common in many localization methods, it is usually calculated once for every selected time-frequency region. In PHALCOR, however, it is calculated J_τ (the size of the τ-grid) times for every selected time-frequency region. As J_τ controls the maximal detectable delay, there is a trade off between the maximal detectable delay and the computational complexity. By decreasing J_τ and increasing $\Delta\tau$ (τ-grid resolution), one can decrease run time without changing the maximal detectable delay. However, increasing $\Delta\tau$ comes at the cost of poor delay resolution.

The main bottleneck in the delay detection is the calculation of ρ and $\hat{\Omega}'$ (Eqs. (25) and (26)) which requires a global maximum search over the sphere. Again, such calculations are common for many localization methods (including MUSIC and beamforming maps); however, similar to the SVD, in PHALCOR this calculation is required for every τ on the grid. Similarly, since the OMP algorithm is applied for every detected delay, the computational complexity of the DOA estimation step also increases with the number of delays.

In summary, the key strength of PHALCOR - separating reflections of different delays, comes with a computational cost, as it is required to repeat the processing for each delay. Run time examples are presented in Section VI-C.

D. Relation to Other Methods

In this section we discuss some similarities between PHALCOR and other methods in array signal processing.

1) MUSIC and Frequency Smoothing: When $\tau = 0$ and $w_j = 1$, $\bar{\mathbf{R}}(\tau)$ is a frequency-smoothed SCM (as used for example in [30]). Frequency smoothing is a common procedure in source localization in the presence of reverberation, as it can can decorrelate signals, which is necessary for subspace methods such as MUSIC. Furthermore, $\bar{\mathbf{R}}(0)$ is positive semi-definite, and therefore its singular vectors are also eigenvectors, so $\hat{\Omega}'(0)$ is the estimate obtained by MUSIC if the signal subspace dimension is set to 1, and $\rho(0)$ is equivalent to the MUSIC spectrum magnitude at this direction. While the frequency smoothing goal is to decorrelate the sources, PHALCOR actually utilizes this correlation to enhance specific reflections.

2) L1-SVD: Another well known method for source localization that can address correlated sources is L1-SVD [31]. It is based on the observation that the first eigenvectors of the SCM are linear combinations of steering vectors. The DOAs are estimated by decomposing the eigenvectors of the SCM into a sparse combination of steering vectors. This is similar to our method, which decomposes a first left singular vector of the phase aligned SCM to a sparse linear combination of steering vectors. In general, the performance of sparse recovery methods improves as the vectors are more sparse. While in L1-SVD the sparsity is determined by the total number of reflections, in PHALCOR the sparsity is determined by the number of sources at a specific delay, which is significantly lower. Furthermore, like MUSIC, in L1-SVD the number of detectable sources is limited by the number of input channels ($(N+1)^2$ in our case). In PHALCOR, on the other hand, it is possible to detect more reflections than input channels, as each delay is processed independently.

3) Generalized Cross Correlation: The relations of PHALCOR to MUSIC and L1-SVD is related only to DOA estimation; however, PHALCOR is also related to delay estimation methods that are based on generalized cross correlation analysis [32]. It can be shown that the entries of $\bar{\mathbf{R}}(\tau)$ contain a generalized cross correlation between each pair of input channels, at lag τ. Although similar, there are some important distinctions between the two methods. While cross correlation analysis is typically used to estimate the delay between two signals that are observed

