Direct and Residual Subspace Decomposition of Spatial Room Impulse Responses

Citation for the original published paper (version of record):
Deppisch, T., Garí, S., Calamia, P. et al (2023). Direct and Residual Subspace Decomposition of Spatial Room Impulse Responses. IEEE/ACM Transactions on Audio Speech and Language Processing, 31: 927-942. http://dx.doi.org/10.1109/TASLP.2023.3240657

DOI 10.1109/TASLP.2023.3240657

Direct and Residual Subspace Decomposition of Spatial Room Impulse Responses

Thomas Deppisch, *Graduate Student Member, IEEE*, Sebastià V. Amengual Garí, Paul Calamia, and Jens Ahrens, *Senior Member, IEEE*

***Abstract*—Psychoacoustic experiments have shown that directional properties of the direct sound, salient reflections, and the late reverberation of an acoustic room response can have a distinct influence on the auditory perception of a given room. Spatial room impulse responses (SRIRs) capture those properties and thus are used for direction-dependent room acoustic analysis and virtual acoustic rendering. This work proposes a subspace method that decomposes SRIRs into a direct part, which comprises the direct sound and the salient reflections, and a residual, to facilitate enhanced analysis and rendering methods by providing individual access to these components. The proposed method is based on the generalized singular value decomposition and interprets the residual as noise that is to be separated from the other components of the reverberation. Large generalized singular values are attributed to the direct part, which is then obtained as a low-rank approximation of the SRIR. By advancing from the end of the SRIR toward the beginning while iteratively updating the residual estimate, the method adapts to spatio-temporal variations of the residual. The method is evaluated using a spatio-spectral error measure and simulated SRIRs of different rooms, microphone arrays, and ratios of direct sound to residual energy. The proposed method creates lower errors than existing approaches in all tested scenarios, including a scenario with two simultaneous reflections. A case study with measured SRIRs shows the applicability of the method under real-world acoustic conditions. A reference implementation is provided.**

***Index Terms*—Microphone array, room reflections, spatial room impulse response, subspace method, virtual acoustic rendering**

I. Introduction

DIRECTIONAL properties of acoustic environments have been subject to extensive research in recent years as they are a key factor in human auditory perception. It was found that salient reflections, i.e., reflections with sufficiently high energy, can have an individual impact on the perceived spatial impression [1], the apparent source position and width [2], and the timbre [3]. On the other hand, statistical properties of the reverberation, such as the overall energy, its angular distribution, the direct-to-reverberant energy ratio, and the reverberation time, influence the perceived envelopment [4], source distance [5], and room size [6], [7]. Auditory perception is also influenced by the directional energy decay of late reverberation [8], [9].

Spatial room impulse responses (SRIRs) capture the directional properties of an acoustic environment and facilitate the analysis and reproduction thereof. Note that the term SRIR is heavily used in literature but is defined inconsistently. We refer to an SRIR as a set of room impulse responses that is captured by a single microphone array to facilitate the directional analysis or auralization for a single source position and from the perspective of a single receiver position. Hence, suitable microphone arrays have a small aperture, typically smaller than $0.5\,\mathrm{m}$. No other requirements are imposed on the array geometry; however, due to their wide commercial availability and the possibility to perform a spherical harmonic (SH) decomposition of the array signals [10], often spherical microphone arrays are used.

Motivated by their perceptual relevance, SRIR-based directional room acoustic analysis and virtual acoustic rendering methods specifically target salient reflections and directional statistic properties of the reverberation. Common analysis objectives include the direction-of-arrival (DOA) estimation of reflections [11]–[13] and the directional energy decay of the reverberation [9], [14]–[16].

SRIR-based virtual acoustic renderers reproduce the acoustics of an environment by convolving a processed SRIR with source signals. They target multi-channel loudspeaker and/or binaural headphone playback. The renderers either analyze the direction of the frequency-dependent instantaneous acoustic intensity [17] or use a broadband DOA estimator [18]–[20] to impose spatial information onto an omnidirectional signal. An extension of [17] generalizes the method using higher-order SHs and processing in angular sectors [21]. A recently proposed method [22] analyzes DOAs of reflections as in [18] but explicitly cuts out salient reflections from the omnidirectional RIR to resynthesize the early part of an SRIR. The methods render diffuse reverberation either implicitly by a fast modulation of the reproduction direction, or explicitly by using a diffuseness estimate and decorrelating diffuse signal parts to create multi-channel loudspeaker signals. Instead of the decorrelation, the late reverberation might also be replaced by filtered noise [23]. Other methods analyze SRIRs to generate parametric synthetic reverberation [24]–[26].

While most of the renderers are designed to accurately reproduce salient reflections and diffuse reverberation, none of the existing methods achieves an explicit separation of salient

This research was supported by Reality Labs Research at Meta.

Thomas Deppisch and Jens Ahrens are with the Chalmers University of Technology, 412 96 Gothenburg, Sweden (e-mail: thomas.deppisch@chalmers.se; jens.ahrens@chalmers.se).

Sebastià V. Amengual Garí and Paul Calamia are with Reality Labs Research, Meta, Redmond, WA 98052, USA (e-mail: samengual@meta.com; pcalamia@meta.com).

Manuscript received May 09, 2022; revised XXXX XX, 2023.

reflections from the SRIR while preserving the spatio-temporal properties of both the reflections and the residual. To overcome this limitation, the authors recently proposed to use the spatial subtraction method to subtract salient reflections from an SRIR by using a beamformer and a plane-wave prototype [27]. The spatial subtraction method was initially proposed for the separation of direct and diffuse parts in sound scenes [28] and was extended in [27] by employing a comprehensive microphone-array signal model that includes the impacts of scattering and spatial aliasing, which improved the separation performance when applied to SRIRs.

The current work proposes a subspace method to separate SRIRs into a direct part, comprising the direct sound and salient reflections, and a residual. The proposed method is shown to improve the separation performance in comparison to the spatial subtraction method as it avoids the error-prone estimation of reflection parameters. It is free from typical assumptions, such as reflections being plane waves and late reverberation being isotropic, and does not rely on parameter estimation regarding, e.g., the number of simultaneous reflections and their DOAs. In consequence, the method also does not provide an estimation of such parameters but rather generates two SRIRs, one containing the direct part and the other containing the residual, that can then be analyzed and processed independently. Nevertheless, the method may improve the performance of existing parameter estimation algorithms by applying it as pre-preprocessing.

By providing an explicit separation of the direct part and the residual, the method facilitates advanced rendering and extrapolation strategies of SRIRs. A perceptual pilot study of such extrapolation strategies suggests that an efficient SRIR extrapolation to different positions in a room may be achieved using the proposed method by combining a residual SRIR from a single measurement with salient reflections from the target position [29].

II. Subspace Decomposition Theory

A. General Principle

Subspace methods in array processing are based on the assumption that target signals only occupy a limited subspace of the full signal space that is spanned by the multiple, noisy sensor readings. The methods reduce noise by confining the noisy signal to a subspace containing a superposition of signal and noise, called the *signal subspace*, while disregarding components in the orthogonal *noise subspace* that are solely attributed to the noise [30], [31].

Subspace methods essentially exploit the Eckart-Young-Mirsky theorem [32] to find the best low-rank approximation of a signal matrix. In array signal processing, this was first applied by Tufts et al. [33] and they showed that the low-rank approximation can either be performed via the singular value decomposition (SVD) of the data matrix or by an orthogonal projection using eigenvectors of the covariance matrix. Later, the principle was exploited in beamforming [34, Ch. 6.8] and parameter estimation [30], [31]. It was also applied in speech enhancement, first in single-channel [35], [36] and later in array-based methods [37], [38]. In speech enhancement, noise components in the signal subspace are typically further reduced using a signal-dependent post-filter and several estimators have been proposed for that purpose [39]. Those estimators reduce noise at the cost of increased signal distortion and thus will be disregarded in this work.

To motivate subspace methods mathematically, it is beneficial to analyze the covariance matrix of a noisy array signal. Let $\boldsymbol{x}(t)$ be a length-M vector containing the signals that are captured by M microphones at the discrete time t. Following a convolutive multiple-input-multiple-output (MIMO) signal model [40, Ch. 2.1.4], the array signals are convolutive mixtures of the source signals $\boldsymbol{s}(t)$ plus additive noise $\boldsymbol{n}(t)$,

$$\boldsymbol{x}(t) = \boldsymbol{H}\boldsymbol{s}(t) + \boldsymbol{n}(t)\,. \tag{1}$$

In each row, the $M \times QN$ matrix $\boldsymbol{H}$ contains Q finite impulse response (FIR) filters of length N that describe the transfer paths from each of the Q sources to one of the M microphones. Accordingly, N observations of each source signal are stacked in the length-QN source signal vector $\boldsymbol{s}(t)$ such that $\boldsymbol{H}\boldsymbol{s}(t)$ describes the convolution of the source signals with the FIR filters at the time t. The noise vector $\boldsymbol{n}(t)$ is of length M and contains one noise observation per microphone at the time t. This general signal model also forms the basis for the decomposition of SRIRs that will be introduced in Sec. III. In that context, the source signals $\boldsymbol{s}(t)$ represent the components of the direct part of the SRIR, i.e., direct sound and salient reflections, $\boldsymbol{H}$ describes their transfer paths to the microphone array, and $\boldsymbol{n}(t)$ represents the residual SRIR.

The spatial covariance matrix $\boldsymbol{R}_\mathrm{x}$ is defined as the expectation $\mathcal{E}\{\cdot\}$ of the outer vector product,

$$\boldsymbol{R}_\mathrm{x} = \mathcal{E}\{\boldsymbol{x}(t)\boldsymbol{x}(t)^\mathrm{T}\}\,. \tag{2}$$

When assuming that the convolutively mixed signals $\boldsymbol{H}\boldsymbol{s}(t)$ and the noise $\boldsymbol{n}(t)$ are mutually uncorrelated, the covariance $\boldsymbol{R}_\mathrm{x}$ of the noisy signal is the sum of the covariance of the mixed source signals $\boldsymbol{R}_\mathrm{s} = \mathcal{E}\{\boldsymbol{H}\boldsymbol{s}(t)\boldsymbol{s}(t)^\mathrm{T}\boldsymbol{H}^\mathrm{T}\}$ and the covariance of the noise $\boldsymbol{R}_\mathrm{n} = \mathcal{E}\{\boldsymbol{n}(t)\boldsymbol{n}(t)^\mathrm{T}\}$,

$$\boldsymbol{R}_\mathrm{x} = \boldsymbol{R}_\mathrm{s} + \boldsymbol{R}_\mathrm{n}\,. \tag{3}$$

If the noise is spatially white, i.e., it is uncorrelated across microphones and has common variance σ_n^2, its covariance is a scaled identity matrix, $\boldsymbol{R}_\mathrm{n} = \sigma_n^2\boldsymbol{I}$. The eigenvalue decomposition (EVD) of the array signal covariance then shows that the signal and noise covariance matrices share the same set of eigenvectors [39], which are collected in the columns of $\boldsymbol{U}$,

$$\boldsymbol{R}_\mathrm{x} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\mathrm{T} = \boldsymbol{U}(\boldsymbol{\Lambda}_\mathrm{s} + \sigma_\mathrm{n}^2\boldsymbol{I})\boldsymbol{U}^\mathrm{T}\,. \tag{4}$$

Thus, the eigenvalues of the noisy-signal covariance $\boldsymbol{R}_\mathrm{x}$ in the diagonal matrix $\boldsymbol{\Lambda}$ are equal to the eigenvalues of the source signal covariance $\boldsymbol{R}_\mathrm{s}$ in the diagonal matrix $\boldsymbol{\Lambda}_\mathrm{s}$ plus the noise variance σ_n^2. Note that due to the convolutive mixture and possible correlation between the source signals $\boldsymbol{s}(t)$, the rank Q_s of $\boldsymbol{R}_\mathrm{s}$ cannot be assumed to be equal to the number of sources Q. However, if the covariance $\boldsymbol{R}_\mathrm{s}$ is singular, i.e., its rank $Q_\mathrm{s} < M$, then $\boldsymbol{\Lambda}_\mathrm{s}$ contains zero-valued eigenvalues and the smallest $Q_\mathrm{n} = M - Q_\mathrm{s}$ eigenvalues of $\boldsymbol{R}_\mathrm{x}$ are equal to the noise variance σ_n^2.

