Signal Compaction Using Polynomial EVD for Spherical Array Processing with Applications

Vincent W. Neo [ORCID], *Member, IEEE,* Christine Evers [ORCID], *Senior Member, IEEE*
Stephan Weiss [ORCID], *Senior Member, IEEE*, and Patrick A. Naylor [ORCID], *Fellow, IEEE*

Abstract—Multi-channel signals captured by spatially separated sensors often contain a high level of data redundancy. A compact signal representation enables more efficient storage and processing, which has been exploited for data compression, noise reduction, and speech and image coding. This paper focuses on the compact representation of speech signals acquired by spherical microphone arrays. A polynomial matrix eigenvalue decomposition (PEVD) can spatially decorrelate signals over a range of time lags and is known to achieve optimum multi-channel data compaction. However, the complexity of PEVD algorithms scales at best cubically with the number of channel signals, e.g., the number of microphones comprised in a spherical array used for processing. In contrast, the spherical harmonic transform (SHT) provides a compact spatial representation of the 3-dimensional sound field measured by spherical microphone arrays, referred to as eigenbeam signals, at a cost that rises only quadratically with the number of microphones. Yet, the SHT's spatially orthogonal basis functions cannot completely decorrelate sound field components over a range of time lags. In this work, we propose to exploit the compact representation offered by the SHT to reduce the number of channels used for subsequent PEVD processing. In the proposed framework for signal representation, we show that the diagonality factor improves by up to 7 dB over the microphone signal representation with a significantly lower computation cost. Moreover, when applying this framework to speech enhancement and source separation, the proposed method improves metrics known as short-time objective intelligibility (STOI) and source-to-distortion ratio (SDR) by up to 0.2 and 20 dB, respectively.

Index Terms—Data compaction, polynomial matrix eigenvalue decomposition, speech enhancement, spherical harmonics, source separation.

I. INTRODUCTION

IN multi-channel signal processing involving spatially separated sensors, the received signals often contain a high level of data redundancy. Many compact signal representation techniques such as subband coding have been developed to improve storage and processing efficiency [1]. The processing of these compacted signals offers computational advantages and has been widely used for data compression [2], noise reduction [3], and speech and image coding [4], [5]. This

work focuses on the compact representation of speech signals measured by spherical microphone arrays.

A data-adaptive signal representation using an infinite-order principal component filterbank (PCFB) has been shown to be optimal in terms of the mean-square error in signal reconstruction and the coding gain in data compression [6]. A PCFB generally requires a frequency-dependent switching of channels [2], whereby the switching function is not analytic and therefore cannot be well approximated with a finite degree PCFB [7], [8]. If the order of the PCFB is constrained to zero-th order, the Karhunen-Loève transform (KLT), i.e. an eigenvalue decomposition (EVD), gives the optimal solution [6].

An alternative approach uses polynomial matrices, which can simultaneously capture the correlations in space, time and frequency and is, therefore, appropriate for modelling multi-channel broadband signals [9]. The processing of polynomial matrices has motivated the development of polynomial matrix eigenvalue decomposition (PEVD) algorithms in the z-domain based on the second-order sequential best rotation (SBR2) [10] and sequential matrix diagonalization (SMD) [11], [12], and those in the discrete Fourier transform (DFT)-domain [13], [14]. Unlike the KLT, the PEVD can mutually decorrelate signals for all lags [11]; this strong decorrelation together with spectral majorization of the decorrelated sequences guarantees optimality in the coding gain sense [2] Thus, the PEVD presents a solution for a finite order PCFB. The PEVD has also been found useful in multi-channel broadband applications such as source separation [15], source identification [16], localization [17], adaptive beamforming [18] and voice activity detection [19], [20].

The use of the PEVD for speech enhancement has been proposed in [21] for arbitrary arrays. It has been shown to improve noise reduction metrics, dereverberation measures, speech intelligibility and speech quality scores in diverse acoustic environments without introducing noticeable artifacts. This is due to the fact that PEVD algorithms [10]–[12] — while using z-domain notation— operate in the time domain, and therefore maintain spectral coherence of the signals being processed. However, the computational complexity of such PEVD algorithms scales at best cubically with the number of signals used for processing [22].

Spherical array processing [23], [24] has gained significant attention owing to its applicability to hearing aids, sound field decomposition and reproduction for personal sound zones, augmented and virtual reality [25]–[27], and robot audition [28]. The microphone array in these applications is commonly modelled as a spherical array along a rigid sphere. Spherical

microphone arrays enable a compact spatial representation of the sound field in the spherical harmonic (SH) domain using eigenbeam signals. Because the eigenbeam signals are computed using only the array geometry and are decoupled from the sensor arrangements, the beam pattern can be designed to be rotationally invariant, non-data-adaptive, and independent of the number of microphones, making the processing scalable [29]. This is why beamforming using spherical arrays [29]–[31] has been proposed for speech enhancement [32], [33], localization and tracking [34]–[36].

A 3D sound field can be perfectly represented by an infinite number of spatially orthogonal SHs (of infinite order) [24]. In practice, a finite number of microphones limiting the SH order leads to an approximate representation such that it is generally not possible to reconstruct 3D sound fields with high spatial accuracy. Further, reverberation arising from multipaths in enclosed spaces causes target source components in eigenbeam signals from different directions to arrive at other times [37]. Consequently, the spherical harmonic transform (SHT) cannot decorrelate a reverberant sound field for non-zero time delays [25].

In this current manuscript, the representation of microphone signals on a spherical array using SH is first compared to a PEVD representation. This will demonstrate the inability of SH to capture temporal correlations, and the challenge of scaling with the number of signals used for PEVD processing. We then propose to combine both approaches by applying a PEVD to a subset of eigenbeam signals generated from preprocessing by a SHT. This capitalizes on the compact and scalable representation offered by SH, and also exploits the ability of the PEVD to strongly decorrelate signals, i.e., to generate outputs that are mutually decorrelated at all lags. We propose a unified general processing framework to incorporate the spherical microphone array processing motivated by signal compression and representations. We substantiate the motivations using theoretical analysis and experimental validations between the various signal representations, and demonstrate how the current proposed framework based on signal representations can be used for specific applications including blind speech enhancement and informed source separation using target source directions based on our preliminary studies in [38] and [39]. Therefore, in supplement to the earlier studies, the novel contributions of this paper are (i) the comparison of different representations of multi-channel signals from a spherical microphone array signals using the SHT, KLT, and PEVD, (ii) the proposed framework that relates the microphone signals, eigenbeams and beamformer outputs through a compression matrix, (iii) the use of a PEVD to combine eigenbeams and beamformer outputs under the proposed framework, and (iv) the application of the proposed approach for speech enhancement and source separation.

The paper first provides an exposition of the problem in Section II. Section III reviews and theoretically analyses the processing and signal representation afforded by SHT and PEVD for spherical microphone arrays. The proposed approach to combine SHT and PEVD is presented in Section IV and subsequently used for speech enhancement and source separation in Section V. Results are summarised in Section VI, experimentally validating our analysis. Conclusions are drawn in Section VII.

II. PROBLEM FORMULATION AND OVERVIEW

A. Signal Model

The noisy and reverberant speech signals arriving at the q-th microphone on the spherical array at sample index n, are

$$x_q(n) = \sum_{p=1}^{P} h_{p,q}(n) * s_p(n) + v_q(n), \quad q = 1, \ldots, Q, \quad (1)$$

where $h_{p,q}(n)$ is the time-invariant room impulse response (RIR) from the p-th source to the q-th microphone, $s_p(n)$ is the p-th source signal, $v_q(n)$ is the additive noise assumed uncorrelated with the speech component, and $*$ denotes the linear convolution operator. The model in (1) accounts for P sources contributing to Q microphone signals. The data vector of the microphone signals is $\mathbf{x}(n) = \left[x_1(n), \ldots, x_Q(n)\right]^{\mathrm{T}}$ with $\mathbf{v}(n)$ similarly defined, whereby $[\cdot]^{\mathrm{T}}$ denotes the transpose operator.