TABLE II
LIMITATIONS OF CURRENT METHODS AND THE WAY IN WHICH THESE ARE OVERCOME USING THE PROPOSED APPROACH

Limitation	**Details**	**Approach for solution**
Spatial resolution	Practical arrays have restricted spatial resolution, imposing a limit on the spatial separability of reflections.	$\mathbf{R}(f)$ is transformed to $\bar{\mathbf{R}}(\tau,f)$, making $\bar{\mathbf{M}}(\tau,f)$ sparse, and enhancing one or a few reflections relative to the rest, therefore overcoming the spatial resolution limits.
Coherent sources	Room reflections are coherent, see Eq. (5), and methods like MUSIC may fail, or may require frequency smoothing [33]. The latter process may only decorrelate reflections with different delays.	The proposed approach exploits coherency of reflection in an explicit signal model, and does not require de-correlation of source, therefore overcoming this limitation.
Reflections with similar DOA	Reflections with similar DOA and different delay may not be resolved by methods that rely on spatial separation, such as beamforming and MUSIC, and other methods as detailed in section I.	The phase alignment transformation leads to sparse $\bar{\mathbf{M}}(\tau,f)$, with reflections of different delay τ contributing to different sparse matrices $\bar{\mathbf{M}}(\tau,f)$, facilitating estimation of reflections with the same DOA but different delay.
More reflections than microphones	Current methods may require more microphones than reflections.	PHALCOR separates reflections into groups with similar delay. As each group is processed separately, the number of microphones limit the reflections per group, and not the number of reflections in total.

TABLE III
ROOM PARAMETERS USED IN THE SIMULATION STUDY

Room	Dimensions (m)	Reverberation Time (s)	Critical Distance (m)	Average Distance Between Source and Array (m)	Average Number of Reflections With Delay Smaller Than 20 ms
Small	$5 \times 4 \times 2.5$	0.6	0.5	1.7	52
Medium	$7 \times 5 \times 3$	0.8	0.7	2.5	31
Large	$10 \times 7 \times 3.5$	1.1	1	3.8	21

directly, PHALCOR aims to estimate the delay between multiple signals that are observed indirectly - each input channel is a linear combination of the delayed signals, given by the unknown steering matrix, which is estimated as well.

Table II summarizes the limitations of current methods and the way in which PHALCOR overcomes these limitations.

VI. SIMULATION STUDY

In this section, a simulation study is presented, demonstrating the performance of PHALCOR. First, a detailed analysis of the different steps of the algorithm is presented on a specific test case. Next, a Monte Carlo analysis is presented, demonstrating the robustness of PHALCOR.

A. Simulation Setup

The setup of the simulations, common to both the case study and the Monte Carlo study, is as follows. An acoustic scene that consists of a speaker and a rigid spherical microphone array in a shoe box room, was simulated using the image method [23]. The speech signal is a 5 seconds sample, drawn randomly from the TSP Speech Database [34]. The array has 32 microphones, and a radius of 4.2 cm (similar to the Eigenmike [35]), facilitating plane wave decomposition with spherical harmonics order $N = 4$. The microphone signals were sampled at 48 kHz. Sensor noise was added, such that the direct sound to noise ratio is 30 dB. The positions of the speaker and the array were chosen at random inside the room, with the restriction that the distance between each other, and to the room boundaries is no less than 1 m. Three different rooms sizes are considered. Their dimensions and several acoustic parameters, are presented in Table III.

B. Methodology

The signals recorded by the microphones were used to compute $\mathbf{a_{nm}}(f)$ as detailed in Section II-B. An STFT was applied to the PWAD coefficients signals using a Hanning window of 8192 samples, and an overlap of 75%. A frequency range of $[500, 5000]$ Hz was chosen for the analysis. The number of time bins used for averaging, J_t, was set to 6, while the number of frequency bins used for the phase alignment transform, J_F, was set such that $J_f \Delta f = 2000$ Hz. The delay resolution $\Delta\tau$ was set to 83.33 microseconds (equivalent to setting $M = 2048$ in Eq. (32)), while J_τ was chosen such that the maximal delay is 20 ms. With these parameters, PHALCOR, detailed in Section V, was applied to the simulated data. The values of the different hyper-parameters of PHALCOR, including $\rho_{\min}$, ϵ_μ, $S_{\max}$, γ_Ω and γ_τ, were set as detailed in Section V.

The MUSIC algorithm [30] was applied as a reference method for DOA estimation, by selecting the peaks in the MUSIC spectrum $\|\mathbf{y}(\Omega)^{\mathrm{H}}\mathbf{U}\|$, where $\mathbf{U}$ is a matrix whose columns are orthonormal eigenvectors that correspond to the L largest eigenvalues of the time and frequency smoothed SCM. The time and frequency smoothing parameters are the same as in PHALCOR. The dimension of the signal subspace L was determined using the SORTE method [36]. To reduce false positives, peaks for which the MUSIC magnitude height is lower than 0.75 are discarded. The local estimates are clustered using DBSCAN, to obtain global estimates. The delays of the detected reflections are estimated using the following method, which is similar to the one proposed in [5]. First, each reflection signal is estimated by solving Eq. (6) for $\mathbf{s}$ in the least squares sense. Then, the delay of the k'th reflections is estimated by selecting the maximum of the cross correlation values between s_k and s_0. Note that in