This is the core observation that subspace methods exploit to reduce noise as it allows for a separation between eigenvalue-eigenvector pairs that are attributed to signal-plus-noise components and other pairs that are solely attributed to the noise. The eigenvectors corresponding to large eigenvalues are hence referred to as signal eigenvectors and are collected in the columns of $\boldsymbol{U}_\mathrm{s}$, while the eigenvectors corresponding to small eigenvalues equal to σ_n^2 are referred to as noise eigenvectors and are collected in the columns of $\boldsymbol{U}_\mathrm{n}$. Noise reduction is then performed by an orthogonal projection of the noisy signal onto the signal subspace [33],

$$\boldsymbol{x}_\mathrm{s}(t) = \boldsymbol{U}_\mathrm{s}\boldsymbol{U}_\mathrm{s}^\mathrm{T}\boldsymbol{x}(t)\,, \tag{5}$$

and the corresponding noise-only signal is obtained by an orthogonal projection onto the noise subspace,

$$\boldsymbol{x}_\mathrm{n}(t) = \boldsymbol{U}_\mathrm{n}\boldsymbol{U}_\mathrm{n}^\mathrm{T}\boldsymbol{x}(t)\,. \tag{6}$$

B. Subspace Decomposition Including a Noise Estimate

If the noise $\boldsymbol{n}(t)$ is not spatially white, the noise covariance matrix $\boldsymbol{R}_\mathrm{n}$ is not a scaled identity matrix and the attribution of eigenvalues of the covariance $\boldsymbol{R}_\mathrm{x}$ to the signal or noise subspace based on their magnitudes fails. Note that, due to the finite distance between the diaphragms, microphone array signals in rooms are never fully uncorrelated, not even in homogeneous diffuse fields [41, Ch. 2.2]. However, if an estimate of the noise covariance matrix is available, the generalized eigenvalue decomposition (GEVD) of $\boldsymbol{R}_\mathrm{x}$ and $\boldsymbol{R}_\mathrm{n}$ diagonalizes the noisy-signal covariance $\boldsymbol{R}_\mathrm{x}$ and the noise covariance $\boldsymbol{R}_\mathrm{n}$ simultaneously [31],

$$\boldsymbol{\Psi}^\mathrm{T}\boldsymbol{R}_\mathrm{x}\boldsymbol{\Psi} = \boldsymbol{\Delta}\,, \tag{7}$$

$$\boldsymbol{\Psi}^\mathrm{T}\boldsymbol{R}_\mathrm{n}\boldsymbol{\Psi} = \boldsymbol{I}\,. \tag{8}$$

The columns of $\boldsymbol{\Psi}$ contain the generalized eigenvectors and the diagonal matrix $\boldsymbol{\Delta}$ contains the generalized eigenvalues. As under the assumption of spatially white noise that was similarly exploited in (4), the generalized eigenvalues $\boldsymbol{\Delta}$ are equal to the eigenvalues of $\boldsymbol{R}_\mathrm{s}$ offset by one, $\boldsymbol{\Delta} = \boldsymbol{R}_\mathrm{s} + \boldsymbol{I}$. Thus, the generalized eigenvalues can be interpreted as the eigenvalues of a pre-whitened signal and the generalized eigenvectors span the signal subspace [42, Ch. 8.7]. Note that as in the EVD, the eigenvectors are uniquely determined only up to an arbitrary factor. This factor is commonly chosen such that $\boldsymbol{\Psi}^\mathrm{T}\boldsymbol{R}_\mathrm{n}\boldsymbol{\Psi} = \boldsymbol{I}$.

Due to the diagonalization of the noise covariance, a magnitude-based discrimination between signal and noise eigenvalues is possible again but the generalized eigenvectors in the columns of $\boldsymbol{\Psi}$, which are the eigenvectors of $\boldsymbol{R}_\mathrm{n}^{-1}\boldsymbol{R}_\mathrm{x}$, are not orthogonal as $\boldsymbol{R}_\mathrm{n}^{-1}\boldsymbol{R}_\mathrm{x}$ is not necessarily symmetric. The projection onto signal and noise subspace is hence expressed as [43],

$$\boldsymbol{x}_\mathrm{s}(t) = \boldsymbol{\Psi}^{-\mathrm{T}}\boldsymbol{\Gamma}_\mathrm{s}\boldsymbol{\Psi}^\mathrm{T}\boldsymbol{x}(t)\,, \tag{9}$$

$$\boldsymbol{x}_\mathrm{n}(t) = \boldsymbol{\Psi}^{-\mathrm{T}}\boldsymbol{\Gamma}_\mathrm{n}\boldsymbol{\Psi}^\mathrm{T}\boldsymbol{x}(t)\,, \tag{10}$$

where the diagonal binary selection matrices $\boldsymbol{\Gamma}_\mathrm{s} = \boldsymbol{I}_{Q_\mathrm{s}\times M}^\mathrm{T}\boldsymbol{I}_{Q_\mathrm{s}\times M}$ and $\boldsymbol{\Gamma}_\mathrm{n} = \boldsymbol{I} - \boldsymbol{\Gamma}_\mathrm{s}$ contain Q_s ones as the first Q_s diagonal entries and $Q_\mathrm{n} = M - Q_\mathrm{s}$ ones as the last Q_n diagonal entries, respectively, and zeros otherwise, to select the Q_s and Q_n eigenvectors of their corresponding subspace. As the sum of $\boldsymbol{\Gamma}_\mathrm{s}$ and $\boldsymbol{\Gamma}_\mathrm{n}$ is the identity matrix, the sum of the signal and noise subspace signals perfectly reconstructs the original array signal, $\boldsymbol{x}(t) = \boldsymbol{x}_\mathrm{s}(t) + \boldsymbol{x}_\mathrm{n}(t)$. The subspace decomposition using the GEVD is mathematically equivalent to the sequence of pre-whitening the signal, decomposing it into the subspaces, and de-whitening the result [42, Ch. 8.7].

If the noise estimate is obtained by discrete measurements of the noise-only signal, the explicit computation of the signal and noise covariances can be avoided for computational efficiency by employing the generalized singular value decomposition (GSVD) instead of the GEVD [38]. The GSVD relies on data matrices, i.e., matrices containing multiple observations of the signal. Let $\boldsymbol{X}$ and $\boldsymbol{N}$ be $K \times M$ and $L \times M$ matrices that contain K and L observations of their corresponding signals $\boldsymbol{x}(t)$ and $\boldsymbol{n}(t)$. The GSVD then decomposes the noisy-signal data matrix $\boldsymbol{X}$ and the noise data matrix $\boldsymbol{N}$ into an orthogonal matrix $\boldsymbol{V}_\mathrm{x}$ or $\boldsymbol{V}_\mathrm{n}$, a non-negative diagonal matrix $\boldsymbol{\Sigma}_\mathrm{x}$ or $\boldsymbol{\Sigma}_\mathrm{n}$, and a common square matrix $\boldsymbol{\Phi}$,

$$\boldsymbol{X} = \boldsymbol{V}_\mathrm{x}\boldsymbol{\Sigma}_\mathrm{x}\boldsymbol{\Phi}^\mathrm{T}\,, \tag{11}$$

$$\boldsymbol{N} = \boldsymbol{V}_\mathrm{n}\boldsymbol{\Sigma}_\mathrm{n}\boldsymbol{\Phi}^\mathrm{T}\,. \tag{12}$$

The matrices $\boldsymbol{\Sigma}_\mathrm{x}$ and $\boldsymbol{\Sigma}_\mathrm{n}$ contain the singular values, while $\boldsymbol{V}_\mathrm{x}$ and $\boldsymbol{V}_\mathrm{n}$ contain the respective left singular vectors and $\boldsymbol{\Phi}$ contains the common right singular vectors. For notational brevity, we assume an *economy-sized* GSVD and $K \geq M$, $L \geq M$, so that $\boldsymbol{\Sigma}_\mathrm{x}$ and $\boldsymbol{\Sigma}_\mathrm{n}$ are $M \times M$ square matrices.

The generalized eigenvalues of $\boldsymbol{R}_\mathrm{x}$ and $\boldsymbol{R}_\mathrm{n}$ on the diagonal of $\boldsymbol{\Delta}$ are obtained via the GEVD, cf. (7). If the corresponding covariance matrices are estimated via the sample covariance, i.e., $\boldsymbol{R}_\mathrm{x} = \frac{1}{K}\boldsymbol{X}^\mathrm{T}\boldsymbol{X}$ and $\boldsymbol{R}_\mathrm{n} = \frac{1}{L}\boldsymbol{N}^\mathrm{T}\boldsymbol{N}$, the generalized eigenvalues $\boldsymbol{\Delta}$ can equivalently be obtained via the GSVD [44, Ch. 8.7.4],

$$\boldsymbol{\Delta} = \frac{L}{K}\left(\boldsymbol{\Sigma}_\mathrm{x}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{x}\right)\left(\boldsymbol{\Sigma}_\mathrm{n}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{n}\right)^{-1}. \tag{13}$$

For convenience, we further define the vector

$$\boldsymbol{\sigma} = \mathrm{diag}\left(\left(\boldsymbol{\Sigma}_\mathrm{x}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{x}\right)\left(\boldsymbol{\Sigma}_\mathrm{n}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{n}\right)^{-1}\right)\,, \tag{14}$$

that contains the squared generalized singular values (GSVs) in decreasing order. The $\mathrm{diag}\,(\cdot)$ operator transfers the entries from the main diagonal of a matrix to a vector. The squared GSVs are an essential part of the proposed threshold selection mechanism in Sec. III-D and will simply be referred to as GSVs in the following.

In the case of the GSVD, the subspace decomposition is performed as a low-rank approximation, resulting in the $K \times M$ signal subspace and noise subspace matrices $\boldsymbol{X}_\mathrm{s}$ and $\boldsymbol{X}_\mathrm{n}$ [42, Ch. 8.4]. Similar to the subspace decomposition using the GEVD in (9) and (10), the Q_s largest singular values and their corresponding singular vectors are used to obtain the signal subspace components and the last $Q_\mathrm{n} = M - Q_\mathrm{s}$ singular vectors corresponding to the smallest singular values are used to obtain the noise subspace components,

$$\boldsymbol{X}_\mathrm{s} = \boldsymbol{V}_\mathrm{x}\boldsymbol{\Sigma}_\mathrm{x}\boldsymbol{\Gamma}_\mathrm{s}\boldsymbol{\Phi}^\mathrm{T}\,, \tag{15}$$

$$\boldsymbol{X}_\mathrm{n} = \boldsymbol{V}_\mathrm{x}\boldsymbol{\Sigma}_\mathrm{x}\boldsymbol{\Gamma}_\mathrm{n}\boldsymbol{\Phi}^\mathrm{T}\,. \tag{16}$$

Note that some subspace methods in the literature impose more assumptions on the signals and thus are able to gain more information in the decomposition process. They typically either assume narrow-band signals [30], [31] or assume short-term stationary signals and perform blockwise processing in the frequency domain [34, Ch. 5.2]. These assumptions reduce the convolutive MIMO signal model in (1) to a multiplicative model, i.e., the FIR filters in $\boldsymbol{H}$ reduce to single-sample scaling factors. The rank Q_s of the source covariance $\boldsymbol{R}_\mathrm{s}$ is then equal to the number of sources Q and the individual source signals are decorrelated which facilitates a source parameter estimation. However, as established in Sec. III, the assumptions that are imposed on the signals in this work are more restrictive.

III. Subspace Decomposition of Spatial Room Impulse Responses

A. Signal Model

Motivated by their perceptual relevance, the proposed algorithm aims at separating the direct sound and salient reflections from the SRIR. Applying the nomenclature from Sec. II to the present context means that direct sound and salient reflections are considered convolutively mixed source signals and everything else is considered noise. In the following, the direct sound and salient reflections will also be referred to as the direct part or direct subspace components, and the rest, containing the superposition of an increasing amount of reflections and noise will be referred to as the residual.

The spatio-temporal properties of the residual of the SRIR typically change over time. In the early part of the SRIR, the residual mainly contains noise and non-transient components of the room response due to room modes. As time progresses, it is additionally comprised of a superposition of non-salient reflections. Toward the later part of the SRIR, the residual is dominated by the superposition of exponentially increasing numbers of non-salient reflections. This reverberation might exhibit isotropic or anisotropic properties, or a combination of both that varies over time [9], [14]. During the late part of the SRIR, no salient reflections are expected so that the SRIR is composed of only the residual and no direct part.

To adapt to the spatio-temporal variations of the residual, we propose to update the residual estimate whenever no salient reflections are detected within a signal block. The procedure is assumed to be successful if the properties of the residual change slower than the residual estimate is updated. Additionally, we propose to process the SRIR backward in time to be able to obtain a reliable residual estimate before any salient reflection occurs. This process is illustrated in Fig. 2 and a more detailed overview of the algorithm will be given in Sec. III-C.

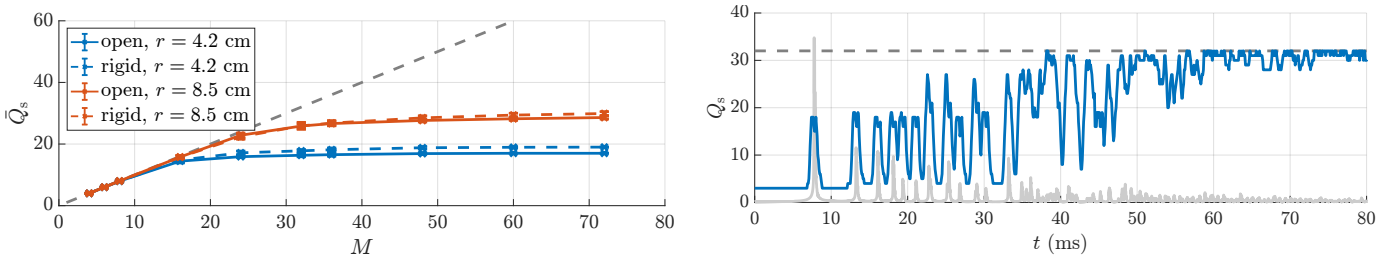
The covariance matrix is assumed to be estimated via the sample covariance of a signal block that contains a limited number of signal observations, cf. Sec. II-B. However, when the GSVD is used, the sample covariance is not calculated explicitly. As the direct sound and the salient reflections are highly transient signals, the rank of the source covariance matrix depends on the correlation of the captured signals and the temporal separation of the transients that are induced by a single or multiple reflections. The correlation depends on the transfer function from the source to the individual microphones, i.e., on properties of the acoustic environment and the array, and the temporal separation depends on the distance between the microphones.