To express the microphone signals in terms of the array geometry explicitly, (1) can be written as $x(n, \mathbf{r}_q)$, where $\mathbf{r}_q = (r, \theta_q, \phi_q)$ is expressed in spherical polar coordinates, r is the radius of the sphere, θ_q and ϕ_q respectively, are the elevation and azimuth angles of the q-th microphone from the array centre measured downwards from the z-axis and from the x-axis towards the y-axis. Accordingly, the data vector is $\mathbf{x}(n, \mathbf{r}) = \left[x(n, \mathbf{r}_1), \ldots, x(n, \mathbf{r}_Q)\right]^{\mathrm{T}}$.

B. Challenge

The number of sources, P, is generally much smaller than the number of microphones, Q. Therefore, this work aims to find a compact signal representation of the microphone signals $\mathbf{x}(n)$ by taking into account the spatial, spectral and temporal information simultaneously using a transformation $\mathbf{G}(n) \in \mathbb{R}^{P \times Q}$ such that $y_p(n) = \sum_{q=1}^{Q} g_{p,q}(n) * x_q(n)$, where $\mathbf{y}(n) = [y_1(n), \ldots, y_P(n)]^{\mathrm{T}}$ is maximally compact and $g_{p,q}(n)$ is the $(p, q)^{\mathrm{th}}$ element in $\mathbf{G}(n)$. While $\mathbf{G}(n)$ is ideally square and unitary, i.e., $P = Q$, it is more efficient to process P principal components for $P < Q$ in practice when the remaining $(Q - P)$ processed signals have small energy compared to the P principal components. The assumption is that the mixture of P linearly independent signals in $\mathbf{x}(n)$ can be compacted into P non-zero components in the processed outputs $y_p(n)$, $p = 1, \ldots, P$. These signals are mutually orthogonalized by $\mathbf{G}(n)$; as such, they do not necessarily align with the original source signals but can permit a reduction in dimensionality together with a potential diagonalization of the signal covariance matrix. This can facilitate more efficient processing for applications such as signal enhancement or the separation of sources.

III. MULTI-CHANNEL ARRAY PROCESSING

A. Spherical Harmonic Transform (SHT)

The real-valued SHT of the spatially sampled sound field is approximated by [31]

$$\chi_\ell^{(m)}(n) = \sum_{q=1}^{Q} \alpha_q x(n, \mathbf{r}_q) \Upsilon_\ell^{(m)}(\mathbf{r}_q) , \qquad (2)$$

where α_q is the quadrature weight for the q-th microphone and $\chi_\ell^{(m)}(n)$ is the ℓ-th order, m-th degree time-domain eigenbeam. The latter is associated with the real-valued SH basis function, $\Upsilon_\ell^{(m)}(\mathbf{r}_q)$, defined as

$$\Upsilon_\ell^{(m)}(\mathbf{r}) = \begin{cases} \sqrt{2}(-1)^m \Im\{Y_\ell^{(|m|)}(\mathbf{r})\} & m < 0 \\ Y_\ell^0(\mathbf{r}) & m = 0 , \\ \sqrt{2}(-1)^m \Re\{Y_\ell^{(m)}(\mathbf{r})\} & m > 0 \end{cases} \quad (3)$$

based on the complex-valued SH basis expressed as

$$Y_\ell^{(m)}(\mathbf{r}) = \sqrt{\frac{(2\ell+1)}{4\pi} \frac{(\ell-m)!}{(\ell+m)!}} P_\ell^{(m)}(\cos\theta) e^{jm\phi} , \quad (4)$$

where $P_\ell^{(m)}(\cdot)$ is the associated Legendre function, $|\cdot|$ denotes modulus, and $\Re\{\cdot\}$ and $\Im\{\cdot\}$ extract the real and imaginary parts of a complex number. The coefficients $\chi_\ell^{(m)}(n)$ are also called eigenbeams because SH are eigen-solutions of the wave equation in spherical coordinates [24], [29], [40]–[42]. Because of the completeness and orthogonality properties of SH, any function in Hilbert space comprising the set of all square integrable functions on the unit sphere can be expressed as a weighted combination of SHs using [24]

$$x(n, \mathbf{r}_q) \approx \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} \chi_\ell^{(m)}(n) \Upsilon_\ell^{(m)}(\mathbf{r}_q) . \qquad (5)$$

Equality in (5), and therefore an alias-free representation, is achieved if $Q \geq (L+1)^2$, where L is the maximum SH order of the sound field. In practice, a finite order SHT leads to a small approximation error; with a sufficient number of microphones, this error — and hence the inequality in (5) — is negligible for typical applications.

Dropping the sample index n for brevity, the vector of eigenbeams is $\boldsymbol{\chi} = [\chi_0^{(0)}, \chi_1^{(-1)}, \chi_1^{(0)}, \ldots, \chi_L^{(L)}]^T \in \mathbb{R}^{(L+1)^2}$, with elements arranged in ascending SH order and degree. More compactly,

$$\boldsymbol{\chi} = \boldsymbol{\Upsilon}^T \mathrm{diag}(\boldsymbol{\alpha}) \mathbf{x}(n, \mathbf{r}) \qquad (6)$$

is written in matrix-vector form, where

$$\boldsymbol{\Upsilon} = \begin{bmatrix} \Upsilon_0^{(0)}(\mathbf{r}_1), & \ldots, & \Upsilon_L^{(L)}(\mathbf{r}_1) \\ \vdots & \ddots & \vdots \\ \Upsilon_0^{(0)}(\mathbf{r}_Q) & \ldots & \Upsilon_L^{(L)}(\mathbf{r}_Q) \end{bmatrix} \in \mathbb{R}^{Q \times (L+1)^2} , \quad (7)$$

and $\mathrm{diag}(\boldsymbol{\alpha})$ creates a diagonal matrix from the vector of microphone weights, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_Q]^T$.

The generation of $\mathcal{L} = (L+1)^2$ eigenbeams using all microphone signals in (2) is also called an eigen-beamformer [29], [42]. Modal beamforming, or the judicious linear combination

of the eigenbeams (see Section V for elaboration), produces a beam pattern directed at a desired source direction using

$$\psi(n) = \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} w_\ell^{(m)} \chi_\ell^{(m)}(n) , \qquad (8)$$

where $w_\ell^{(m)}$ is the beamformer weight associated with $\chi_\ell^{(m)}(n)$. In vectorial form, $\psi(n) = \mathbf{w}^T \boldsymbol{\chi}$, where $\mathbf{w} = \left[w_0^{(0)}, w_1^{(-1)}, w_1^{(0)}, \ldots, w_L^{(L)}\right]^T$. With $\mathscr{P}$ different combinations of weights, the beamformer can generate $\mathscr{P}$ outputs

$$\boldsymbol{\psi} = \mathbf{W}^T \boldsymbol{\chi} , \qquad (9)$$

where $\boldsymbol{\psi} = [\psi_1, \ldots, \psi_{\mathscr{P}}]^T$, $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_{\mathscr{P}}]$ and $\mathbf{w}_p$ contains the set of weights associated with ψ_p, $p = 1, \ldots, \mathscr{P}$. Ideally, the selected beams should mainly contain the target source signal while minimizing any unwanted signals.

B. Polynomial Matrix Eigenvalue Decomposition

The PEVD approach considers the space-time covariance matrix [10], parameterized by a time lag τ, which is defined as

$$\mathbf{R}_{\mathbf{xx}}(\tau) = \mathbb{E}\{\mathbf{x}(n)\mathbf{x}^T(n-\tau)\} , \qquad (10)$$

where $\mathbb{E}\{\cdot\}$ is the expectation operator over n. Each element, $r_{p,q}(\tau)$, is computed using the cross-correlation sequence between the p-th and q-th microphone signals. Therefore, $\mathbf{R}_{\mathbf{xx}}(\tau)$ contains auto- and cross-correlations on its diagonals and off-diagonals, respectively.