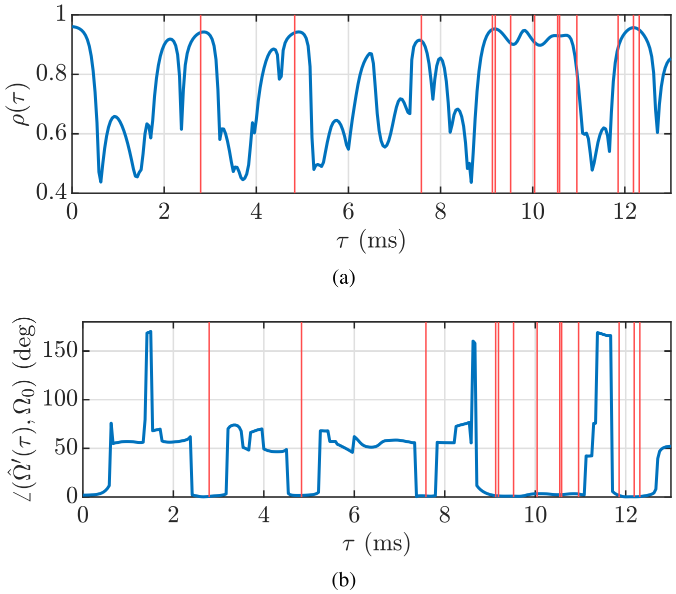


Fig. 3. (a) ρ as a function of τ (Eq. (25)). (b) The angle between $\hat{\Omega}'(\tau)$ (Eq. (26)) and true direction of the direct sound, as a function of τ. The red vertical lines correspond to the true delays of the reflections.

contrast to PHALCOR, the delays are estimated only after the clustering.

For both PHALCOR and the reference method, we consider a detected reflection as a true positive if its delay and DOA match simultaneously the delay and DOA of a true reflection. The matching tolerance was chosen to be 500 μs for the delay, and 15 degrees for the DOA. The probability of detection (PD) at a given maximal delay t is defined as the fraction of true positive detections of reflections r with a delay smaller than or equal to t, out of the total number of reflections with a delay smaller than or equal to t:

$$\mathrm{PD}_t \triangleq \frac{|\{r \in \mathcal{D} : r \text{ is true positive with delay} \leq t\}|}{|\{r \in \mathcal{GT} : r \text{ has delay} \leq t\}|},$$

where $\mathcal{GT}$ is the set of all ground truth reflections, and $\mathcal{D}$ is the set of all estimated reflections. The false positive rate (FPR) at a given maximal delay t is defined as the fraction of false positive detections with a delay smaller than or equal to t, out of the total number of detections with a delay smaller than or equal to t:

$$\mathrm{FPR}_t \triangleq \frac{|\{r \in \mathcal{D} : r \text{ is false positive with delay} \leq t\}|}{|\{r \in \mathcal{D} : r \text{ has delay} \leq t\}|}.$$

Here, $|\cdot|$ denotes the cardinality of the set.

C. Results of a Case Study

The test case presented in this section is of a female speaker in the medium sized room. There are $K = 31$ reflections with a delay less than 20 ms in this case.

Figs. 3(a) and 3(b) illustrate the delay detection routine, as detailed in Section V-A2. Fig. 3(a) shows $\rho(\tau)$ as a function of τ. Since $\rho(\tau)$ measures the similarity between $\mathbf{v}_\tau$, a first right singular vector of $\bar{\mathbf{R}}(\tau)$, and a steering vector (see Eq. (25)), it is high for values of τ that are close to a reflection's delay, indicated on the plot using red vertical lines. There are also values of τ that are not close to a reflection's delay, for which $\rho(\tau)$ is high. These correspond to delays between two reflections (as opposed to delays between a reflection and the direct sound). For example, the peak near $\tau = 2$ ms, corresponds to the delay between the second and third reflections, whose delays are about 3 ms and 5 ms, respectively. As discussed in Section V-A2, such cases may be identified by testing $\angle(\hat{\Omega}'(\tau), \Omega_0)$, the angle between the DOA of the steering vector that is most similar to $\mathbf{v}_\tau$, and the DOA of the direct sound. As shown in Fig. 3(b), $\angle(\hat{\Omega}'(\tau), \Omega_0)$ is small for values of τ which are close to a reflection's delay, and high otherwise. Therefore, false detections such as $\tau = 2\mathrm{ms}$, will be discarded during the first step of the second part of the algorithm.