The only requirement for the microphone array is that it has a small aperture so that sound pressures that are generated by a reflection are captured by a single signal block. Determining the duration of a reflection is not a straightforward task as it is influenced by the array aperture, diffraction, and scattering. However, the propagation delay of a sound wave across the maximum array dimension is often a good approximation. In practice, often spherical arrays are employed and their signals are transformed to the spherical harmonic (SH) domain, where they are radial filtered to compensate for the array radius and the scattering of the array baffle [10, Ch. 2.6]. This leads to the typical assumption that spherical arrays in SH-domain processing have frequency-independent steering vectors so that, as with narrow-band signals, a multiplicative signal model is sufficient. However, the necessary regularization of the radial filters and spatial aliasing in practice limit this property to a narrow frequency region [27]. Thus, these assumptions are not made in the proposed broadband algorithm so that it can either be directly applied to the microphone signals or to an SH decomposition thereof. We demonstrate the application of the proposed algorithm with both signal representations in Sec. IV-A.

B. Rank Analysis of the Covariance Matrix

The core assumption that facilitates noise reduction via subspace methods is that the source signals, which are in the present case the direct sound and the salient reflections, only occupy a subspace of the full signal space. In other words, noise reduction is only possible if the source covariance matrix $\boldsymbol{R}_\mathrm{s}$ is singular, $Q_\mathrm{s} < M$. To determine if a subspace decomposition is feasible for salient reflections in an SRIR, in the following, the rank of the source covariance matrix is analyzed, first for an individual plane wave impinging on different microphone arrays and then for a simulated SRIR.

Fig. 1 (a) shows the mean and the standard deviation of the rank Q_s of the covariance matrix for a plane wave that impinges on different spherical microphone arrays under anechoic conditions. Note that the standard deviations are small and thus hardly visible. The dashed gray line illustrates the maximum rank M, which is equal to the number of microphones in the array. The mean rank $\bar{Q}_\mathrm{s}$ is obtained by averaging the rank over 240 incidence directions that are distributed according to a t-design [45]. The rank for each incidence direction is calculated as the number of eigenvalues of the covariance matrix that are less than $100\,\mathrm{dB}$ below the largest eigenvalue. The plane waves and the spherical scattering were simulated using the spherical microphone array impulse response generator (SMIRGen) [46]. The simulated array configurations include 10 different spherical arrangements comprising between 4 and 72 microphones that all are arranged according to t-designs. All 10 arrangements were



(a) (b)

Fig. 1. The subspace decomposition can be performed if the source covariance matrix $\boldsymbol{R}_\mathrm{s}$ is singular, i.e., its rank $Q_\mathrm{s} < M$. **(a)** The mean rank $\bar{Q}_\mathrm{s}$ due to a single impinging plane wave depends on the number of microphones M, on the array radius r, and on the array surface being rigid or open. It is singular if it stays below the dashed gray line illustrating the number of microphones. **(b)** The source covariance matrix of an SRIR was simulated using the image source method. It is singular in the early part. The summed magnitude of the SRIR is shown in gray for reference.

simulated as open and rigid arrays, and with radii of $4.2\,\mathrm{cm}$ and $8.5\,\mathrm{cm}$.

The covariance matrix on average has full rank for the array configurations that comprise 4, 6, and 8 microphones. The mean rank $\bar{Q}_\mathrm{s}$ is close to being full in the case of the array configurations with 16 microphones and in the case of the arrays with the larger radius and 24 microphones. For all other configurations, the mean rank is clearly singular. Thus, a subspace decomposition for a single impinging plane wave can be performed with the simulated arrays of radius $4.2\,\mathrm{cm}$ that comprise $M = 24$ or more microphones and with the arrays of radius $8.5\,\mathrm{cm}$ that comprise $M = 32$ or more microphones. With an increasing number of microphones M, all tested configurations converge to a maximum rank that depends on the array configuration and is equal to 29 and 30 for the open and rigid sphere configurations with a radius of $8.5\,\mathrm{cm}$, and equal to 17 and 19 for the configurations with a radius of $4.2\,\mathrm{cm}$. The covariance matrices from microphone arrays with the larger radius generally have a higher rank Q_s than the ones from the smaller arrays since their signals contain larger temporal delays and since they capture less-correlated signals because those arrays are large compared to the wavelength down to lower frequencies. The observed rank is also slightly higher in the case of rigid arrays because the scattering of sound waves off their surface additionally decorrelates the captured signals.

To determine if a subspace decomposition is still feasible if more than one reflection occurs per analysis window, we analyze the evolution of the rank Q_s for an SRIR. The SRIR was generated using the image source method [47] and SMIRGen, assuming a shoe-box room of dimensions $8 \times 7 \times 6$ m. Note that the goal of this simulation is not to render a highly-realistic SRIR but to investigate the rank of the direct-part (source) covariance matrix $\boldsymbol{R}_\mathrm{s}$ due to the direct sound and individual reflections, not the full covariance $\boldsymbol{R}_\mathrm{x}$ that will be used in the subspace decomposition. The separation of eigenvalues can be attempted as in (4) only if the source covariance $\boldsymbol{R}_\mathrm{s}$ is singular. The simulated array is of radius $4.2\,\mathrm{cm}$ and comprises 32 microphones that are arranged according to a t-design. Fig. 1 (b) shows Q_s during the first $80\,\mathrm{ms}$ of the SRIR. The summed magnitude of the SRIR is shown in gray for reference. The sample covariance matrix was calculated in 32-sample ($0.7\,\mathrm{ms}$) rectangular windows with a hop size of 4 samples. The rank Q_s was calculated as the number of eigenvalues of the covariance matrix that are less than $100\,\mathrm{dB}$ below the largest eigenvalue in each window.

During the first $38\,\mathrm{ms}$, the rank Q_s consistently stays below the maximum possible rank of $M = 32$, which is again shown as a dashed gray line. Between $38\,\mathrm{ms}$ and $59\,\mathrm{ms}$, Q_s fluctuates over a wide range of values and approaches the maximum rank multiple times. After $59\,\mathrm{ms}$, the rank stays close to the maximum rank. Hence, for the given SRIR a subspace decomposition can separate the direct part from the residual in the early part of the SRIR until $38\,\mathrm{ms}$ and might be able to separate some salient reflections until $59\,\mathrm{ms}$. These specific time spans do not generalize to other SRIRs and microphone arrays. However, it can be assumed that the time span where a separation is effective increases with an increasing number of microphones and that the subspace decomposition is effective in the early part of typical SRIRs if a suitable microphone array is used. This is also shown in the case study in Sec. V, where the proposed method is applied to three measured SRIRs with different acoustic properties. In practice, the decomposition of the SRIR might be mainly relevant up to the perceptual mixing time [48] and thus the method could be limited to the early part of the SRIR according to a mixing time estimate.

C. Algorithm Overview

Fig. 2 illustrates the application of the proposed subspace decomposition algorithm to a simulated SRIR. The SRIR is simulated as direct sound and first-order image-source reflections in exponentially decaying noise. For convenience and without loss of generality, the decaying noise starts before the direct sound occurs. The simulated, rigid spherical microphone array is the same that was used in Fig. 1 (b). It has a radius of $4.2\,\mathrm{cm}$ and comprises 32 microphones that are arranged according to a spherical t-design of degree 7 [45]. Again, the image-source method was employed using SMIRGen [46]. The microphone signals of the exponentially decaying noise tail were generated to exhibit the spatial coherence of the array in an isotropic spherical noise field using the method from [49].

Fig. 2 (a) shows five channels of the full 32-channel SRIR. The algorithm first takes a signal block from the end of the SRIR as an initial residual estimate and then performs

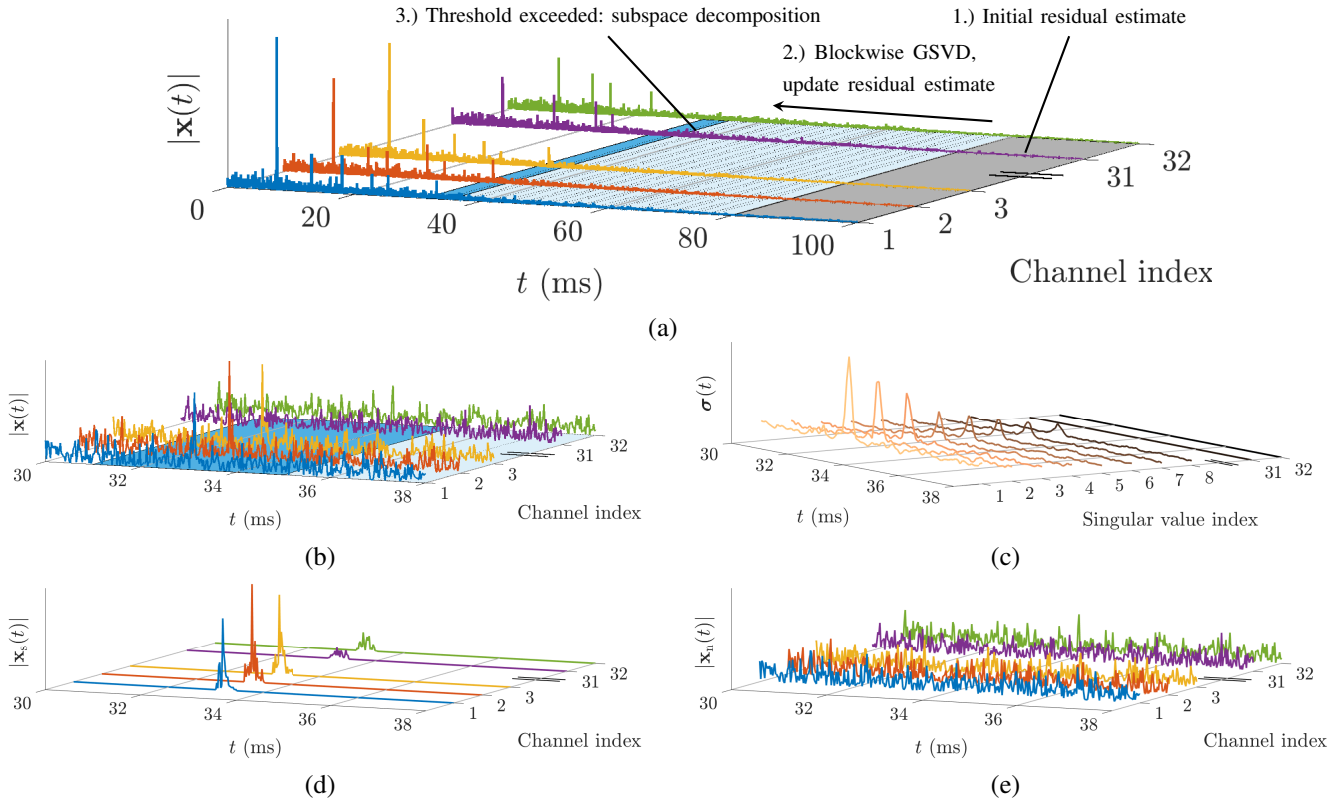


Fig. 2. Direct and residual subspace decomposition of a 32-channel SRIR $\boldsymbol{x}(t)$. **(a)** The proposed algorithm first takes an initial residual estimate from the end of the SRIR. It then proceeds toward the beginning of the SRIR and performs the GSVD on every signal block. If the sum of the GSVs $\boldsymbol{\sigma}(t)$ is below the detection threshold, the residual estimate is updated. If their sum exceeds the threshold, the subspace decomposition is performed. **(b)** A zoomed-in part of the SRIR contains a salient reflection. **(c)** The eight largest GSVs of the zoomed-in part exhibit a distinct peak at the location of the reflection. The two smallest GSVs do not exhibit a visible peak. **(d)** The direct signal $\boldsymbol{x}_\mathrm{s}(t)$ contains the salient reflection from (b). **(e)** The residual signal $\boldsymbol{x}_\mathrm{n}(t)$ does not contain the reflection.

a blockwise GSVD while proceeding toward the beginning of the SRIR. If the sum of the generalized singular values (GSVs) exceeds the detection threshold, the SRIR is decomposed into the direct part and the residual. Otherwise, only the residual estimate is updated.