Classical subspace-based approaches for narrowband signals typically consider a special case of (10) by evaluating $\mathbf{R}_{\mathbf{xx}}(\tau)$ only at $\tau = 0$. The received signals are then decorrelated using an EVD. The resulting instantaneous spatial covariance matrix, $\mathbf{R}_{\mathbf{xx}}(0)$, does not fully capture the second-order statistics of the sensor signals, and has been shown to be not fully adequate for broadband signals such as speech [17], [21], which naturally exhibit temporal correlations especially in reverberant environments. Consequently, the proposed PEVD approach considers the decorrelation of speech signals over a range of discrete time lags. Accordingly, the concatenation of the covariance matrices in (10) for all values of $\tau \in \{-N, \ldots, N\}$ can be represented by a 3D-tensor of dimension $Q \times Q \times (2N+1)$.

Speech signals are typically processed in the short-time Fourier transform domain. However, this approach divides the broadband into multiple narrowband signals, ignoring the spectral coherence or correlation that exists between different DFT bins, and neglecting phase coherence across bands [43], [44]. As an alternative representation, the z-transform of (10),

$$\mathscr{R}_{\mathbf{xx}}(z) = \mathscr{Z}\{\mathbf{R}_{\mathbf{xx}}(\tau)\} = \sum_{\tau=-\infty}^{\infty} \mathbf{R}_{\mathbf{xx}}(\tau) z^{-\tau} , \qquad (11)$$

or $\mathbf{R}_{\mathbf{xx}}(\tau) \; \circ\!\!-\!\!\bullet \; \mathscr{R}_{\mathbf{xx}}(z)$, is a para-Hermitian Laurent polynomial matrix satisfying $\mathscr{R}_{\mathbf{xx}}(z) = \mathscr{R}_{\mathbf{xx}}^P(z) = \mathscr{R}_{\mathbf{xx}}^H(1/z^*)$, where $[\cdot]^*$, $[\cdot]^H$ and $[\cdot]^P$ are the complex conjugate, Hermitian and para-Hermitian operators. The PEVD of (11) is [10]

$$\mathscr{R}_{\mathbf{xx}}(z) \approx \mathcal{U}_{\mathbf{x}}(z)\boldsymbol{\Lambda}_{\mathbf{x}}(z)\mathcal{U}_{\mathbf{x}}^P(z), \qquad (12)$$

where the columns of $\mathcal{U}_{\mathbf{x}}(z)$ are the eigenvectors and the elements on the diagonal matrix $\mathbf{\Lambda}_{\mathbf{x}}(z)$ are the eigenvalues. To compute (12), iterative approaches based on the SBR2 [10] and the SMD [11], [12] families of algorithms have been proposed, which are proven to converge, and encourage spectral majorization such that the eigenvalues in $\mathbf{\Lambda}_{\mathbf{x}}(z)$ are strictly ordered on the unit circle. The decomposition in (12) holds with equality for an analytic $\mathcal{R}_{\mathbf{xx}}(z)$ with analytic factors $\mathcal{U}_{\mathbf{x}}(z)$ and $\mathbf{\Lambda}_{\mathbf{x}}(z)$, which in most cases are infinite Laurent series [45]. Analyticity implies that an arbitrarily close approximation by Laurent polynomials in (12) is possible with a sufficiently large order of factors [46].

At each iteration, the SBR2 PEVD algorithm [10] first searches for the off-diagonal element with the largest magnitude. If its magnitude exceeds a predefined threshold, a delay polynomial matrix is applied to bring the element to the z^0 plane. A unitary matrix, which is designed to zero out the element, is applied to the entire polynomial matrix. A trimming procedure [10] is also used to keep the polynomial order compact. The algorithm terminates when the magnitudes of all off-diagonal elements fall below a pre-set threshold, or when a user-defined maximum number of iterations is reached.

C. Comparison of Signal Representations

The theoretically optimal compaction filter $\mathbf{G}(n)$, whose coding gain is guaranteed by its strong decorrelation and spectral majorization properties [2], [7], is allowed to have infinite order and is also the principal component filter bank used for data compression [2]. These principal components could then be extracted for speech enhancement and source separation, where the signals of interest reside. For a spherical array geometry, the two practical designs of $\mathbf{G}(n)$ include the SHT and the PEVD. The SHT provides a compact spatial representation of the 3D sound field sampled by microphones on a spherical array while the PEVD can compactly represent space-time information in the signal eigenspace.

1) Analysis of Signal Representations: In noisy, reverberant environments, each microphone signal on a spherical array comprises the combination of many delayed and attenuated versions of the speech signal due to multi-path propagation. Consequently, microphone signals are highly correlated in space and time, as evident from a non-diagonalized space-time covariance matrix. The PEVD can diagonalize the space-time covariance matrix. Strong decorrelated signals can be obtained using the polynomial eigenvectors as filters. Selecting only the subspaces corresponding to polynomial eigenvalues of significant magnitude leads to a signal representation based on the PEVD that uses fewer signals than the number of microphone signals. When there are P independent, spatially separate sources, the ideal PEVD is expected to generate P out of Q non-zero outputs; while these outputs very likely do not match the source signals, a compaction of the data is still achieved.

For a spherical array, the SHT decomposes a 3D sound field using spatially orthogonal basis functions to generate the eigenbeam signals. Due to multi-paths, sound field components in different eigenbeams exhibit temporal correlations.

Consequently, an EVD or KLT, which is typically applied to the spatial covariance matrix $\mathbf{R}(0)$, only removes instantaneous correlations i.e., at time lag $\tau = 0$, and is insufficient for processing (10). Instead, the PEVD can be used to achieve better compression and complete diagonalization across a range of time lags, as will be illustrated in Section VI-C.

2) Measures of Signal Representation Quality: An instrumental measure of compactness is the coding gain γ, which is the ratio between the arithmetic and geometric mean of the variances of the elements in $\mathbf{y}(n)$, computed using [1], [7],

$$\gamma = \frac{\frac{1}{N}\mathrm{tr}\{\mathbf{R}_{\mathbf{yy}}(0)\}}{\det\{\mathbf{R}_{\mathbf{yy}}(0)\}^{\frac{1}{N}}} , \qquad (13)$$

where $\mathrm{tr}\{\cdot\}$ and $\det\{\cdot\}$ compute the trace and determinant of a matrix. Ideally, $\mathbf{G}(n)$ should be a para-unitary or lossless filterbank, such that the total powers of $\mathbf{x}(n)$ and $\mathbf{y}(n)$ remain the same. In this case, for $\mathbf{y}(n)$ to attain maximum coding gain, two conditions must be satisfied [47]:

1) strong decorrelation, such that $\mathbb{E}\{y_i(n)y_j(n-\tau)\} = 0$, $\forall \tau, i \neq j$, i.e., the elements of $\mathbf{y}(n)$ are decorrelated for all lags τ;
2) spectral majorization such that for all normalized angular frequencies Ω,

$$S_{y_{i-1}y_{i-1}}(e^{j\Omega}) \geq S_{y_i y_i}(e^{j\Omega}), \; i = 2, \ldots, Q, \quad (14)$$

where $S_{y_i y_i}(e^{j\Omega})$ is the power spectral density of $y_i(n)$. Both conditions are met by PEVD algorithms such as SBR2 and SMD, where strong decorrelation, and in case of SBR2 also spectral majorization [48], are proven to be enforced.

The quality of the signal representation can be quantified by the reconstruction error ε, expressed as a percentage, using

$$\varepsilon = \frac{\sum_{q=1}^{Q}(\hat{x}(n,\mathbf{r}_q) - x(n,\mathbf{r}_q))^2}{\sum_{q=1}^{Q}(x(n,\mathbf{r}_q))^2} \times 100\% , \qquad (15)$$

that is, the total squared error between the received x and reconstructed microphone signals $\hat{x}$, normalized by the total energy of the signals in Q microphones. The reconstructed microphone signals $\hat{\mathbf{x}}(n)$ are computed using (5) for SHT and $\hat{\mathbf{x}}(n) = \sum_{k,\nu} \mathbf{U}_{\mathbf{x}}(k)\mathbf{U}_{\mathbf{x}}^{\mathrm{T}}(k-\nu)\mathbf{x}(n-\nu)$ for the PEVD.