Fig. 4 illustrates the process of DOA estimation, as detailed in Section V-A3. Each plot shows a different function on the sphere, which is projected onto the 2D page using the Hammer projection. In the top row, the function $|\mathbf{y}(\Omega)\mathbf{v}_\tau|$ is shown, where each column corresponds to a different value of τ. Recall that when τ equals a reflection's delay, we expect the direction that maximizes the response to be that of the direct sound. Indeed, as τ varies across columns, the location of the peak remains, and is equal to Ω_0, the DOA of the direct sound.

In the middle row, the function $|\mathbf{y}(\Omega)\mathbf{u}_\tau|$ is shown. It is similar to the top row, except that a first left singular vector is used instead of a right one. Recall that when τ is a reflection's delay, $\mathbf{u}_\tau$ is approximately equal to a linear combination of the steering vectors of reflections with delays of approximately τ. When the DOAs are sufficiently separated, they can be identified as peaks in $|\mathbf{y}(\Omega)\mathbf{u}_\tau|$. For τ_1 and τ_2, only one such peak is apparent, and its location matches the DOA of the corresponding reflection. When $\tau = \tau_4$, it is apparent that there are two dominant peaks, at directions Ω_4 and Ω_5. This is due to the fact that the 4th and 5th reflections have similar delays. Similarly, since the 8th and 9th reflections have similar delays, when $\tau = \tau_8$ the two peaks correspond to Ω_8 and Ω_9.

Fig. 4 demonstrates the effectiveness of PHALCOR in separating reflections from the direct sound, as well as reflections with different delays. This is in contrast to the MUSIC spectrum (shown ± the same figure), which shows only a few peaks, which are much less separable; as a result, fewer reflections are potentially identified.

Figs. 5 and 6 present the local estimates obtained with PHALCOR (as detailed in Section V-A) and the reference methods (as detailed in Section VI-B), respectively. It is apparent that, compared to the reference, PHALCOR is able to detect significantly more reflections. PHALCOR detected successfully 29 reflections, while the MUSIC based method could only detect 8 (not including the direct sound). Furthermore, Figs. 5 and 6 illustrate the advantage of simultaneously estimating DOA and delay for cluster analysis.

Finally, we note that while the performance of the proposed method is superior, the difference in computation time is quite significant: 303 seconds for the proposed method, and only 19 seconds for the reference method (as obtained using MATLAB 2020a, on a MacBook Pro 2019 with a 2.3 GHz 8-Core Intel Core i9 processor, 16 GB RAM).