Fig. 2 (b) shows a magnified section of the SRIR that includes a salient reflection. A subset of the corresponding GSVs is shown in Fig. 2 (c). The first eight GSVs exhibit a distinct peak at the location of the reflection while the smallest two GSVs do not exhibit a visible peak. All Q_s GSVs above a given threshold are attributed to the reflection whereas the $Q_\mathrm{n} = 32 - Q_\mathrm{s}$ smaller GSVs are attributed to the residual. In the present case, Q_s was chosen to be 6. A method that determines this threshold is proposed in Sec. III-D. The direct part SRIR that contains the reflection is shown in Fig. 2 (d) and the residual SRIR in Fig. 2 (e). The sum of the two reconstructs the original SRIR.

D. Threshold Selection

The selection of appropriate thresholds for the detection of salient reflections and the estimation of the number of direct subspace components Q_s is key to a successful decomposition. Common criteria to find the number of signal subspace components are either based on the ratio of the geometric mean to the arithmetic mean of a number of small eigenvalues [50], in a nutshell rating the *equality* of a subset of eigenvalues, or using measures to find the gap between a set of larger and a set of smaller eigenvalues [51], [52]. The methods are based on the assumption of eigenvalues of the noise subspace being similar in size and do not exploit prior information like a noise estimate.

To facilitate a robust threshold selection by incorporating information from the residual estimate, we propose a threshold measure based on the cumulative sum of the GSVs. It is inspired by [53], where the number of components Q_s is selected so that the reconstruction error is close to an estimate of the noise variance. The Frobenius norm of a matrix is the root of the sum of its squared elements and is equal to the root of the sum of its squared singular values [44, Ch. 2.4.2]. In the case of data matrices containing microphone array signals, the Frobenius norm can be interpreted as the energetic sum of all microphone signals. A threshold that keeps the total energy of the residual in the presence of salient reflections equal to the energy of the full signal in the absence of salient reflections can hence be defined via the sum of squared singular values.

The proposed threshold is based on this idea, however, two further observations lead to its precise definition: i) In contrast to the orthogonal left and right singular vectors of the SVD, the GSVD involves the non-orthogonal right singular vectors $\boldsymbol{\Phi}$, cf. (11). In consequence, the rooted sum of the squared singular values in $\boldsymbol{\Sigma}_\mathrm{x}$ is not equal to the Frobenius norm of

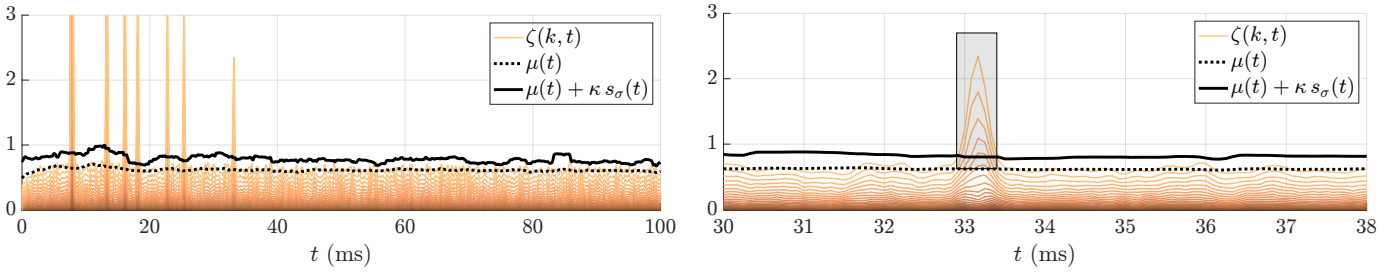


Fig. 3. The weighted cumulative sums $\zeta(k,t)$ of the GSVs, shown in shades from orange to brown, exhibit distinct peaks at times where reflections occur. **(a)** The GSV sum, which is the largest of the cumulative sums $\zeta(k,t)$, exceeds the detection threshold, drawn as a solid black line, for the direct sound and each of the 6 reflections. Thus, all direct components are detected. **(b)** Zoomed-in section of (a) around a reflection. The left and right borders of the gray rectangle mark the time instances where a reflection is detected. The number of direct subspace components is determined as the number of weighted, cumulatively summed GSVs $\zeta(k,t)$ that exceed the time-averaged sum $\mu(t)$ of the GSVs, which is shown as a dotted black line. At the peak, this results in 6 direct components and 26 residual components.

the signal matrix $\boldsymbol{X}$. However, the GSVs $\boldsymbol{\sigma}$ can be interpreted as the singular values of the pre-whitened signal and the SVD of the pre-whitened signal, which is not explicitly calculated, involves orthogonal singular vectors. Thus, by choosing the number of residual components Q_n during the decomposition so that the sum of the corresponding Q_n GSVs equals the sum of all GSVs in the absence of salient reflections, the energy of the whitened residual can be kept constant.

ii) While the residual subspace only contains residual components, the direct subspace contains a superposition of direct and residual components. The singular values of the whitened signal are the GSVs and they all carry equal parts of the variance of the whitened residual. Thus, the sum of the Q_n GSVs during the decomposition needs to be smaller than the sum of all GSVs in the absence of salient reflections. More precisely, if Q_n GSVs are attributed to the residual, they should carry a fraction of Q_n/M of the full energy that is determined when no salient reflections are present. Recall that M denotes the number of microphones.

In mathematical terms, the proposed criterion determines the number of residual components Q_n as the maximum integer k for which the cumulative sum of k GSVs $\zeta(k,t)$ is still smaller than the time-averaged sum of the GSVs $\mu(t)$ in the absence of salient reflections,

$$Q_\mathrm{n}(t) = \max(\{k \in \mathbb{N}^+ : \zeta(k,t) < \mu(t)\})\,, \tag{17}$$

where

$$\zeta(k,t) = \frac{M}{k} \sum_{m=M-k+1}^{M} [\boldsymbol{\sigma}(t)]_m \tag{18}$$

is the cumulative sum of the k smallest GSVs weighted by M/k and $[\boldsymbol{\sigma}(t)]_m$ denotes the m-th element of the vector $\boldsymbol{\sigma}(t)$, i.e., the m-th largest GSV. The weighting M/k stems from the above reasoning that showed that the whitened-residual energy should be a Q_n/M part of the GSV sum. As Q_n is being determined at this point, it has been replaced by the index k. The fraction k/M has further been transferred to the left side of the inequality in (17) to become M/k. The time-averaged GSV sum

$$\mu(t) = \mathcal{F}_\mathrm{CMA}\left\{\sum_{m=1}^{M} [\boldsymbol{\sigma}(t)]_m\right\} \tag{19}$$

is obtained by applying the constrained moving average filter $\mathcal{F}_\mathrm{CMA}\{\cdot\}$ that time-averages the sum of the GSVs and is only updated if no salient reflection is present. Once the number of residual components Q_n is known, the number of direct subspace components is obtained as $Q_\mathrm{s} = M - Q_\mathrm{n}$.

A second thresholding mechanism is needed to detect the presence of reflections. Only if reflections are detected, the estimation of the number of subspace components is performed. Similar to the estimator for the number of subspace components, the proposed detection threshold is based on the sum of the GSVs that is averaged over time instants without salient reflections. Reflections are detected if the sum of the GSVs of the current observation is larger than the time-averaged sum of previous GSVs $\mu(t)$ plus a multiple κ of their standard deviation $s_\sigma(t)$,

$$\sum_{m=1}^{M} [\boldsymbol{\sigma}(t)]_m > \mu(t) + \kappa\, s_\sigma(t)\,. \tag{20}$$

If that concept is to be implemented, a first-in-first-out (FIFO) buffer that contains a number of observations of the sum of the GSVs and is only updated during time instances without salient reflections is beneficial. The averaged GSV sum $\mu(t)$ and the standard deviation $s_\sigma(t)$ are then calculated as the arithmetic mean and the standard deviation of all observations in the buffer.

Fig. 3 shows the weighted cumulative GSV sums $\zeta(k,t)$ in shades from orange to brown, the time-averaged sum of GSVs $\mu(t)$ as a dotted black line, and the reflection detection threshold $\mu(t) + \kappa\, s_\sigma(t)$ as a solid black line for the same simulated SRIR as in Fig. 2. The detection threshold is calculated using $\kappa = 4$. The parameter selection process is further described in Sec. III-E and in the supplementary material referenced therein. Due to the iterative update of the residual estimate, the GSVs and also their cumulative sums stay constant in the absence of reflections although the reverberation is exponentially decaying, see Fig. 2 (a). Reflections are detected whenever the sum of the GSVs, which is equivalent to the largest cumulative sum, is larger than the detection threshold. As shown in Fig. 3 (a), the GSV sum exceeds the detection threshold for the direct sound and all 6 reflections and hence all direct components are detected.

Fig. 3 (b) shows a zoomed-in section around the occurrence of the last reflection. The gray rectangle illustrates the estimation of the number of direct subspace components Q_s. The left and right boundaries of the rectangle illustrate the temporal bounds in which the GSV sum is larger than the detection threshold. Within these bounds, the subspace decomposition is performed. The number of direct components is the number of weighted, cumulatively summed GSVs $\zeta(k,t)$ that is larger than the averaged GSV sum $\mu(t)$, resulting in $Q_\mathrm{s} = 6$ direct components and $Q_\mathrm{n} = 26$ residual components at the peak location.

E. Influence of the Parameters

This section discusses the influence of the different parameters including the block size, the detection threshold offset κ, the amount of GSV averaging for the calculation of the thresholds, and the length of the residual estimate. In practice, the optimal parameter values depend on the acoustic environment and the employed microphone array, and they can be found by analyzing the decomposition results, the evolution of GSV sums, and the proposed thresholds as in Fig. 3. For brevity, we discuss the influence of the parameters here and provide examples that illustrate the influence of the different parameters and the parameter selection process as supplementary material[1].

The block size determines the temporal resolution of the subspace decomposition. A lower bound of the block size in samples is given by the number of microphones (or SH coefficients) of the employed array to be able to exploit the full signal space. Additionally, the block size should capture the full propagation delay of a sound wave across the maximum array dimension. For instance, in the case of the spherical Eigenmike em32 array with a radius of $4.2\,\mathrm{cm}$, we assume a maximum dimension of $8.4\,\mathrm{cm}$ and a corresponding propagation delay of $0.24\,\mathrm{ms}$. Larger block sizes decrease the temporal resolution of the calculated GSVs and thus may reduce the temporal precision of the extraction of salient reflections from the residual. In this contribution, we use a sampling rate of $48\,\mathrm{kHz}$ and set the block size to either 32 samples, for microphone arrays with 32 or fewer microphones, or to 64 samples, for arrays with more than 32 microphones. The hop size between consecutive blocks is set to $1/8$ of the block size to frequently update the thresholds, the residual estimate, and a possible decomposition.

The detection threshold offset κ determines the number of standard deviations by which the GSV sum in a signal block needs to exceed the averaged GSV sum such that a reflection is detected. With smaller values of κ, weaker energetic peaks are treated as reflections and with too small values most of the energy in the early part of the SRIR may be assigned to the direct part. If, on the other hand, κ is chosen too large, only very strong reflections will be extracted from the residual. Values of $\kappa = 3$ or 4 yielded good results in our experiments and are used in all examples in this contribution.

[1]A MATLAB Live Script and a corresponding PDF document are provided at https://github.com/thomasdeppisch/SRIR-Subspace-Decomposition.

The averaging of the GSVs ensures that the proposed thresholds change smoothly over time and are not strongly influenced by individual reflections, cf. Sec. III-D. If too little averaging is applied, the thresholds fluctuate strongly when reflections are detected and reflections that appear slightly earlier in time than other reflections might not be detected due to the raised detection threshold. (Recall that the algorithm proceeds backward in time.) If too much averaging is applied, the thresholds do not account for overall changes in energy in the residual. All examples in this work use averaging lengths between 32 and 64 blocks.

The length of the residual estimate determines how fast the GSV reacts to changes in the overall covariance of the residual. Appropriate lengths result in GSV sums that stay constant in the absence of reflections and exhibit strong peaks in the presence of reflections, cf. Fig. 3. Too short estimates prevent the implicit pre-whitening of the residual so that the GSV sums do not stay constant over time in the absence of reflections. Very long estimates reduce the relative peak height of GSV sums and thus make the separation between direct part and residual more difficult. All examples in this contribution use a residual estimate with a length of $20\,\mathrm{ms}$.

F. Algorithm Summary

This section summarizes the proposed algorithm with reference to the pseudocode in Algorithm 1.

As illustrated in Fig. 2 (a), the proposed algorithm starts by taking a signal block from the end of the SRIR as an initial residual estimate $\boldsymbol{N}_0$ and then advances in a blockwise manner toward the beginning of the SRIR, starting with the signal block J that directly precedes the initial residual estimate. The FIFO buffer $\boldsymbol{\rho}$ that will later contain observations of the sum of the GSVs in the absence of salient reflections is initialized with very large values, or infinity, so that the detection threshold will not be exceeded within the first signal blocks.