Since with iterative PEVD algorithms, the diagonalization in (12) is only approximate, similarly to [10], we define the total on- and off-diagonal energies E_{on} and E_{off} in the ideally diagonalized eigenvalue term $\boldsymbol{S}[\tau] \approx \mathbf{\Lambda}_{\mathbf{x}}(\tau)$ as

$$E_{\mathrm{on}} \triangleq \sum_{\forall \tau} \sum_{i=1}^{Q} |s_{ii}(\tau)|^2, \quad E_{\mathrm{off}} \triangleq \sum_{\forall \tau} \|\boldsymbol{S}[\tau]\|_{\mathrm{F}}^2 - E_{\mathrm{on}} , \; (16)$$

where $s_{ii}[\tau]$ is the i-th diagonal element of $\boldsymbol{S}[\tau]$, and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. Using (16), the diagonality factor δ, which measures the level of diagonalization of a matrix $\boldsymbol{S}[\tau]$, is defined as

$$\delta = 5 \log_{10}\left(\frac{E_{\mathrm{on}}}{E_{\mathrm{off}}}\right) , \qquad (17)$$

expressed in dB. The unusual scaling by 5 accounts for the fact that the terms of $\boldsymbol{S}[\tau]$ in (17) are already of a quadratic nature [7]. The lower and upper bounds, $0 \leq E_{\mathrm{on}}/E_{\mathrm{off}} < \infty$, correspond to the case when only the off-diagonal elements
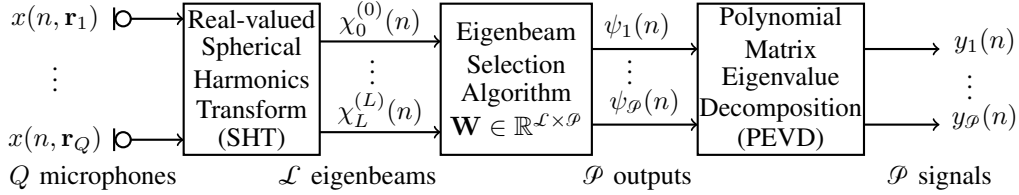
Fig. 1: Block diagram of the proposed method which uses combinations of eigenbeams for PEVD processing.

are non-zero and the case when the polynomial matrix is fully diagonalized, respectively.

IV. PEVD IN THE SPHERICAL HARMONIC DOMAIN

A. Motivation for Proposed Approach

While PEVD addresses the optimum coding gain problem in multi-channel systems [7], it is data-dependent and, in practice, likely to face numerical difficulties for a large number of sensors, Q [49], as is the case with the spherical microphone array here. The space-time covariance matrix in (11), on which the PEVD operates, contains $Q^2(2N + 1)$ elements and give rise to a computational complexity that is at least $\mathcal{O}(Q^3 N)$ due to matrix multiplications applied to every lag [22]. Hence, it will be computationally advantageous to reduce the number of channels Q if the accuracy of the signal representation is not significantly affected.

In contrast to the PEVD, the SHT scales well w.r.t. the spatial dimension. The normalized truncation or approximation error depends only on the product between wavenumber k and radius r [29], [30]. Particularly for a desired truncated SH order L, the product $L = kr$ gives a normalized truncation error of roughly 4% for all values of kr, which is sufficient for most practical applications [50]. Further, for a sufficiently large SH order L, instead of using Q microphones, a lower dimension of $\mathcal{L}$ eigenbeams can be used to represent the sound field. In terms of compaction, particularly for sound sources arriving at the spherical microphone array from well-defined directions, the SHT outputs can be linearly combined via beamforming to further compact the source's power, and is therefore widely regarded as a tool for dimensionality reduction of spherical microphone array data [25], [51], [52]. This compaction is difficult to assess via the coding gain, since the SHT is not a unitary transform or otherwise norm-preserving. We can state, though, that the SHT is sub-optimal w.r.t. the coding gain in (13), as it only operates instantaneously. Consequently, it does not target decorrelation over a range of time lags and cannot remove correlations for other noise types and reverberation exhibiting temporal correlations.

With a view to combine the benefits of SHT and PEVD, we propose to utilize a preprocessor comprising the SHT together with a beamformer prior to applying a PEVD. This preprocessor performs a spatial decomposition, with the aim of compacting as much energy as possible into as few outputs as necessary while exploiting the benefits of modal beamforming in terms of scalability and efficiency. Note that when the source directions are unknown and beamforming is not performed, the signal-independent SHT, which has a low computational

complexity, performs a dimensionality reduction. Regardless, the SHT preprocessor reduces the number of signals used in the PEVD and significantly reduces the complexity compared to the PEVD-only approach in Section III-B. Our method is summarized in Fig. 1 and Algorithm 1.

B. Proposed Algorithm

Assuming stationarity, the space-time covariance matrix of the modal beamformer outputs $\boldsymbol{\psi}(n)$ can be computed using

$$\mathbf{R}_{\boldsymbol{\psi\psi}}(\tau) = \mathbb{E}\{\boldsymbol{\psi}(n)\boldsymbol{\psi}^{\mathrm{T}}(n - \tau)\} . \quad (18)$$

Using (2) and (9), the z-transform of (18), $\mathcal{Z}\{\boldsymbol{R}_{\boldsymbol{\psi\psi}}(\tau)\}$, is

$$\begin{aligned}
\mathcal{R}_{\boldsymbol{\psi\psi}}(z) &= \mathbf{W}^{\mathrm{T}}\mathcal{R}_{\boldsymbol{\chi\chi}}(z)\mathbf{W} \\
&= \mathbf{W}^{\mathrm{T}}\boldsymbol{\Upsilon}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{\alpha})\mathcal{R}_{\mathbf{xx}}(z)\mathrm{diag}(\boldsymbol{\alpha})\boldsymbol{\Upsilon}\mathbf{W} \\
&= \mathbf{C}^{\mathrm{T}}\mathcal{R}_{\mathbf{xx}}(z)\mathbf{C} ,
\end{aligned} \quad (19)$$

where $\mathcal{R}_{\boldsymbol{\chi\chi}}(z) \; \bullet\!\!-\!\!\circ \; \mathbf{R}_{\boldsymbol{\chi\chi}}(\tau)$ is the space-time covariance matrix of the eigenbeam signals. The compression matrix $\mathbf{C} = \mathrm{diag}(\boldsymbol{\alpha})\boldsymbol{\Upsilon}\mathbf{W}$ depends on the microphone weights $\boldsymbol{\alpha}$, eigenbeam weights $\mathbf{W}$, and the SH order, L. The design and impact of $\mathbf{C}$ will be discussed in Section V.

In practice, (18) can be estimated with bias using

$$\mathbf{R}_{\boldsymbol{\psi\psi}}(\tau) \approx \frac{1}{T} \sum_{n=0}^{T-1} \boldsymbol{\psi}(n)\boldsymbol{\psi}^{\mathrm{T}}(n - \tau) , \quad (20)$$

where T is the frame size and (19) is approximated by

$$\mathcal{R}_{\boldsymbol{\psi\psi}}(z) \approx \sum_{\tau=-W}^{W} \mathbf{R}_{\boldsymbol{\psi\psi}}(\tau)z^{-\tau} . \quad (21)$$

The estimation error incurred in (20) depends on both the ground truth $\mathbf{R}_{\boldsymbol{\psi\psi}}(\tau)$ and frame size T [53]. In turn, the time support W, over which the estimation is performed, can be optimized to balance truncation and estimation errors [54].