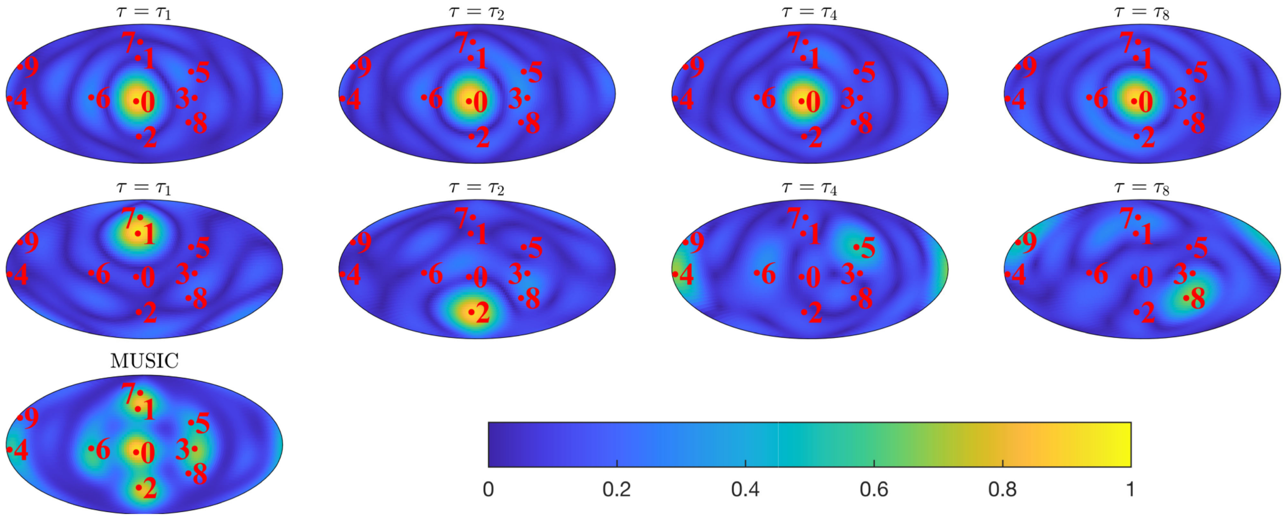


Fig. 4. Top row: $|\mathbf{y}(\Omega)^{\mathrm{H}}\mathbf{v}_\tau|$ as a function of Ω, for different values of τ. Middle row: $|\mathbf{y}(\Omega)^{\mathrm{H}}\mathbf{u}_\tau|$ as a function of Ω, for different values of τ. Bottom row: MUSIC spectrum. The red markers correspond to ground truth DOAs $\Omega_0, \ldots, \Omega_9$, with the numbers indicating the index. The Hammer projection is used to project the sphere onto the plane.

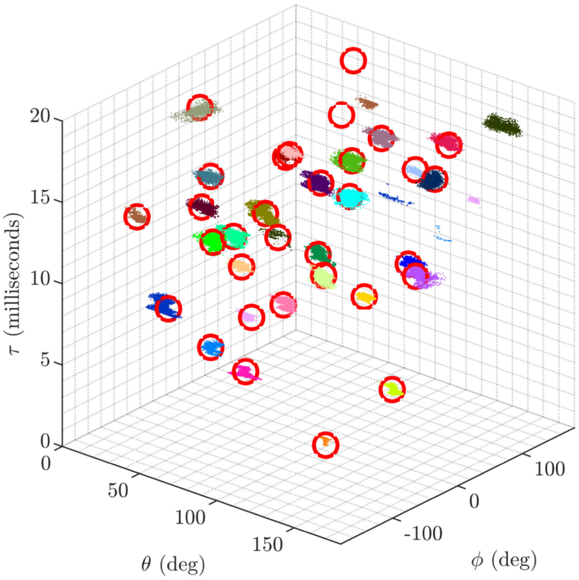


Fig. 5. Local DOA and delay estimates obtained by PHALCOR, colored by assigned cluster number. The θ and ϕ axes are for elevation and azimuth, respectively. The red circles correspond to true DOAs and delays.

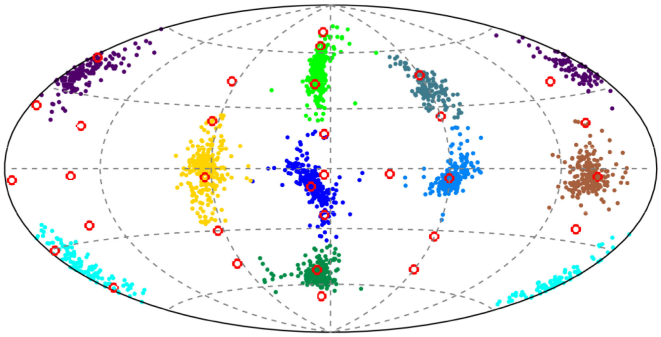


Fig. 6. Local DOA estimates obtained by MUSIC, colored by assigned cluster number. The red circles correspond to $\Omega_0, \ldots, \Omega_{31}$. The Hammer projection is used to project the sphere onto the plane.