The signal is assumed to be divided into overlapping blocks before the processing and the GSVD is performed for each signal block $\boldsymbol{X}_i$ and the residual estimate $\boldsymbol{N}$. The number of observations of the residual estimate $\boldsymbol{N}$ and the number of signal observations in the blocks $\boldsymbol{X}_i$ can be chosen independently. The exemplary SRIR from Figs. 2 and 3 was decomposed using a block size of $K = 32$ samples, a hop size of 4 samples and a residual estimate with a length of 960 samples.

The sum of the GSVs is then compared to the detection threshold that is calculated from the average of the observations of the GSV sum in $\boldsymbol{\rho}$ plus a multiple κ of their standard deviation. For the exemplary SRIR, we set $\kappa = 4$ and averaged the GSV sum over 32 observations. If the current GSV sum exceeds the detection threshold, the cumulative sum of the GSVs is calculated, summing from the smallest toward the largest GSV. The estimated direct subspace dimension Q_s is obtained once the weighted cumulative GSV sum exceeds the average of the GSV sums in $\boldsymbol{\rho}$. From the dimension of the direct subspace, the binary direct subspace selection matrix $\boldsymbol{\Gamma}_\mathrm{s}$ is calculated. It is a diagonal matrix, containing ones as the first Q_s diagonal entries and zeros otherwise.

Algorithm 1 SRIR Subspace Decomposition

$\boldsymbol{N} = \boldsymbol{N}_0$ ▷ initial residual estimate
$i = J$ ▷ initial block index
$\boldsymbol{\rho} = \mathbf{inf}$ ▷ initialize GSV sums as infinity
while $i > 1$ **do** ▷ iterate over signal blocks
  $\boldsymbol{X}_i = \boldsymbol{V}_\mathrm{x}\boldsymbol{\Sigma}_\mathrm{x}\boldsymbol{\Phi}^\mathrm{T}$ ▷ GSVD of signal and residual data
  $\boldsymbol{N} = \boldsymbol{V}_\mathrm{n}\boldsymbol{\Sigma}_\mathrm{n}\boldsymbol{\Phi}^\mathrm{T}$
  $\boldsymbol{\sigma} = \mathrm{diag}\left(\boldsymbol{\Sigma}_\mathrm{x}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{x}(\boldsymbol{\Sigma}_\mathrm{n}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{n})^{-1}\right)$ ▷ GSVs
  $\xi = \sum_{m=1}^{M}[\boldsymbol{\sigma}]_m$ ▷ sum of GSVs
  if $\xi > \mathrm{mean}(\boldsymbol{\rho}) + \kappa\,\mathrm{std}(\boldsymbol{\rho})$ **then** ▷ reflection detected
    $c_\sigma = [\boldsymbol{\sigma}]_M$
    $k = 1$
    while $c_\sigma\, M/k < \mathrm{mean}(\boldsymbol{\rho})$ **do**
      c_σ += $[\boldsymbol{\sigma}]_{M-k}$ ▷ cumulative GSV sum
      k++
    end while
    $Q_\mathrm{s} = M - k + 1$ ▷ direct subspace dimension
    $\boldsymbol{\Gamma}_\mathrm{s} = \boldsymbol{I}_{Q\times M}^\mathrm{T}\boldsymbol{I}_{Q\times M}$
  else ▷ no reflection detected
    $\boldsymbol{\Gamma}_\mathrm{s} = \boldsymbol{0}_{M\times M}$
    $\boldsymbol{N} \circlearrowleft \boldsymbol{X}_i$ ▷ update residual estim. (FIFO)
    $\boldsymbol{\rho} \circlearrowleft \xi$ ▷ update sum of GSVs (FIFO)
  end if
  $\boldsymbol{\Gamma}_\mathrm{n} = \boldsymbol{I}_{M\times M} - \boldsymbol{\Gamma}_\mathrm{s}$
  $\boldsymbol{X}_{\mathrm{s},i} = \boldsymbol{V}_\mathrm{x}\boldsymbol{\Sigma}_\mathrm{x}\boldsymbol{\Gamma}_\mathrm{s}\boldsymbol{\Phi}^\mathrm{T}$ ▷ direct signal block
  $\boldsymbol{X}_{\mathrm{n},i} = \boldsymbol{V}_\mathrm{x}\boldsymbol{\Sigma}_\mathrm{x}\boldsymbol{\Gamma}_\mathrm{n}\boldsymbol{\Phi}^\mathrm{T}$ ▷ residual signal block
  i--
end while

If the current GSV sum does not exceed the detection threshold, $\boldsymbol{\Gamma}_\mathrm{s}$ is set to a zero matrix and the residual estimate $\boldsymbol{N}$ is updated by replacing its oldest rows by the rows of the current signal data matrix $\boldsymbol{X}_i$ that were not already included in previous signal blocks, using the FIFO principle. Similarly, the current GSV sum replaces the oldest element in the vector $\boldsymbol{\rho}$.

Subsequently, the residual subspace selection matrix $\boldsymbol{\Gamma}_\mathrm{n}$ is calculated from the direct subspace selection matrix $\boldsymbol{\Gamma}_\mathrm{s}$. If the detection threshold was not exceeded, $\boldsymbol{\Gamma}_\mathrm{n}$ is the identity matrix, attributing all GSVs to the residual subspace. Otherwise, $\boldsymbol{\Gamma}_\mathrm{n}$ is a diagonal matrix, whose first Q_s diagonal entries are zero and whose last Q_n diagonal entries are one. Finally, the current signal block is decomposed into a direct part $\boldsymbol{X}_{\mathrm{s},i}$ and a residual part $\boldsymbol{X}_{\mathrm{n},i}$, by performing low-rank approximations of the signal matrix $\boldsymbol{X}_i$.

IV. Quantitative Evaluation

This section comprises an evaluation of the proposed subspace decomposition method using simulated SRIRs. A perceptual evaluation of the method is beyond the scope of this contribution, however, a perceptual evaluation of an application of the herein proposed method is available in [29]. The evaluation starts in Sec. IV-A with an illustration showing that the application of the subspace decomposition is possible with unprocessed microphone array signals as well as with an SH decomposition thereof. The following sections apply the method to SH-domain signals to make it directly comparable to the spatial subtraction method whose signal model relies on SH-domain processing. In Sec. IV-B magnitude spectra of extracted reflections that are obtained by the proposed method and the spatial subtraction method using two different signal models are analyzed. Then, in Sec. IV-C, the proposed method is evaluated using a spatio-spectral error measure and is compared to the spatial subtraction method and to a temporal cut-out approach for different rooms, microphone arrays, and levels of the residual. The evaluation ends with a comparison of the performance of the methods in the presence of two simultaneous reflections in Sec. IV-D. A case study with measured SRIRs follows in Sec. V.

A. Raw vs. SH-Domain Processing

In a nutshell, the proposed subspace decomposition method achieves the separation of the direct part and the residual by comparing the energy of singular values of the array signals to the energy of a residual estimate. The method does not assume a particular arrangement or directivity of the employed microphones. An SH decomposition of the microphone signals can be interpreted as signals captured by microphones with a specific directivity, e.g., the zeroth-order SH has an omnidirectional directivity and first-order SHs have figure-of-eight directivities that are aligned with the Cartesian axes. Thus, the subspace decomposition method can be applied to unprocessed microphone signals or an SH decomposition thereof.

Fig. 4 (a) shows the norms of the ground truth direct part $\|\boldsymbol{x}_\mathrm{s}(t)\|$ and of the ground truth residual $\|\boldsymbol{x}_\mathrm{n}(t)\|$ of a simulated SRIR, i.e., seven simulated reflections are treated as direct part ground truth and noise with the spatial coherence of the array in an isotropic spherical noise field is treated as residual ground truth. The simulated, rigid array is again of radius $4.2\,\mathrm{cm}$ and comprises 32 microphones that are arranged according to a t-design. Figs. 4 (b) and (c) show the norms of the direct part $\|\boldsymbol{x}_\mathrm{s}(t)\|$ and of the residual $\|\boldsymbol{x}_\mathrm{n}(t)\|$ that are obtained by applying the subspace decomposition to the unprocessed SRIR and to an SH decomposition using up to fourth-order SHs. The subspace decomposition method does not have access to the individual parts shown in Fig. 4 (a) but only to their sum. All SH decompositions in this work are accompanied by radial filtering using Tikhonov regularization [54]. The radial filtering reduces the influence of scattering on the SH signals and may improve the separability of reflections and the residual in more complex scenarios. A detailed analysis of this is however beyond the scope of this contribution and is left for future work. Following the reasoning from Sec. III-E, the subspace decomposition was in both cases performed using a block size of 32 samples ($0.7\,\mathrm{ms}$), a hop size of 4 samples, a residual estimate of $20\,\mathrm{ms}$, GSV averaging of 32 blocks and $\kappa = 4$. A comparison of Figs. 4 (a), (b), and (c) shows that the decomposition is successful with unprocessed signals and with an SH decomposition thereof: in both cases, the seven reflections are extracted from the rest of the SRIR. A detailed performance evaluation using a numerical error measure follows in Sec. IV-C.

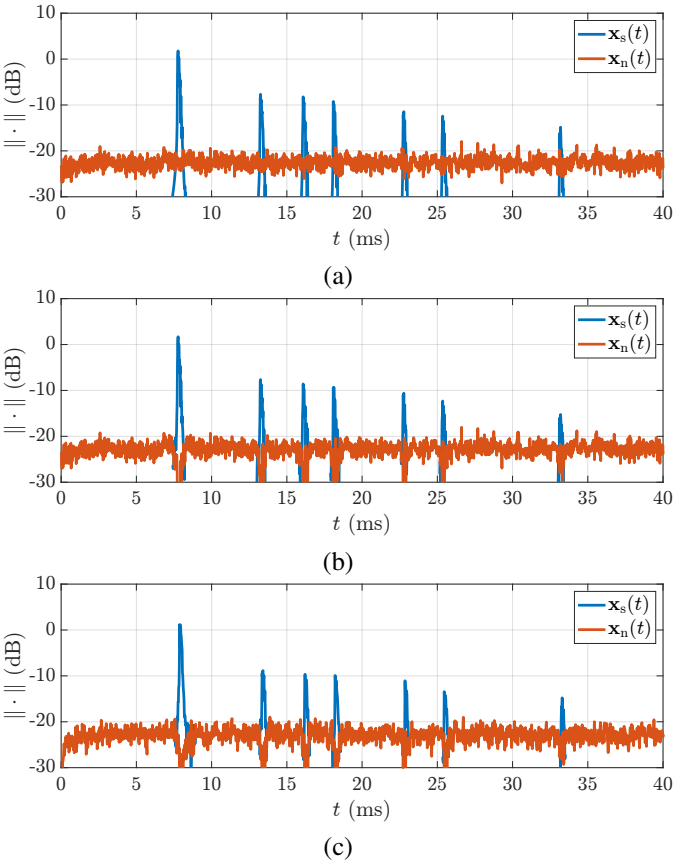


Fig. 4. Norms of the direct part $\boldsymbol{x}_\mathrm{s}(t)$ and the residual $\boldsymbol{x}_\mathrm{n}(t)$ of, **(a)**, the ground truth, **(b)**, the proposed method applied to unprocessed microphone signals and, **(c)**, the proposed method applied to an SH decomposition of the array signals.