The PEVD of the modal beamformer outputs in (19) is

$$\mathcal{R}_{\boldsymbol{\psi\psi}}(z) \approx \mathcal{U}_{\boldsymbol{\psi}}(z)\boldsymbol{\Lambda}_{\boldsymbol{\psi}}(z)\mathcal{U}_{\boldsymbol{\psi}}^{\mathrm{P}}(z) . \quad (22)$$

The eigenvector filterbank, $\mathcal{U}_{\boldsymbol{\psi}}(z) \; \bullet\!\!-\!\!\circ \; \mathbf{U}_{\boldsymbol{\psi}}(n)$ is para-unitary by construction, i.e. $\mathcal{U}_{\boldsymbol{\psi}}^{\mathrm{P}}(z)\mathcal{U}_{\boldsymbol{\psi}}(z) = \mathcal{U}_{\boldsymbol{\psi}}(z)\mathcal{U}_{\boldsymbol{\psi}}^{\mathrm{P}}(z) = \mathbf{I}$. Thus, $\mathcal{U}_{\boldsymbol{\psi}}(z)$ can only redistribute spectral power among channels and not change the total signal powers. The generated outputs are also spectrally majorized and sorted in descending order of their signal energy [10]. Consequently, in the case when the spectrally majorized solution corresponds to the target source, this signal can be extracted from the first output in $\mathbf{y}(n)$ using

$$y_1(n) = \sum_{k=0}^{K} \mathbf{u}_1^{\mathrm{T}}(-k)\boldsymbol{\psi}(n-k) \,, \qquad (23)$$

where $\mathbf{u}_1(n)$ represents the first column of $\mathbf{U}_{\boldsymbol{\psi}}(n)$ and is of order K. The filtered outputs $\mathbf{y}(n)$ are strongly decorrelated due to diagonalization achieved by $\mathscr{R}_{\mathbf{yy}}(z) \approx \boldsymbol{\Lambda}_{\boldsymbol{\psi}}(z)$.

C. Modal Beamformer Designs for Eigenbeams

The spatial dimension of $\mathscr{R}_{\mathbf{xx}}(z)$ is Q. For a sufficiently large SH order L, the rank of the SH basis matrix $\boldsymbol{\Upsilon}$ is $\mathscr{L}$. Consequently, the SHT generates $\mathscr{L}$ eigenbeam signals and its space-time covariance polynomial matrix, $\mathscr{R}_{\boldsymbol{\chi}\boldsymbol{\chi}}(z)$ in (19), has a spatial dimension of $\mathscr{L}$. Furthermore, if prior information such as direction-of-arrivals (DoAs) is available, $\mathbf{W}$ in (19) may no longer be a square matrix, e.g., $\mathscr{P} \leq \mathscr{L}$, and contains modal beamformer weights as its elements. The resulting $\mathscr{R}_{\boldsymbol{\psi}\boldsymbol{\psi}}(z)$ has dimension $\mathscr{P}$, thereby achieving further compression. The design of $\mathbf{W}$ depends on the application, and this will be demonstrated for speech enhancement and source separation in Section V. Therefore, the complete proposed practical design for $\mathbf{G}(n)$ is $\mathbf{U}_{\boldsymbol{\psi}}^{\mathrm{T}}(n)\mathbf{C}^{\mathrm{T}}$.

Algorithm 1 PEVD processing in the SHT domain.

Inputs: $\mathbf{x}(n) \in \mathbb{R}^Q, \alpha, L, \mathbf{W}, T, W$.
$\quad \boldsymbol{\chi}(n) \leftarrow \mathbf{x}(n)$ // *real-valued SHT, see* (2)
$\quad \boldsymbol{\psi}(n) \leftarrow \mathbf{W}^{\mathrm{T}}\boldsymbol{\chi}(n)$ // *modal beamforming*
$\quad \mathbf{R}_{\boldsymbol{\psi}\boldsymbol{\psi}}(\tau) \leftarrow E\{\boldsymbol{\psi}(n)\boldsymbol{\psi}^{\mathrm{T}}(n-\tau)\}$ // *see* (20)
$\quad \mathscr{R}_{\boldsymbol{\psi}\boldsymbol{\psi}}(z) \leftarrow \mathscr{Z}\{\mathbf{R}_{\boldsymbol{\psi}\boldsymbol{\psi}}(\tau)\}$ // *see z-transform* (21)
$\quad \mathcal{U}_{\boldsymbol{\psi}}(z), \boldsymbol{\Lambda}_{\boldsymbol{\psi}}(z) \leftarrow \mathrm{PEVD}\{\mathscr{R}_{\boldsymbol{\psi}\boldsymbol{\psi}}(z)\}$ // *use SMD* [11]
$\quad y_1(n) \leftarrow \mathrm{filter}\{\mathbf{u}_1^{\mathrm{T}}(n), \boldsymbol{\psi}(n)\}$ // *extract target, see* (23)
$\quad$ **return** $y_1(n)$.

V. APPLICATION EXAMPLES

The proposed compaction system is used for two illustrative application examples. The first is a speech enhancement application using the proposed system to enhance the signal of a single speaker in noise, where the beamformer operates akin to a KLT to compact energy. The second application is source separation of two speakers. The target speaker direction is assumed to be known to inform the construction of several beamformers which coherently combine the target speaker but not the second source. This enables the extraction of the target speaker as the principal component in the PEVD stage.

A. Single Source Speech Enhancement

The problem of unsupervised speech enhancement for a single speaker is now discussed. A significant amount of data redundancy is expected when sampling the sound field of a single source by several microphones at different spatial positions. In fact, the signal subspace due to a single speech source in $\boldsymbol{\Lambda}_{\mathbf{x}}(z)$ or $\boldsymbol{\Lambda}_{\boldsymbol{\psi}}(z)$, is expected to be of rank one, thereby suggesting an opportunity for compression. If the direction of the source is not known, the full set of eigenbeams generated from the SHT can be chosen up to the L-th order
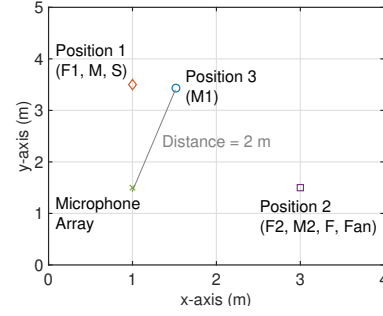

Fig. 2: Setup of the $4\,\mathrm{m} \times 5\,\mathrm{m} \times 6\,\mathrm{m}$ simulated room.

SH for PEVD processing. This is equivalent to using $\mathbf{W} = \mathbf{I}$, where $\mathbf{I} \in \mathbb{R}^{\mathscr{L} \times \mathscr{L}}$ is the identity matrix and $\mathbf{C} \in \mathbb{R}^{Q \times \mathscr{L}}$ only depends on the SH order approximation if all microphones are used for processing. The speech enhancement performance using different sets of complete eigenbeam signals will be investigated in Section VI-C3.

B. Separation of Multiple Sources

The problem of separating multiple directional sources in a reverberant environment is next considered. Applying a PEVD for data compaction directly to the microphone signals is prohibitive in terms of computational complexity; further, due to spectral majorization, the PEVD may extract a mix of signal components that is counterproductive when performing source separation. For this reason, we assume that the target source direction is either known or can be estimated [55], [56], such that only a subset of eigenbeam signals is required, and can be further compressed by a modal beamformer. This reduces the input dimension to the PEVD stage, and also aligns the target signal while incoherently combining the secondary, interfering source(s). Thus, the PEVD becomes computationally viable and can compact the target source in its principal eigenvectors.

With $\mathscr{P}$ beamformer outputs for each target source, $\mathbf{C} \in \mathbb{R}^{Q \times \mathscr{P}}$ represents the dimensionality reduction. Consequently, the spectrally majorized outputs generated by the PEVD enable the extraction of the target source in the first channel. This process can be repeated for all sources. The source separation performance will be demonstrated in Section VI-C4.

VI. SIMULATIONS AND RESULTS

To demonstrate the benefits of the proposed framework, first a comparison of signal representations using theoretical and realistic examples generated from measured data illustrates the achievable compression and practical benefits. The proposed framework is then used for blind speech enhancement and informed source separation using the approaches discussed in Section V. Listening examples are available at [57].

A. Experimental Setup

The TIMIT corpus [58] provides 16 kHz anechoic speech signals. For each speech source, short utterances from the same randomly selected speaker were concatenated to generate signals of 8 to 10 s duration. In experiments involving simulated rooms, the SMIRgen tool in [59] was used to generate the

RIRs for a 32 microphone spherical array. The room setup is shown in Fig. 2, and the room reverberation time T_{60} [37] was varied between 0 s, 0.3 s and 0.7 s. Spatially and temporally white Gaussian noise was used to simulate sensor noise. In experiments using measured RIRs, Lecture Room 2 with $T_{60} = 1.22$ s and babble noise signals, which were recorded using the 32-channel Eigenmike spherical array [60], were taken from the ACE corpus [61].