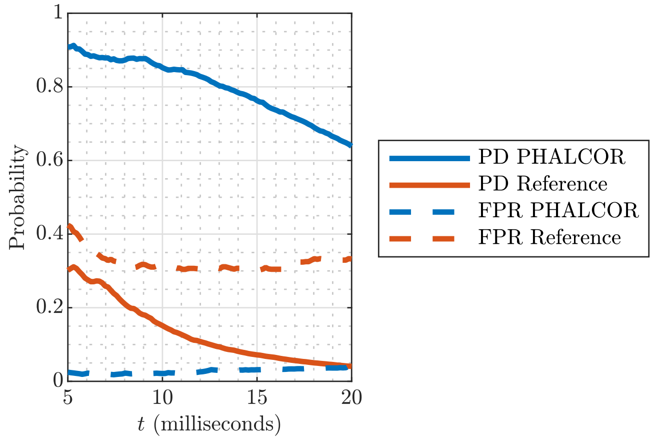


Fig. 7. PD and FPR, as defined in Section VI-B, of PHALCOR and the reference method.

D. Monte Carlo Analysis

The simulation described above is repeated 50 times for each of the 3 rooms, varying the speakers, their location, and the microphone array location, as detailed in Section VI-A. Fig. 7 presents PD and FPR, as defined in Section VI-B, for different values of t, the maximum delay of the identified reflections. Compared with the reference method, the performance of PHALCOR is significantly better, both in terms of probability of detection and false positive rates, by a factor ranging from 3 to 20. As the delay of a reflection increases, the probability of detection decreases. This is since later reflections usually have lower amplitudes. Furthermore, the reflection density is higher as the delay increases, making it more difficult to separate the reflections spatially. Figs. 8 and 9 show how the distribution of reflection amplitudes and reflection temporal densities vary between detected and undetected reflections. Here, we define the

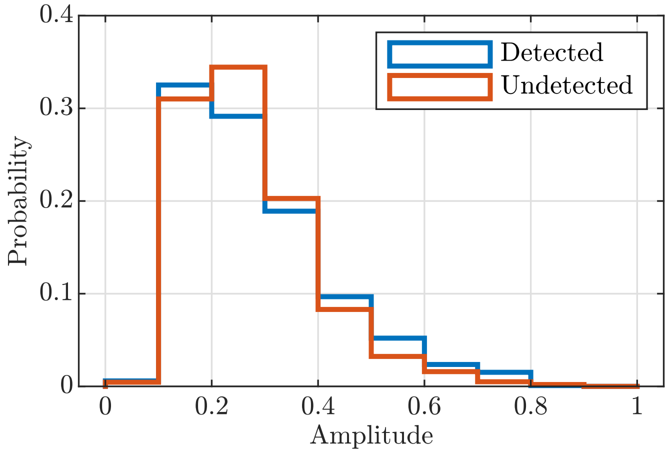


Fig. 8. Histogram of reflection amplitudes for detected reflections and undetected reflections.

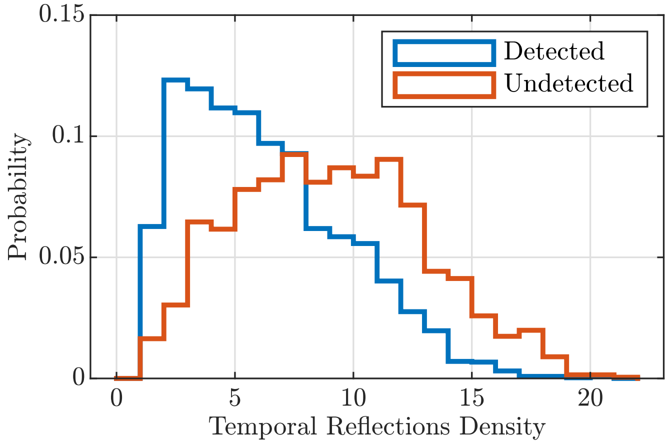


Fig. 9. Histogram of reflection temporal densities for detected reflections and undetected reflections. The temporal density for a given reflection is defined as the number of neighboring reflections within 1 ms of the reflection delay.

TABLE IV
AVERAGE ESTIMATION ERRORS FOR THE ENTIRE MONTE CARLO SIMULATION

Method	DOA Error RMS (deg)	Delay Error RMS (μs)	Average Number Of True Positive Detections
PHALCOR	4.3	77	24.8
Reference	6.5	43	3.8

temporal density of a reflection with a delay τ as the number of reflections with a delay within 1 ms of τ. Note that, as described in Section III, the amplitudes are normalized such that the direct sound amplitude is 1. As is evident, detected reflections tend to be of higher amplitude, and lower temporal density, compared to undetected reflections. It is also apparent that this effect is more significant for temporal density.