B. Analysis of Extracted Reflection Spectra

In this section, the norms of the spectra $\boldsymbol{\chi}_\mathrm{s}(f)$ of two of the ground truth reflections from Fig. 4 (a) are compared to extracted spectra from the direct part $\boldsymbol{x}_\mathrm{s}(t)$ obtained either via the spatial subtraction method [27] or the subspace decomposition. The frequency-domain vector $\boldsymbol{\chi}_\mathrm{s}(f)$ contains the spectrum of all SH-domain signal channels during the presence of a reflection. Specifically, we analyze the spectra of the first and the last reflection in Fig. 4 (a) to show results for different ratios of reflection and residual energy. According to Parseval's theorem for the spherical Fourier transform [55], the norm of an SH-domain signal vector equals the total signal energy integrated over the surface of the unit sphere and is thus a suitable measure to illustrate the overall results of the different methods. For the spatial subtraction method, we assume that the time-of-arrival (TOA) of the respective reflection is known and use SH-MUSIC [56] to estimate its DOA. For both reflections, the DOA estimation errors are small, they amount to $1.6°$ for the first reflection and to $2.1°$ for the seventh reflection. After the DOA estimation, the spatial subtraction method is applied using two different signal models, the one originally proposed in the context of sound scenes in [28], in the following referred to as *SpatSub1*, and the comprehensive signal model from [27] that includes the influences of scattering, radial filtering, and spatial aliasing and is referred to as *SpatSub2*. Note that the proposed subspace

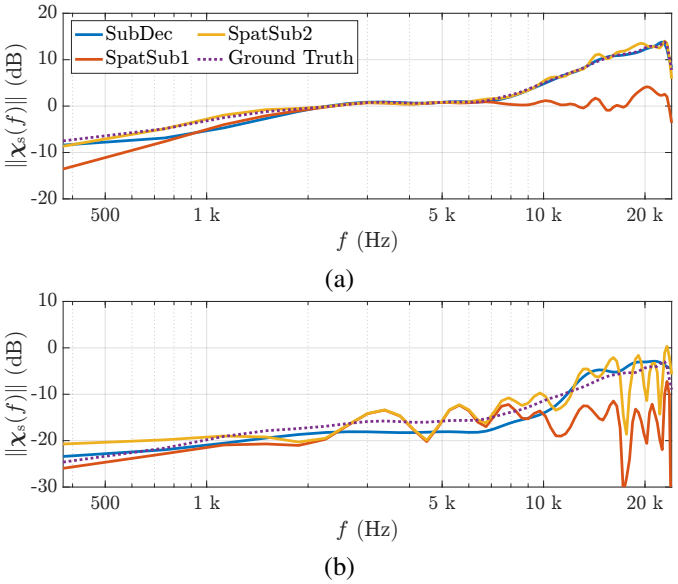


Fig. 5. Norms of the ground truth spectra $\boldsymbol{\chi}_\mathrm{s}(f)$ of, **(a)**, the first and, **(b)**, the seventh reflection from Fig. 4 (a) and of extracted reflection spectra using the spatial subtraction method with two different signal models, *SpatSub1* and *SpatSub2*, as well as the proposed subspace decomposition method *SubDec*.

decomposition method, which is in the following also referred to as *SubDec*, does not have access to the true TOAs and does not utilize the DOA estimates. The spatial subtraction method was applied using the discrete Fourier transform (DFT) of the array signals within a $1\,\mathrm{ms}$ rectangular window centered around the respective reflection and the shown spectra are calculated within the same window. The $1\,\mathrm{ms}$ window ensures that the full reflection is captured while limiting the amount of noise, i.e., the norm of the simulated reflection decays by about $30\,\mathrm{dB}$ within the window and is at least $6\,\mathrm{dB}$ below the noise floor at the edge of the window. The subspace decomposition was performed using the same parameters as in Sec. IV-A.

Fig. 5 (a) shows the obtained norms of the spectra $\boldsymbol{\chi}_\mathrm{s}(f)$ for the first reflection. The norms of the spectra obtained by both *SubDec* and *SpatSub2* follow the norm of the ground truth closely. The subspace decomposition (*SubDec*) shows a maximum deviation from the ground truth of about $2\,\mathrm{dB}$ around $1\,\mathrm{kHz}$ while *SpatSub2* has a maximum deviation of about $1.5\,\mathrm{dB}$ around $19\,\mathrm{kHz}$. The method *SpatSub1* deviates more strongly from the ground truth. Its underlying signal model does not include the influence of non-ideal radial filtering, leading to a deviation of about $6\,\mathrm{dB}$ around $400\,\mathrm{Hz}$, and also neglects the influence of spatial aliasing and the SH order truncation, leading to large deviations above $8\,\mathrm{kHz}$, with a maximum deviation of $12\,\mathrm{dB}$ around $16\,\mathrm{kHz}$.

The norms of the spectra for the extraction of the seventh reflection are shown in Fig. 5 (b). The ratio of reflection peak to residual energy is in this case much lower in comparison to the first reflection, cf. Fig. 4 (a), making the extraction task more difficult. Again *SubDec* and *SpatSub2* follow the spectrum of the ground truth closely but this time *SpatSub2* exhibits strong fluctuations that increase with frequency, resulting in a maximum deviation of $13\,\mathrm{dB}$ at $17\,\mathrm{kHz}$. The spectrum of the proposed method *SubDec* does not show such fluctuations and has a maximum deviation from the ground truth of about

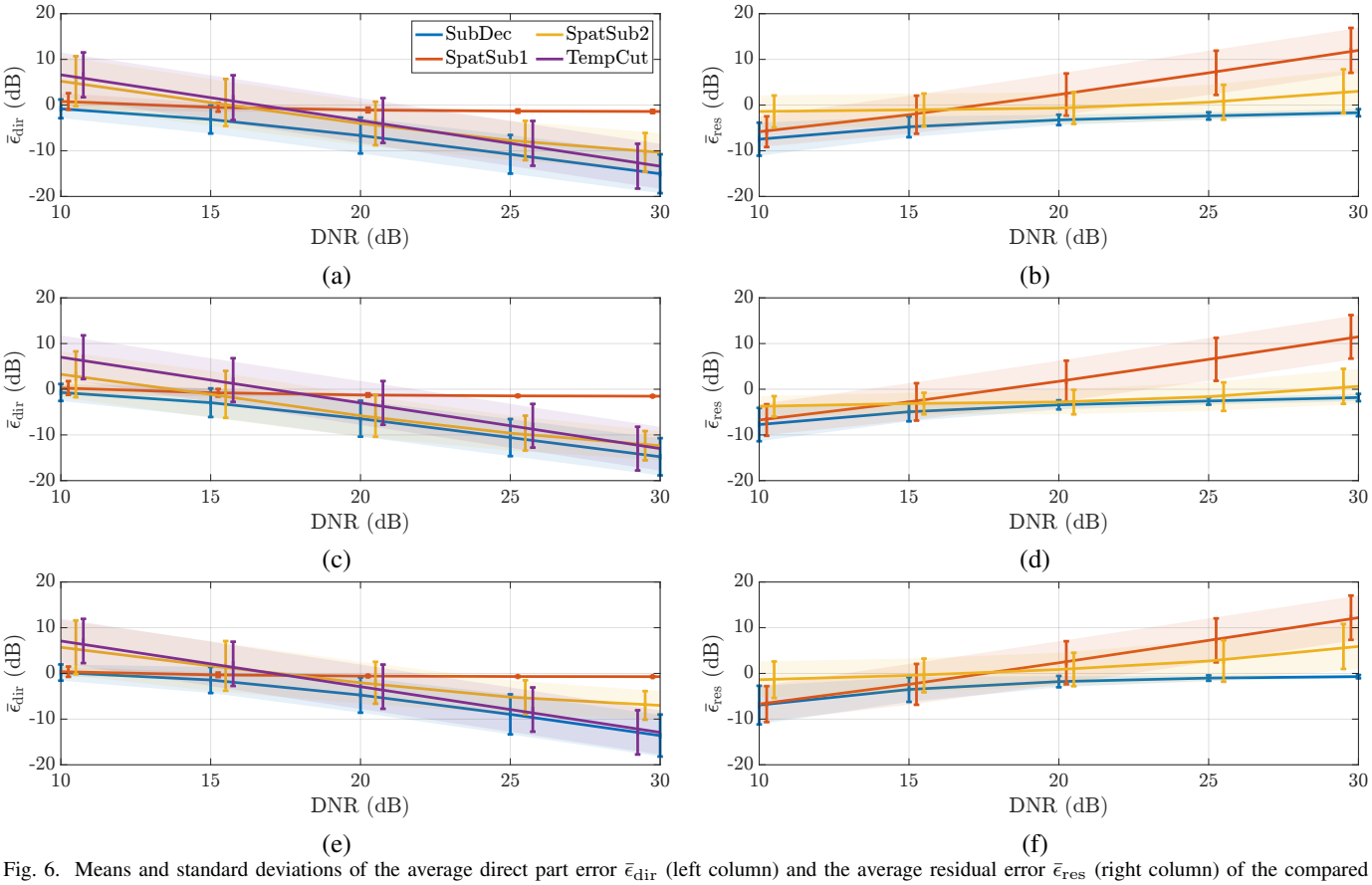


Fig. 6. Means and standard deviations of the average direct part error $\bar{\epsilon}_{\mathrm{dir}}$ (left column) and the average residual error $\bar{\epsilon}_{\mathrm{res}}$ (right column) of the compared methods as a function of the omnidirectional direct-sound-peak-to-residual-noise-RMS ratio (DNR), for, **(a)** and **(b)**, *Array1* with 24 microphones, **(c)** and **(d)**, *Array2* with 32 microphones, and, **(e)** and **(f)**, *Array3* with 48 microphones.

2.6 dB around 6 kHz. The method *SpatSub1* shows similar fluctuations as *SpatSub2* but additionally deviates strongly from the ground truth at high frequencies above 10 kHz.

C. Simulation Study

To systematically evaluate the performance of the proposed method, a simulation study is performed. SMIRGen was again used to simulate the direct part of SRIRs as first-order image sources. No higher-order image sources were calculated to be able to assume that all generated reflections can be considered salient and therefore be assigned to the direct part. This was further achieved by setting the broadband absorption coefficient to 0.3 in all simulations. The simulation study comprises the combination of, (i), 15 shoebox-shaped rooms with random dimensions, (ii), 3 different spherical microphone arrays, (iii) the direct sound and six first-order reflections per SRIR, and, (iv), five different ratios of direct-sound-peak to residual root-mean-square (RMS) energy. The rooms were generated with uniformly-distributed dimensions between $4 \times 4 \times 2$ m and $15 \times 15 \times 10$ m. The source and microphone array positions were randomly generated with the constraints of having a distance of at least 1 m to any room boundary and at least 2 m from each other. Additionally, it was ensured that the generated reflections arrive at the microphone array with a time difference of at least 1 ms so that for the subspace decomposition methods it can be assumed that each subtraction window contains a single reflection. The case of two simultaneously arriving reflections will be investigated in Sec. IV-D. The residual part was generated as noise with the coherence of the simulated arrays in an isotropic diffuse field with a decay of 60 dB per second. An SH decomposition was performed for both the direct part and the residual, and both parts were radial filtered before being added together with varying energy ratios. For this purpose, we define the omnidirectional direct-sound-peak-to-residual-noise-RMS ratio (DNR) that comprises the ratio between the maximum absolute value of the zeroth-order SH channel and the zeroth-order-SH RMS value of the generated residual noise. Note that we define the DNR as a measure per SRIR, meaning that the direct sound of the simulated SRIR is ensured to stand out against the residual RMS but this is not necessarily the case for the six first-order reflections. The DNR was varied in 5 dB steps between 10 dB and 30 dB. The three microphone arrays under test are all rigid, spherical arrays of radii 4.2 cm, 4.2 cm, and 8.5 cm. They comprise 24, 32, and 48 microphones that are arranged according to t-designs and allow for SH decompositions of maximum order 3, 4, and 5. They will also be referred to as *Array1*, *Array2*, and *Array3* in the following.

To facilitate a numerical evaluation, we define a spatio-spectral error measure that comprises the ratio of the norm of the difference of the spectrum of the ground truth reflection

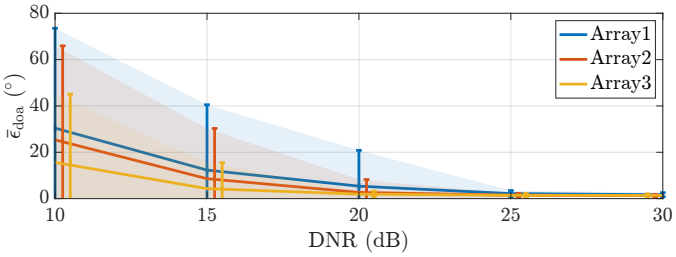


Fig. 7. Mean and standard deviation of the average DOA estimation error $\bar{\epsilon}_{\mathrm{doa}}$ for the three microphone arrays and different DNRs.

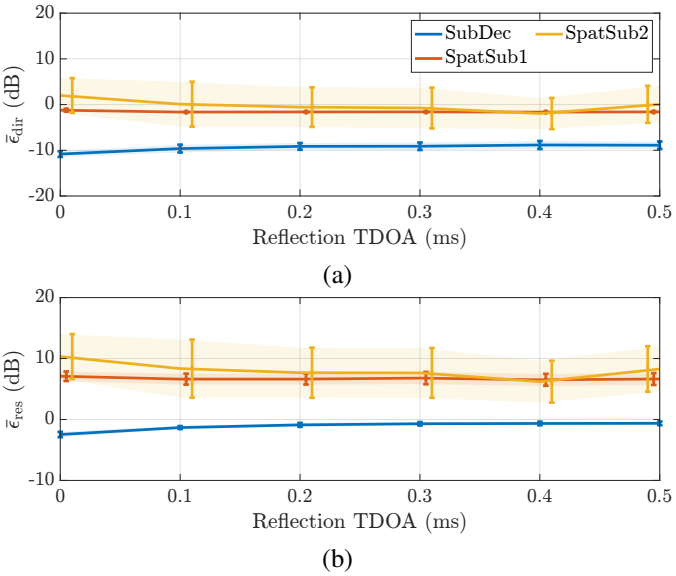


Fig. 8. Means and standard deviations of, **(a)**, the average direct part error $\bar{\epsilon}_{\mathrm{dir}}$ and, **(b)**, the average residual error $\bar{\epsilon}_{\mathrm{res}}$ for two simultaneous reflections with different TDOAs. The simulated array is *Array2*.