In each experiment, 50 trials were conducted for which the speech signals were varied. For each trial, each anechoic speech signal was convolved with the RIRs before adding noise to generate the microphone signals at a specified input SNR using [62]. The SNR ranged from -10 dB to 20 dB.

The SMD algorithm was used to compute a PEVD using $\rho = N_1/\sqrt{q} \times 10^{-2}$ for the maximum off-diagonal column-norm, where $q = 32$ is the number of signals used for processing, $N_1 = \text{tr}\{\mathbf{R}(0)\}$ is the total energy in those q signals, a trim factor $\mu = 10^{-3}$, 500 iterations and $T = W = 1600$ samples. These parameters are selected following [21]. The quadrature weights $\boldsymbol{\alpha}$ are computed using [63]. We now describe the three tests.

1) Compression and Signal Representation Comparison: The microphone signals were directly processed using the KLT and PEVD. Addtionally, SHT was applied in order to generate eigenbeam signals (SHT), with subsequent processing of eigenbeams using KLT (SHT+KLT) and PEVD (SHT+PEVD).

2) Single Source Speech Enhancement: The source was located at Position 2 in Fig. 2. The proposed approach using the complete set of eigenbeams for $L = 1$ (PEVD L1) and $L = 2$ (PEVD L2) are compared against the PEVD-based algorithm which uses the raw microphone signals (RAW PEVD) [21]. Consequently, 4 and 9 eigenbeam signals are passed as inputs to the PEVD algorithm. Eigenbeams $\chi_0^{(0)}$ and $\chi_1^{(1)}$ are also evaluated because the former can provide some noise reduction [24] and the latter represents a dipole directed at the source for the simulated room. To compare the ability of the PEVD to strongly decorrelate signals, the KLT was also applied to the single eigenbeam $\chi_0^{(0)}$ in the time-domain using [64] (KLT$\{\chi_0^{(0)}\}$) since spatial and temporal decorrelation are separately achieved by using the SHT and KLT, respectively.

3) Source Separation: For each target source, $\mathscr{P} = 4$ instead of 32 microphone signals were used as inputs to the PEVD to demonstrate its effectiveness while reducing computational complexity. The modal signals for the source at Position 1 were $\chi_1^{(-1)}, \chi_3^{(-3)}, \chi_3^{(-1)}$ and the modified hyper-cardioid (MHCARD) [39] beam directed at $(\frac{\pi}{2}, \frac{\pi}{2})$. For the source at Position 2, the modal signals were $\chi_1^{(1)}, \chi_3^{(1)}, \chi_3^{(3)}$ and the MHCARD directed at $(\frac{\pi}{2}, 0)$. The proposed approach was compared against well-known approaches such as the third-order hyper-cardioid modal beamformer optimized for maximum directivity index (MaxDir) [29], auxiliary function-based independent vector analysis (AuxIVA) [65], independent low-rank matrix analysis (ILRMA) [66] and fast multi-channel non-negative matrix factorization (MNMF) (FastMNMF) [67], implemented in [68]. During experimentation, it was found that the independent component analysis (ICA) and MNMF-based algorithms did not perform well when all 32 micro-

TABLE I: RESULTS FOR CODING GAIN EXAMPLE.

Processing	γ [dB]
Unprocessed microphones	0
KLT on microphones	14.6
PEVD on microphones	18.3
Eigenbeams (SHT on microphones)	10.9
KLT on Eigenbeams	14.9
PEVD on Eigenbeams	19.1
Optimal	**18.7**

phones were used. Therefore, signals from two microphones closest to each source were chosen for these methods in the following results.

B. Evaluation Metrics

To measure the quality of the signal representation, the reconstruction error ε, coding gain γ and diagonality factor δ, defined in Section III-C2, are computed. The PEVD complexity factor β relative to Q microphones, estimated using $\beta = (\frac{\mathcal{L}}{Q})^3$, quantifies the computational savings.

Speech enhancement is popularly evaluated using segmental signal to noise ratio (SegSNR) and frequency-weighted SegSNR (FwSegSNR) for noise reduction [69], normalized signal-to-reverberant ratio (NSRR) and Bark spectral distortion (BSD) for dereverberation [70], short-time objective intelligibility (STOI) for speech intelligibility [71] and perceptual evaluation of speech quality (PESQ) [72] for quality.

To evaluate the source separation performance, source-to-distortion ratio (SDR), source-to-interferences ratio (SIR), and source-to-artifacts ratio (SAR) are used to measure the overall source separation ability, interference rejection and processing artifacts, respectively [73].

The metrics are computed for the microphone signals and the processed signals, and their difference Δ is reported for each measure. Positive Δ values indicate improvements in all measures except ΔBSD, for which a negative value implies a reduction in spectral distortions.

C. Experiments and Discussion

1) Theoretical Example for Compression: To demonstrate the achievable compaction of the various signal representations, a single uncorrelated source is considered, i.e., $P = 1$, illuminating a spherical array with $Q' = 36$ microphones, of which initially only $Q = 25$ elements are utilized. The propagation environment is anechoic, and the array is assumed to be sufficiently small to not suffer from attenuation loss across the array. Thus, the array signal can be expressed as in (1), with $\mathbf{h}_1(n) \in \mathbb{R}^Q$ consisting of fractional delay filters [74], [75], and a normalization such that, for $\mathbf{h}_1(n) \; \circ\!\!-\!\!\bullet \; \boldsymbol{\hbar}_1(z)$, $\boldsymbol{\hbar}_1^{\mathrm{P}}(z)\boldsymbol{\hbar}_1(z) = 1$. With a source power of σ_s^2, and spatially and temporally uncorrelated noise of power σ_v^2 corrupting each microphone, the SNR at each microphone is $\rho = \sigma_s^2/\sigma_v^2$.

The ideal PEVD uses a para-unitary polynomial matrix $\boldsymbol{\mathcal{U}}(z) = [\boldsymbol{\hbar}_1(z)\boldsymbol{\mathcal{H}}(z)]^{\mathrm{P}}$, such that $\boldsymbol{\mathcal{H}}^{\mathrm{P}}(z)\boldsymbol{\hbar}_1(z) = \mathbf{0}$, with $\boldsymbol{\mathcal{H}}(z) : \mathbb{C} \to \mathbb{C}^{Q \times (Q-1)}$ generated via a para-unitary matrix completion method [18]. Then, the polynomial eigenvalues are

$$\lambda_m(z) = \begin{cases} Q\mathcal{S}(z) + \sigma_v^2 & m = 1, \\ \sigma_v^2 & m > 1, \end{cases} \tag{24}$$
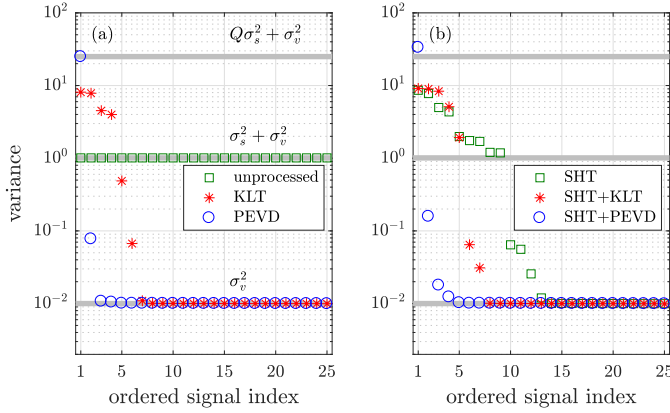
Fig. 3: Variances achieved by various data compaction approaches based on (a) unprocessed microphone signals and (b) eigenbeams produced by the SHT. Grey horizontal lines show the upper and lower variances provided by the ideal PEVD in (24), and the variance of the unprocessed microphone signals.

where $\mathcal{S}(z) \bullet\!\!-\!\!\circ S(\tau)$ and $S(\tau)$ is the auto-correlation function of the source signal. With $\mathcal{S}(z) = \sigma_s^2$, the optimal coding gain can be shown to take the form

$$\gamma_{\mathrm{opt}} = \frac{\sigma_s^2 + \sigma_v^2}{\sqrt[Q]{(Q\sigma_s^2 + \sigma_v^2)(\sigma_v^2)^{Q-1}}} = \frac{1+\rho}{\sqrt[Q]{1+Q\rho}} \ . \qquad (25)$$

With $Q = 25$ and $\rho = 100$, i.e. an SNR of $20\,\mathrm{dB}$, $\gamma_{\mathrm{opt}} = 18.7\,\mathrm{dB}$. A KLT, performing an optimum instantaneous spatial decorrelation, achieves a coding gain of $\gamma_{\mathrm{KLT}} = 14.6\,\mathrm{dB}$, while an approximate PEVD via the SMD algorithm applied to the microphone data yields $\gamma_{\mathrm{SMD}} = 18.3\,\mathrm{dB}$ because of diagonalization achieved over a range of time lags. The approximate nature of the SMD algorithm leads to a small loss in coding gain compared to γ_{opt} shown in Table I, but provides significantly better coding gain than the KLT. Fig. 3(a) illustrates the variances of the microphone signals, $(\sigma_s^2 + \sigma_v^2)$, and the outputs of the KLT and PEVD, demonstrating the effect of compaction with the KLT packing the source signal into approximately six and the PEVD into approximately two outputs before the noise floor of σ_v^2 is reached.