The root mean square (RMS) for DOA and delay estimation errors for each method are computed and averaged for all the estimates in this Monte Carlo simulation, and are presented in Table IV. The RMS is calculated excluding the direct sound. Table IV shows that the performance in terms of DOA estimation error is comparable between the two methods. In terms of delay estimation error, the reference method is superior, but note that the errors are calculated only on true positive detections, which are considerably more frequent in PHALCOR, as is evident from Fig. 7 and the last column of Table IV. Furthermore, in applications where high delay accuracy is important, it may be possible to apply cross correlation (as in the reference method) for delay estimation as a post processing step of PHALCOR. Nevertheless, this is proposed for future work.

VII. Conclusion

In this work, PHALCOR, a novel method for estimating the DOA and delay of early reflections, is proposed. The method is based on a phase alignment transform of the spatial correlation matrices, which enables the detection of reflections with similar DOAs. A simulation study showed that the proposed method is able to detect and localize a large number of reflections compared to existing methods. The estimation of reflection amplitudes and the validation of the method performance on measured data are proposed for future work. Future research can focus on extending PHALCOR to accept microphone signals from compact arrays of different configurations and wearable arrays, as well as to handle multiple simultaneous sources.

References

[1] K. Kowalczyk, S. Kacprzak, and M. Ziółko, “On the extraction of early reflection signals for automatic speech recognition,” in *Proc. IEEE 2nd Int. Conf. Signal Image Process.*, 2017, pp. 351–355.

[2] Y. Peled and B. Rafaely, “Method for dereverberation and noise reduction using spherical microphone arrays,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 113–116.

[3] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, “From blind to guided audio source separation: How models and side information can improve the separation of sound,” *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 107–115, May 2014.

[4] H. A. Javed, A. H. Moore, and P. A. Naylor, “Spherical microphone array acoustic rake receivers,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 111–115.

[5] E. Mabande, K. Kowalczyk, H. Sun, and W. Kellermann, “Room geometry inference based on spherical microphone array eigenbeam processing,” *J. Acoustical Soc. America*, vol. 134, no. 4, pp. 2773–2789, 2013.

[6] J. Catic, S. Santurette, and T. Dau, “The role of reverberation-related binaural cues in the externalization of speech,” *J. Acoustical Soc. Amer.*, vol. 138, no. 2, pp. 1154–1167, 2015.

[7] M. Vorländer, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. Berlin, Germany: Springer, 2007.

[8] V. Pulkki, S. Delikaris-Manias, and A. Politis, *Parametric Time-Frequency Domain Spatial Audio*. Hoboken, NJ, USA: Wiley, 2018.

[9] P. Coleman, A. Franck, P. Jackson, R. J. Hughes, L. Remaggi, and F. Melchior, “Object-based reverberation for spatial audio,” *J. Audio Eng. Soc.*, vol. 65, no. 1/2, pp. 66–77, 2017.

[10] H. Kuttruff, *Room Acoustics*. Boca Raton, FL, USA: CRC Press, 2016.

[11] H. Sun, E. Mabande, K. Kowalczyk, and W. Kellermann, “Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing,” *J. Acoustical Soc. Amer.*, vol. 131, no. 4, pp. 2828–2840, 2012.

[12] B. Jo and J.-W. Choi, “Robust localization of early reflections in a room using semi real-valued eb-esprit with three recurrence relations and laplacian constraint,” in *Proc. 23rd Int. Congr. Acoust.*, Sep. 2019, pp. 4949–4956.

[13] D. Ciuonzo, G. Romano, and R. Solimene, “Performance analysis of time-reversal music,” *IEEE Trans. Signal Process.*, vol. 63, no. 10, pp. 2650–2662, May 2015.

[14] D. Ciuonzo, “On time-reversal imaging by statistical testing,” *IEEE Signal Process. Lett.*, vol. 24, no. 7, pp. 1024–1028, Jul. 2017.