$\boldsymbol{\chi}_{\mathrm{s}}^{\mathrm{gt}}(f_b)$ and the spectrum of an extracted reflection $\boldsymbol{\chi}_{\mathrm{s}}(f_b)$, to the norm of the ground truth reflection,

$$\epsilon_{\mathrm{dir}} = \frac{\sum_b \|\boldsymbol{\chi}_{\mathrm{s}}(f_b) - \boldsymbol{\chi}_{\mathrm{s}}^{\mathrm{gt}}(f_b)\|}{\sum_b \|\boldsymbol{\chi}_{\mathrm{s}}^{\mathrm{gt}}(f_b)\|}, \tag{21}$$

where the sum over b denotes the sum over all frequency bins of a 128-point DFT. The division by the norm of the ground truth spectrum ensures that the calculated error is relative to the ground truth energy, i.e., low-energy reflections do not automatically generate a lower error. The spatio-spectral residual error ϵ_{res} is defined similarly by replacing the reflection spectrum $\boldsymbol{\chi}_{\mathrm{s}}(f_b)$ and its ground truth $\boldsymbol{\chi}_{\mathrm{s}}^{\mathrm{gt}}(f_b)$ by the spectrum of the residual $\boldsymbol{\chi}_{\mathrm{n}}(f_b)$ and its ground truth $\boldsymbol{\chi}_{\mathrm{n}}^{\mathrm{gt}}(f_b)$. All spectra are calculated within a $1\,\mathrm{ms}$ window that is centered around the ground truth TOA of the reflection.

As before, the spatial subtraction methods have access to the true TOA of the reflections and use SH-MUSIC for the DOA estimation. The subspace decomposition method is applied without access to any additional information from the ground truth and does not require DOA estimation. It is performed using a block size of 32 samples for *Array1* and *Array2*, and 64 samples for *Array3*, a hop size of $1/8$ of the block size, a residual estimate of $20\,\mathrm{ms}$, GSV averaging of 32 blocks and $\kappa = 4$. To increase the interpretability of the results, another approach is added to the comparison that involves a temporal cut-out of the reflections. It is similarly performed in [22] and, using an omnidirectional RIR, in [57]. The approach is in the following also referred to as *TempCut* and comprises cutting out individual reflections via a $1\,\mathrm{ms}$ window that is centered around the reflection. The cut-out is equally applied to all SH channels and is a straightforward approach that avoids the need for beamforming or a subspace decomposition. However, the *TempCut* approach cannot provide a residual SRIR and is thus only considered in the direct-part comparison of the methods.

Fig. 6 shows the mean and the standard deviation of the average direct part error $\bar{\epsilon}_{\mathrm{dir}}$ and the average residual error $\bar{\epsilon}_{\mathrm{res}}$, i.e., both errors are averaged over the different rooms and over the individual reflections. Errors are shown for the three different microphone arrays and for different DNRs. The proposed subspace decomposition method *SubDec* outperforms the compared methods in all tested cases. The two spatial subtraction methods perform differently, depending on the DNR. For lower DNRs, *SpatSub1* outperforms *SpatSub2* while for higher DNRs *SpatSub2* outperforms *SpatSub1*. For lower DNRs, the residual noise prevents an accurate DOA estimate and distorts the estimate of the reflection spectrum of the underlying plane-wave signal model. These errors more severely influence the results from *SpatSub2* as small inaccuracies have a big influence on the estimate of the spatial aliasing. The mean and standard deviation of the average DOA estimation error $\bar{\epsilon}_{\mathrm{doa}}$ are shown in Fig. 7.

Two overall trends can be observed for all methods: the average direct part errors $\bar{\epsilon}_{\mathrm{dir}}$ tend to decrease with increasing DNR and the average residual errors $\bar{\epsilon}_{\mathrm{res}}$ tend to increase with increasing DNR. With increasing DNR, individual reflections stand out more against the residual and are hence easier to extract. Further, as shown in Fig. 7, the DOA estimates that are required for the spatial subtraction methods get more accurate with higher DNR. For high DNRs, $\bar{\epsilon}_{\mathrm{dir}}$ of the *TempCut* method approaches the results of *SubDec* because at high DNRs the residual energy is negligible in comparison to the direct part energy. At low DNRs, *TempCut* performs worst as the residual dominates over the reflection and a simple temporal cut-out thus creates a large error.

In the case of the average residual errors $\bar{\epsilon}_{\mathrm{res}}$, an increase in error can be observed with increasing DNR. Although the reflections stand out more against the residual with higher DNRs and thus the extraction task becomes simpler, the generated errors increase as the employed error measure, cf. (21), is normalized by the energy of the ground truth residual. At high DNRs, the ground truth residual carries low energy and thus relative errors tend to increase with the DNR.

D. Two Simultaneous Reflections

The reflection density in acoustic environments typically increases exponentially with time. Thus, multiple reflections are likely to occur within one analysis signal block of the different decomposition algorithms. In the following, we refer to multiple reflections within one signal block as simultaneous reflections and investigate the decomposition performance of the different algorithms for two simultaneous reflections. While this does not require a modification of the subspace

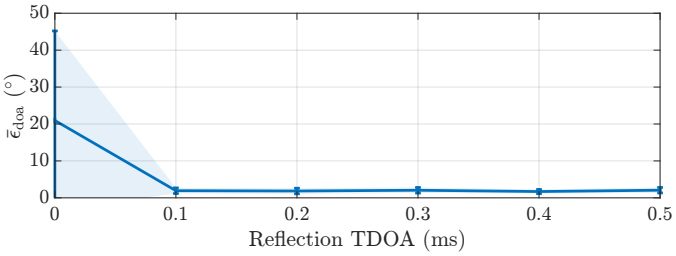


Fig. 9. Means and standard deviations of the average DOA estimation error $\bar{\epsilon}_{\mathrm{doa}}$ for two simultaneously arriving reflections, *Array2* and different TDOAs.

decomposition method, the spatial subtraction method using both signal models is extended as in [28] to support the simultaneous subtraction of multiple reflections. The number of simultaneous reflections and their average time of arrival is assumed to be known by the spatial subtraction methods and the spatial subtraction window of $1\,\mathrm{ms}$ length is centered around the average TOA of the two reflections. We perform simulations for time-differences-of-arrival (TDOAs) of the reflections at the array center between $0\,\mathrm{ms}$ and $0.5\,\mathrm{ms}$. Both reflections are created with the same magnitude and different angles of arrival. For each simulated TDOA, 100 repetitions with random, unique incidence angles that are drawn from the vertices of a dodecahedron are performed. This ensures that the two reflections have at least an angular separation of $41°$. The subspace decomposition is performed using the same parameters as in Sec. IV-C. The simulated rigid, spherical array is the *Array2* from the previous simulation, i.e., its 32 microphones are distributed according to a t-design and a fourth-order SH decomposition is performed. Non-decaying noise with the coherence of the array in an isotropic diffuse field is added to achieve a DNR of $20\,\mathrm{dB}$.

Fig. 8 shows the means and the standard deviations of the average direct part errors $\bar{\epsilon}_{\mathrm{dir}}$ and the average residual errors $\bar{\epsilon}_{\mathrm{res}}$ for TDOAs between $0\,\mathrm{ms}$ and $0.5\,\mathrm{ms}$ and Fig. 9 shows the corresponding average DOA estimation errors $\bar{\epsilon}_{\mathrm{doa}}$. Again, the subspace decomposition method outperforms both spatial subtraction methods in terms of both direct part error and residual error for all TDOAs, although the average DOA estimation error means are equal to or below $2°$ for TDOAs of $0.1\,\mathrm{ms}$ or more. In contrast to the previous simulations of the array with a DNR of $20\,\mathrm{dB}$, cf. Figs. 6 (c) and (d), *SpatSub1* now achieves lower errors than *SpatSub2* except for the case with a TDOA of $0.4\,\mathrm{ms}$. The plane-wave model parameters of the comprehensive signal model of *SpatSub2* cannot be estimated accurately due to the interference of the two reflections, which creates the observed error.

V. Case Study With Measured Spatial Room Impulse Responses

To demonstrate the practical applicability of the proposed method, we apply the subspace decomposition to three SRIRs that were measured in different acoustic environments. The three SRIRs cover a large variety of acoustic conditions and include some variation in terms of the employed microphone arrays. All three SRIRs are publicly available. Binaural renderings of the original SRIRs and the direct part and residual

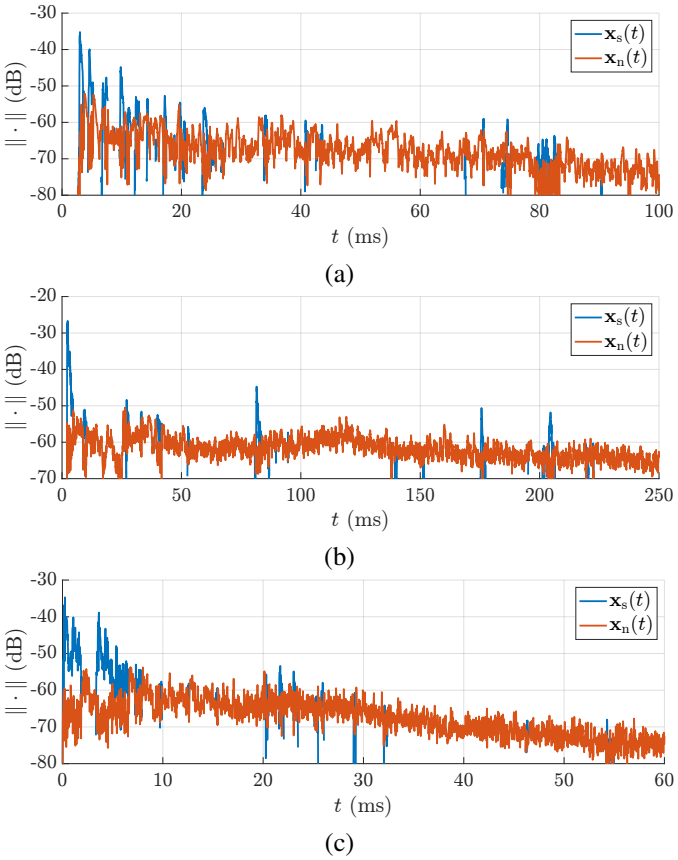


Fig. 10. Norms of the direct part $\boldsymbol{x}_{\mathrm{s}}(t)$ and the residual $\boldsymbol{x}_{\mathrm{n}}(t)$ of SRIRs measured, **(a)**, in a small conference room, **(b)**, in a concert hall, and, **(c)**, in the entrance of an anechoic chamber with the measurement loudspeaker located in the adjacent office.

SRIRs from the subspace decomposition are provided on a companion website[2].

The first SRIR was measured in a $10.3 \times 5.8 \times 3.1$ m conference room with a broadband reverberation time of $0.63\,\mathrm{s}$ [58]. The measurement was performed using the Eigenmike em32 32-channel rigid-sphere microphone array with a radius of $4.2\,\mathrm{cm}$. The subspace decomposition was performed using a block size of 32 samples ($0.7\,\mathrm{ms}$) and a hop size of 4 samples. The residual estimate had a length of $20\,\mathrm{ms}$ and the thresholds were calculated with a GSV averaging length of 64 blocks and using $\kappa = 3$. Fig. 10 (a) shows the direct and residual subspace decomposition for measurement position 2 of the data set.

The second SRIR was measured in a $30.3 \times 16.5 \times 11.6$ m concert hall with a broadband reverberation time of $1.46\,\mathrm{s}$ [59]. It was measured using a sequential 50-channel rigid-sphere microphone array with a radius of $8.75\,\mathrm{cm}$. The subspace decomposition used a block size of 64 samples ($1.3\,\mathrm{ms}$), a hop size of 8 samples and GSV averaging of 48 blocks. The other parameters were the same as previously. Fig. 10 (b) shows the direct and residual subspace decomposition for the measurement that was performed using a PA loudspeaker located at the center of the stage.

The third SRIR was measured at the entrance of an anechoic chamber with the measurement loudspeaker placed within

[2]http://www.ta.chalmers.se/srir-subspace-decomposition/

line-of-sight in the adjacent $6 \times 3.8 \times 2.8$ m office [60]. The measurement again was performed using the Eigenmike em32 microphone array but in this case, the SRIR is provided as 25-channel SRIR in the SH domain. The subspace decomposition was performed in blocks of 32 samples (0.7 ms), with a hop size of 4 samples and the thresholds were calculated with a GSV averaging length of 32 blocks. The other parameters were the same as previously. Fig. 10 (c) shows the direct and residual subspace decomposition for a measurement that was taken 50 cm from the open door inside the anechoic chamber and contains strongly anisotropic reverberation.

In the case of the SRIRs from smaller rooms, cf. Figs. 10 (a) and (c), salient reflections are mainly extracted within the first 30 ms but some weaker reflections are extracted until 100 ms after the direct sound. In the case of the SRIR from the concert hall, Fig. 10 (b), salient reflections are extracted until 200 ms after the direct sound. Although all three SRIRs, stemming from a small conference room, a concert hall, and from the transition between an office and an anechoic chamber, exhibit vastly different reverberation characteristics, the algorithm successfully separates the direct part and the residual in all three cases. Thus, the proposed algorithm proves to be applicable also when using measurement data from a variety of acoustic environments.