Since the SHT is generally not orthonormal, $Q' = 36$ microphone signals were used to generate a reduced set of $Q = 25$ eigenbeams, i.e., a 4th order SH representation. This can be achieved by a transform matrix $\mathbf{S} \in \mathbb{R}^{Q \times Q'}$, with $\mathbf{S}\mathbf{S}^{\mathrm{T}} = \mathbf{I}$ while noting that $\mathbf{S}^{\mathrm{T}}\mathbf{S} \neq \mathbf{I}$. If the input is spatially uncorrelated and $\mathbf{x}(n) = \mathbf{S}\mathbf{x}'(n)$, then $\mathbb{E}\{\|\mathbf{x}(n)\|_2^2\} = \frac{Q}{Q'}\mathbb{E}\{\|\mathbf{x}'(n)\|_2^2\}$. For spatially correlated data, such as the above source signal, no input-output relation for the energy can be stated: in the best case, the signal $\mathbf{x}'(n)$ occupies the row space of $\mathbf{S}$; in the worst case, it will lie in the nullspace of $\mathbf{S}$ resulting in zero power for $\mathbf{x}(n)$. The reconstruction error ϵ for 25 eigenbeams is 5.7%.

Therefore, due to the incomplete preservation of power in the SHT, the coding gain cannot yield a direct comparison. Ignoring this, Table I shows that the SH signal representation yields a coding gain of $10.9\,\mathrm{dB}$ over the unprocessed microphone signals. A KLT or PEVD operating on the eigenbeams increases this coding gain further. Fig. 3(b) shows that the
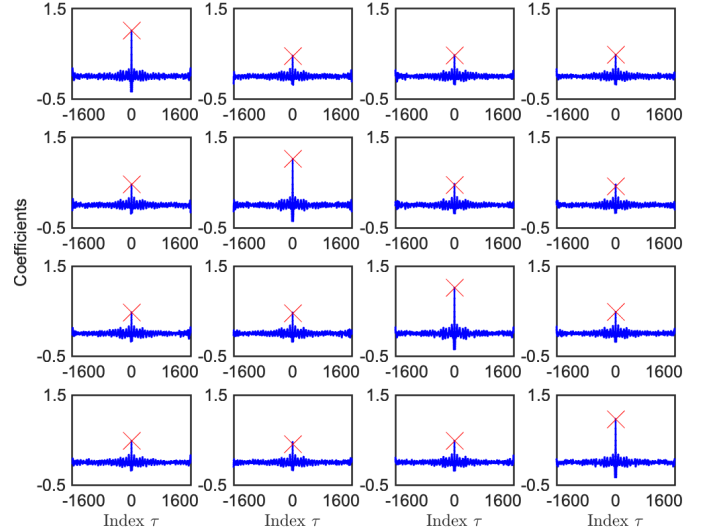

Fig. 4: Portion of $\mathbf{R}_{\mathbf{xx}}(\tau)$ corresponding to the first 4 signals from the Eigenmike with $\mathbf{R}_{\mathbf{xx}}(0)$ marked by red crosses.
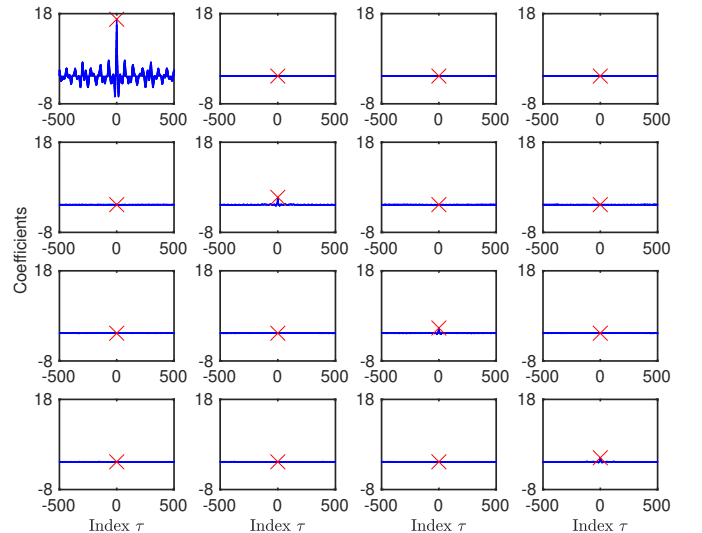

Fig. 5: First 4 principal eigenvalues in $\mathbf{\Lambda}_{\mathbf{x}}(z)$ computed from $\mathcal{R}_{\mathbf{xx}}(z)$ in Fig. 4 and red crosses are the zero-lag terms.

SHT+PEVD coding gain exceeds the optimum value according to (25), because the best-compacted signal breaks the upper limit in (24). Importantly, though, Fig. 3(b) illustrates three key aspects of our proposed approach: (i) without being data adaptive, the SHT still manages to compress the signal into approximately 12 eigenbeams; (ii) processing this reduced set of eigenbeams further by a KLT or PEVD yields similar compression when compared to the results in Fig. 3(a); thus, (iii) particularly the PEVD can operate on a significantly reduced spatial dimension and therefore at a much reduced computational cost after SHT preprocessing.

2) Signal Representation Example Using Measured Data: To compare the signal representation of speech signals captured by spherical microphone arrays in practice, an illustrative
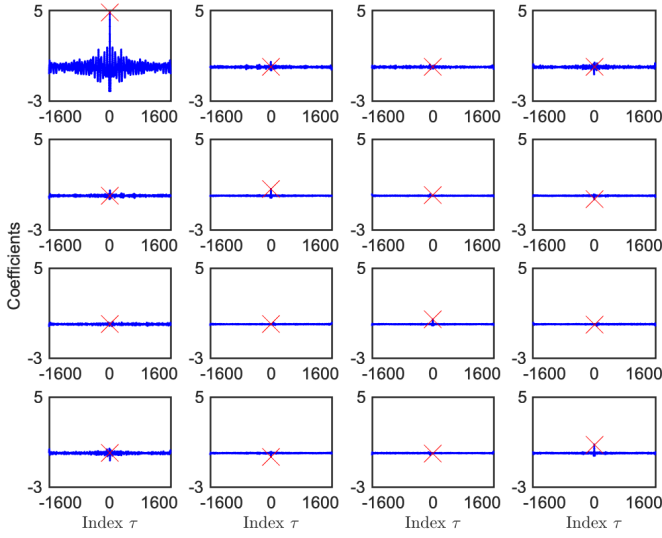
Fig. 6: $\mathscr{R}_{\chi\chi}(z)$ of eigenbeams for $L = 1$ computed using signals that generate Fig. 4. Red crosses are the zero-lag terms.
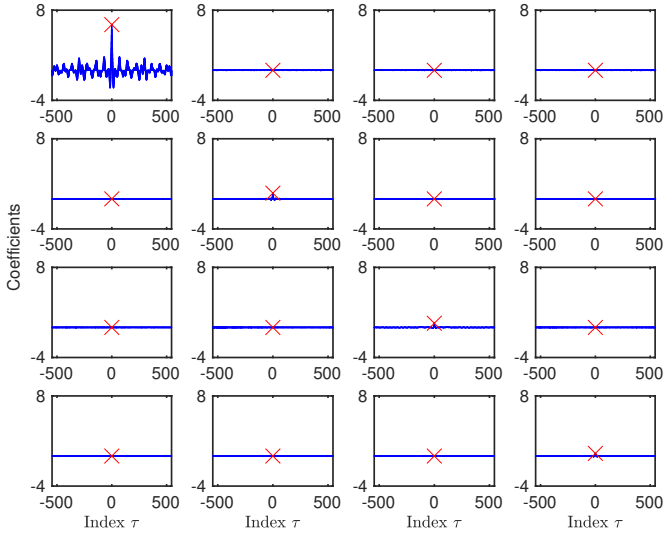