[15] P. K. T. Wu, N. Epain, and C. Jin, “A dereverberation algorithm for spherical microphone arrays using compressed sensing techniques,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 4053–4056.

[16] Y. Hu, J. Lu, and X. Qiu, "Direction of arrival estimation of multiple acoustic sources using a maximum likelihood method in the spherical harmonic domain," *Appl. Acoust.*, vol. 135, pp. 85–90, 2018.

[17] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. Hoboken, NJ, USA: Wiley, 2004.

[18] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2510–2526, Nov. 2007.

[19] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," *in Microphone Arrays*. Springer, 2001, pp. 157–180.

[20] H. Do and H. F. Silverman, "SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 125–128.

[21] T. Shlomo and B. Rafaely, "Blind localization of early room reflections from reverberant speech using phase aligned spatial correlation," *J. Acoustical Soc. Amer.*, 148, no. 4, pp. 2547–2547.

[22] B. Rafaely, *Fundamentals of Spherical Array Processing*. Berlin, Germany: Springer, 2015, vol. 8.

[23] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustical Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[24] E. Fisher and B. Rafaely, "Near-field spherical microphone array processing with radial filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 256–265, Feb. 2011.

[25] A. V. Oppenheim, *Discrete-Time Signal Processing*. New York, NY, USA: Pearson Education India, 1999.

[26] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.

[27] T. T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4680–4688, Jul. 2011.

[28] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Kdd*, vol. 96, no. 34, pp. 226–231, 1996.

[29] J. Fliege and U. Maier, "A two-stage approach for computing cubature formulae for the sphere," in *Mathematik 139 T*, Universitat Dortmund, Fachbereich Mathematik, Universitat Dortmund, 44221. Citeseer, 1996.

[30] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2009, pp. 221–224.

[31] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, Aug. 2005.

[32] J. Hassab and R. Boucher, "Optimum estimation of time delay by a generalized correlator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 4, pp. 373–380, Aug. 1979.

[33] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detecP-tion and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASS33, no. 4, pp. 823–831, Aug. 1985.

[34] P. Kabal, "Tsp speech database," McGill University, *Database Version*, vol. 1, no. 0, pp. 09–02, 2002.

[35] M. Acoustics, "Em32 Eigenmike Microphone Array Release Notes (v17. 0)," *25 Summit Ave*, Summit, NJ 07901, USA, 2013.

[36] K. Han and A. Nehorai, "Improved source number detection and direction estimation with nested arrays and ulas using jackknifing," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 6118–6128, Dec. 2013.

Tom Shlomo (Student Member, IEEE) received the B.Sc. (cum laude) and M.Sc. degrees in electrical and computer engineering from Ben-Gurion University of the Negev, Beer-Sheva, Israel, in 2019 and 2020, respectively. His research interests include acoustical signal processing and spatial audio.

Boaz Rafaely (Senior Member, IEEE) received the B.Sc. degree (cum laude) in electrical engineering from Ben-Gurion University, Beer-Sheva, Israel, in 1986, the M.Sc. degree in biomedical engineering from Tel Aviv University, Tel Aviv, Israel, in 1994, and the Ph.D. degree from the Institute of Sound and Vibration Research (ISVR), Southampton University, Southampton, U.K., in 1997. In 1997, he was appointed as a Lecturer with ISVR and a Senior Lecturer in 2001, working on active control of sound and acoustic signal processing. In 2002, he spent six months as a Visiting Scientist with the Sensory Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, investigating speech enhancement for hearing aids. In 2003, he joined the Department of Electrical and Computer Engineering, Ben-Gurion University, as a Senior Lecturer, and appointed as an Associate Professor and a Professor in 2010 and 2013, respectively. He is currently heading the Acoustics Laboratory, investigating methods for audio signal processing and spatial audio. During 2010–2014, he was as an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, and during 2013–2018, was a Member of the IEEE Audio and Acoustic Signal Processing Technical Committee. He was also an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS during 2015–2019, the *IET Signal Processing* during 2016–2019, and currently for the *Acta Acustica*. During 2013–2016, he was the Chair of the Israeli Acoustical Association, and is currently chairing the Technical Committee on Audio Signal Processing in the European Acoustical Association. He was the recipient of the British Council's Clore Foundation Scholarship.