VI. Conclusion

In this work, we proposed a subspace method for the decomposition of SRIRs into a direct part, containing the direct sound and salient reflections, and a residual. The method does not rely on a specific microphone array geometry but the array configuration needs to guarantee a singular rank of the covariance matrix in the presence of a plane wave, which, for instance, is shown to be the case for spherical arrangements with a radius of 4.2 cm and 24 or more microphones. The proposed method does not assume a specific wave model and does not rely on corresponding parameter estimates. It outperforms existing methods that rely on DOA estimation, beamforming, and the assumption of plane waves in all simulated scenarios, which include different rooms, microphone arrays, and ratios of direct sound to residual. It further generates lower direct part and residual errors than the compared methods in scenarios with two simultaneous reflections. The proposed subspace decomposition can be applied to SH-domain SRIRs without modification and guarantees the perfect reconstruction of the original SRIR by summing up the direct part and the residual. The method facilitates novel ways of SRIR-based virtual acoustic rendering and might enhance the performance of established parameter estimation methods when applied as pre-processing. A reference implementation is provided at `https://github.com/thomasdeppisch/SRIR-Subspace-Decomposition`.

Acknowledgment

Thomas Deppisch would like to thank Franz Zotter for numerous fruitful discussions about linear algebra.

References

[1] M. Barron, "The subjective effects of first reflections in concert halls-The need for lateral reflections," *Journal of Sound and Vibration*, vol. 15, no. 4, pp. 475–494, 1971.

[2] S. E. Olive and F. E. Toole, "The Detection of Reflections in Typical Rooms," *J. Audio Eng. Soc.*, vol. 37, no. 7/8, pp. 539–553, 1989.

[3] S. Bech, "Timbral aspects of reproduced sound in small rooms. I," *The Journal of the Acoustical Society of America*, vol. 97, no. 3, pp. 1717–1726, 1995.

[4] J. S. Bradley and G. A. Soulodre, "Objective measures of listener envelopment," *The Journal of the Acoustical Society of America*, vol. 98, no. 5, pp. 2590–2597, 1995.

[5] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1832–1846, 2002.

[6] S. Hameed, J. Pakarinen, K. Valde, and V. Pulkki, "Psychoacoustic Cues in Room Size Perception," in *116th Conv. Audio Eng. Soc.*, 2004.

[7] M. Yadav, D. Cabrera, L. Miranda, W. L. Martens, D. Lee, and R. Collins, "Investigating auditory room size perception with autophonic stimuli," in *135th Conv. Audio Eng. Soc.*, 2013.

[8] D. Romblom, C. Guastavino, and P. Depalle, "Perceptual thresholds for non-ideal diffuse field reverberation," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. 3908–3916, 2016.

[9] B. Alary, P. Massé, S. J. Schlecht, M. Noisternig, and V. Välimäki, "Perceptual analysis of directional late reverberation," *The Journal of the Acoustical Society of America*, vol. 149, no. 5, pp. 3189–3199, 2021.

[10] B. Rafaely, *Fundamentals of Spherical Array Processing*, 2nd ed. Springer, 2019.

[11] D. Khaykin and B. Rafaely, "Acoustic analysis by spherical microphone array processing of room impulse responses," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 261–270, 2012.

[12] N. Huleihel and B. Rafaely, "Spherical array processing for acoustic analysis using room impulse responses and time-domain smoothing," *The Journal of the Acoustical Society of America*, vol. 133, no. 6, pp. 3995–4007, 2013.

[13] S. Tervo and A. Politis, "Direction of arrival estimation of reflections from room impulse responses using a spherical microphone array," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 10, pp. 1539–1551, 2015.

[14] B. N. Gover, J. G. Ryan, and M. R. Stinson, "Measurements of directional properties of reverberant sound fields in rooms using a spherical microphone array," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2138–2148, 2004.

[15] M. Berzborn and M. Vorländer, "Investigations on the Directional Energy Decay Curves in Reverberation Rooms," *Proceedings of Euronoise*, p. 2005–2010, 2018.

[16] J. Pätynen, S. Tervo, and T. Lokki, "Analysis of concert hall acoustics via visualizations of time-frequency and spatiotemporal responses," *The Journal of the Acoustical Society of America*, vol. 133, no. 2, pp. 842–857, 2013.

[17] J. Merimaa and V. Pulkki, "Spatial impulse response rendering I: Analysis and synthesis," *J. Audio Eng. Soc.*, vol. 53, no. 12, pp. 1115–1127, 2005.

[18] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *J. Audio Eng. Soc.*, vol. 61, no. 1/2, pp. 17–28, 2013.

[19] M. Zaunschirm, M. Frank, and F. Zotter, "BRIR synthesis using first-order microphone arrays," in *144th Conv. Audio Eng. Soc.*, 2018.

[20] S. V. Amengual Garí, J. M. Arend, P. T. Calamia, and P. W. Robinson, "Optimizations of the spatial decomposition method for binaural reproduction," *J. Audio Eng. Soc.*, vol. 68, no. 12, pp. 959–976, 2020.

[21] L. McCormack, V. Pulkki, A. Politis, O. Scheuregger, and M. Marschall, "Higher-Order Spatial Impulse Response Rendering: Investigating the Perceived Effects of Spherical Order, Dedicated Diffuse Rendering, and Frequency Resolution," *J. Audio Eng. Soc.*, vol. 68, no. 5, pp. 338–354, 2020.

[22] O. Puomio, T. Pihlajakuja, and T. Lokki, "Sound rendering with early reflections extracted from a measured spatial room impulse response," in *Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, 2021.

[23] N. Meyer-Kahlen and S. J. Schlecht, "Parametric Late Reverberation from Broadband Directional Estimates," in *Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, 2021.

[24] P. Coleman, A. Franck, P. J. 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.

[25] P. Stade, J. M. Arend, and C. Pörschmann, "Perceptual evaluation of synthetic early binaural room impulse responses based on a parametric model," in *142nd Conv. Audio Eng. Soc.*, 2017.

[26] F. Brinkmann, H. Gamper, N. Raghuvanshi, and I. Tashev, "Towards encoding perceptually salient early reflections for parametric spatial audio rendering," in *148th Conv. Audio Eng. Soc.*, 2020.

[27] T. Deppisch, J. Ahrens, S. V. Amengual Garí, and P. Calamia, "Spatial Subtraction of Reflections from Room Impulse Responses Measured with a Spherical Microphone Array," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 346–350.

[28] A. Politis, S. Tervo, and V. Pulkki, "COMPASS: Coding and multidirectional parameterization of ambisonic sound scenes," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6802–6806.

[29] T. Deppisch, S. V. Amengual Garí, P. Calamia, and J. Ahrens, "Perceptual Evaluation of Spatial Room Impulse Response Extrapolation by Direct and Residual Subspace Decomposition," in *AES Conference on Audio for Virtual and Augmented Reality (AVAR)*, 2022.

[30] R. O. Schmidt, "Multiple emitter location and parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[31] R. Roy and T. Kailath, "ESPRIT - Estimation of Signal Parameters Via Rotational Invariance Techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[32] C. Eckart and G. Young, "The Approximation of One Matrix by Another of Lower Rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[33] D. W. Tufts, R. Kumaresan, and I. Kirsteins, "Data Adaptive Signal Estimation By Singular Value Decomposition of a Data Matrix," *Proceedings of the IEEE*, vol. 70, no. 6, p. 684–685, 1982.

[34] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, 2002.

[35] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Communication*, vol. 10, no. 1, pp. 45–57, 1991.

[36] Y. Ephraim and H. L. Van Trees, "A Signal Subspace Approach for Speech Enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, 1995.

[37] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 497–507, 2000.

[38] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2230–2244, 2002.

[39] K. Hermus, P. Wambacq, and H. Van Hamme, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *EURASIP J. Adv. Signal Process*, 2007.

[40] Y. Huang, J. Benesty, and J. Chen, *Acoustic MIMO Signal Processing*. Springer, 2006.

[41] W. Herbordt, *Sound Capture for Human/Machine Interfaces*. Springer, 2005.

[42] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. CRC Press, 2013.

[43] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, 2003.

[44] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed. The Johns Hopkins University Press, 2013.

[45] R. H. Hardin and N. J. Sloane, "McLaren's improved snub cube and other new spherical designs in three dimensions," *Discrete and Computational Geometry*, vol. 15, no. 4, pp. 429–441, 1996.

[46] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: Algorithm and applications," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1462–1472, 2012.

[47] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[48] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses," *J. Audio Eng. Soc.*, vol. 60, no. 11, pp. 887–898, 2012.

[49] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, 2008.

[50] M. Wax and T. Kailath, "Detection of Signals by Information Theoretic Criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985.

[51] Z. He, A. Cichocki, S. Xie, and K. Choi, "Detecting the number of clusters in n-way probabilistic clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 2006–2021, 2010.

[52] F. Cong, A. K. Nandi, Z. He, A. Cichocki, and T. Ristaniemi, "Fast and effective model order selection method to determine the number of sources in a linear transformation model," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 1870–1874.

[53] S. Bakamidis, M. Dendrinos, and G. Carayannis, "SVD Analysis by Synthesis of Harmonic Signals," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 472–477, 1991.

[54] S. Moreau, J. Daniel, and S. Bertet, "3D Sound Field Recording with Higher Order Ambisonics - Objective Measurements and Validation of a 4th Order Spherical Microphone," in *120th Conv. Audio Eng. Soc.*, 2006.

[55] B. Rafaely, B. Weiss, and E. Bachmat, "Spatial aliasing in spherical microphone arrays," *IEEE Transactions on Signal Processing*, vol. 55, no. 3, pp. 1003–1010, 2007.

[56] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009.

[57] C. Pörschmann, P. Stade, and J. M. Arend, "Binauralization of omnidirectional room impulse responses - algorithm and technical evaluation," in *Proc. of DAFx-17*, Edinburgh, UK, Sep. 2017, pp. 345–352.

[58] C. Schneiderwind, A. Neidhardt, F. Klein, and S. Fichna, "Data set: Eigenmike-DRIRs, KEMAR 45BA-BRIRs, RIRs and 360° pictures captured at five positions of a small conference room," in *Proc. of the German Annual Conference on Acoustics (DAGA)*, 2019.

[59] P. Stade, B. Bernschütz, and M. Rühl, "A Spatial Audio Impulse Response Compilation Captured at the WDR Broadcast Studios," in *27th Tonmeistertagung - VDT International Convention*, 2012.

[60] T. McKenzie, S. J. Schlecht, and V. Pulkki, "Acoustic Analysis and Dataset of Transitions Between Coupled Rooms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 481–485.

Thomas Deppisch received the M.Sc. degree in Electrical Engineering and Audio Engineering jointly from Graz University of Technology and the University of Music and Performing Arts, Graz, Austria, in 2020. Since 2020, he has been working toward the Ph.D. degree at the Division of Applied Acoustics at Chalmers University of Technology, Gothenburg, Sweden. His main research interests lie in signal processing methods for the virtual reproduction of acoustic environments, including capture, analysis, rendering, and perception.

Sebastià V. Amengual Garí is currently a research scientist at Reality Labs Research working on room acoustics, spatial audio and auditory perception. He received a Diploma Degree in Telecommunications with major in Sound and Image in 2014 from the Polytechnic University of Catalonia (UPC) in 2014, completing his Master's Thesis at the Norwegian University of Science and Technology (NTNU). His doctoral work at the Detmold University of Music focused on investigating the interaction of room acoustics and live music performance using virtual acoustic environments. His research interests lie in the intersection of audio, perception and music.

Paul Calamia is a research science manager on the Audio Team at Reality Labs Research, where he supports research in room acoustics for augmented-reality applications. Previously he was a member of the Technical Staff at MIT Lincoln Laboratory in the Bioengineering Systems and Technologies Group, with a focus on auditory health and hearing protection, and the Advanced Undersea Systems and Technology Group, working on sonar signal processing. His other prior positions include Assistant Professor in the Graduate Program in Architectural Acoustics at Rensselaer Polytechnic Institute in Troy, NY, and Consultant and Head of R&D at Kirkegaard Associates in Chicago, IL. He holds a BS degree in mathematics from Duke University, an MS degree in electrical and computer engineering from the Engineering Acoustics Program at the University of Texas at Austin, and a Ph.D. in computer science from Princeton University.

Jens Ahrens has been an Associate Professor within the Division of Applied Acoustics at Chalmers since 2016. He received a Diplom (equivalent to a M.Sc.) in Electrical Engineering and Audio Engineering jointly from Graz University of Technology and the University of Music and Performing Arts, Graz, Austria, in 2005. He completed his Doctoral Degree (Dr.-Ing.) at the Technische Universität Berlin, Germany, in 2010. From 2011 to 2013, he was a Postdoctoral Researcher at Microsoft Research in Redmond, Washington, USA, and in the fall and winter terms of 2015/16, he was a Visiting Scholar at the Center for Computer Research in Music and Acoustics (CCRMA) at Stanford University, California, USA. He is an Associate Editor of the IEEE Signal Processing Letters and of the EURASIP Journal on Audio, Speech, and Music Processing.