Fig. 7: $\mathbf{\Lambda}_{\psi}(z)$ of eigenbeam signals in Fig. 6 and red crosses represent the zero-lag terms.

example is generated using RIRs and 10 dB babble noise measured by the 32-element Eigenmike in ACE Lecture Room 2. A portion of the space-time covariance matrix corresponding to the first 4 microphones in Fig. 4 shows that correlations exist at non-zero lags. The PEVD is used to generate outputs which are spatially decorrelated over a range of time lags, resulting in a space-time covariance matrix equivalent to $\mathbf{\Lambda}_{\mathbf{x}}(z)$ in Fig. 5.

With $Q = 32$ microphones, the sound field can be approximated up to SH order 4. Note that in the theoretical example of compression achieved by different signal representations in Section VI-C1, the number of channels before and after processing is the same, i.e. 25 channels. Practical microphone arrays like the Eigenmike, however, are usually spatially over-

TABLE II: Signal representation using SHT for 10 dB babble noise Lecture Room example in Fig. 4.

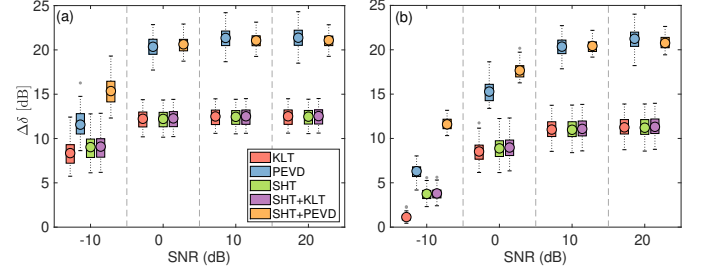| L	$\mathcal{L}$	Complexity, β	Error, $\varepsilon(\%)$	δ [dB]	γ [dB]
0	1	-	31.7	-	-
1	4	0.002	13.3	5.16	3.03
2	9	0.022	8.4	5.00	5.89
3	16	0.125	5.4	4.93	7.03
4	25	0.477	2.3	4.87	7.16

Fig. 8: Diagonality factor improvement for (a) white noise in simulated room with $T_{60} = 0.3$ s and (b) babble noise in ACE Lecture Room 2.

sampled as they use more microphones than $(L + 1)^2$, represented by an over-determined non-square (tall) compression matrix $\mathbf{C} \in \mathbb{R}^{Q \times (L+1)^2}$.

While the complexity β increases with L, the reconstruction error ϵ and coding gain γ generally decrease as seen in Table II. For this example, even with $L = 2$, ε is 8.4%, indicating that 9 eigenbeams can capture most of the spatial information. This suggests that the raw microphone signals may be highly redundant, and that the SHT can offer a significant level of compression for an order-limited sound field. For a small number of $\mathcal{L}$ signals, β is also significantly lower and is omitted for $\mathcal{L} = 1$ since the PEVD is a multi-channel algorithm. For reference, applying PEVD on $\mathcal{L} = 25$ eigenbeams yields coding gain $\gamma = 31.5$ dB.

When L increases from 1 to 4, δ reduces from 5.16 dB to 4.87 dB indicating that there is a decrease in the on-diagonal energy E_{on} relative to the off-diagonal energy E_{off}, i.e., the matrix for larger L is less diagonal. This is expected because most higher-order eigenbeams do not contain the target speech component and will not exhibit correlations between many pairs of eigenbeam signals. Some higher-order eigenbeams, however, may exhibit temporal correlations with lower-order eigenbeams pointing in the same direction, as seen from the peaks between $\chi_0^{(0)}(n)$ and $L = 1$ in Fig. 6 occurring at $\tau \neq 0$. Since SH are spatially orthogonal, the eigenbeams may be instantaneously decorrelated, as shown by the red crosses in Fig. 6. To reduce the correlations between eigenbeams across the range of τ, a PEVD is used and leads to $\mathbf{\Lambda}_{\chi}(z)$ in Fig. 7.

The diagonality factor δ defined in Section III-C2 is computed and compared with the processed signals for 50 trials. The results for white noise in the simulated room and babble noise in Lecture Room 2 are shown in Fig. 8. In both cases the difference in diagonality, $\Delta\delta$ is highest by up to 22 dB for the PEVD-based methods because PEVD can diagonalize the space-time covariance matrix over a range of time lags. At -10 dB SNR, however, the proposed SHD+PEVD improves

diagonalization by 7 dB over PEVD using the microphone signals directly. SHD processing alone provides a substantial improvement in diagonalization by up to 12 dB in $\Delta\delta$.

While a different time segment is used in each trial for the recorded babble noise, the speakers generating the babble noise are seated in the same positions and the noise signals are not perfectly diffuse. Therefore, SHT, which performs a spatial decomposition, can be an advantageous preprocessing step and performs more consistently, i.e., a larger improvement in the mean of SHT+PEVD than the PEVD in Fig. 8(b), and a smaller variance for SHT+PEVD in babble noise in Fig. 8(b) than the white noise scenario in Fig. 8(a).

3) Single Source Speech Enhancement: Table III(a) compares the speech enhancement performance for a single example using the methods presented in Section VI-A2. In this simulated setup, the PEVD-based algorithms perform similarly, with PEVD L2 performing best in most metrics. PEVD L1 gives a greater improvement in ΔFwSegSNR and ΔBSD by 5.72 dB and -1.68 dB compared to RAW PEVD, which uses all microphone signals directly. After PEVD L2, the $\chi_1^{(1)}(n)$ eigenbeam directed at the source gives the best ΔSTOI. Applying the KLT to $\chi_0^{(0)}(n)$ improves all measures except for ΔSTOI.

Speech enhancement performance for 50 Monte-Carlo trials in the simulated room is shown in Fig. 9. Across all measures and SNRs, the PEVD-based algorithms (RAW PEVD, PEVD L1, PEVD L2) are ranked first or second after KLT$\{\chi_0^{(0)}(n)\}$, which is optimal for white noise, or the $\chi_1^{(1)}(n)$ eigenbeam, which is pointing directly at the source. The two PEVD-based algorithms perform comparably well, even though different numbers of channels are used for processing. This shows that processing the eigenbeams instead of the raw microphone signals for speech enhancement is effective and computationally advantageous, and is without the need of DoA information.

When recorded signals from the ACE corpus are used, Fig. 10 shows that RAW PEVD provides the greatest improvement across all metrics for SNR≥ 0 dB and is closely followed by PEVD L2 and PEVD L1 even without using any DoA information. The $\chi_0^{(0)}(n)$ eigenbeam shows some improvement and the use of KLT does not offer further improvement, and may even introduce processing artifacts when the noise is not white. This can be seen from a slight reduction in ΔSTOI and ΔPESQ in Fig. 10(c) and Fig. 10(d) for babble noise. The $\chi_1^{(1)}(n)$ eigenbeam does not perform well because it is not directly pointing at the source and may only pick up weaker reverberant components along with noise.

At -10 dB SNR, the PEVD algorithms which use the eigenbeams, PEVD L1 and PEVD L2, perform better than RAW PEVD. In very noisy environments, the generated eigenbeams take advantage of the noise reduction offered by the signal independent SH domain processing, which is not available to RAW PEVD. At other SNRs, they perform comparably even though PEVD L1 and PEVD L2 use 4 and 9 channels respectively compared to the 32 channels used in the RAW PEVD approach.
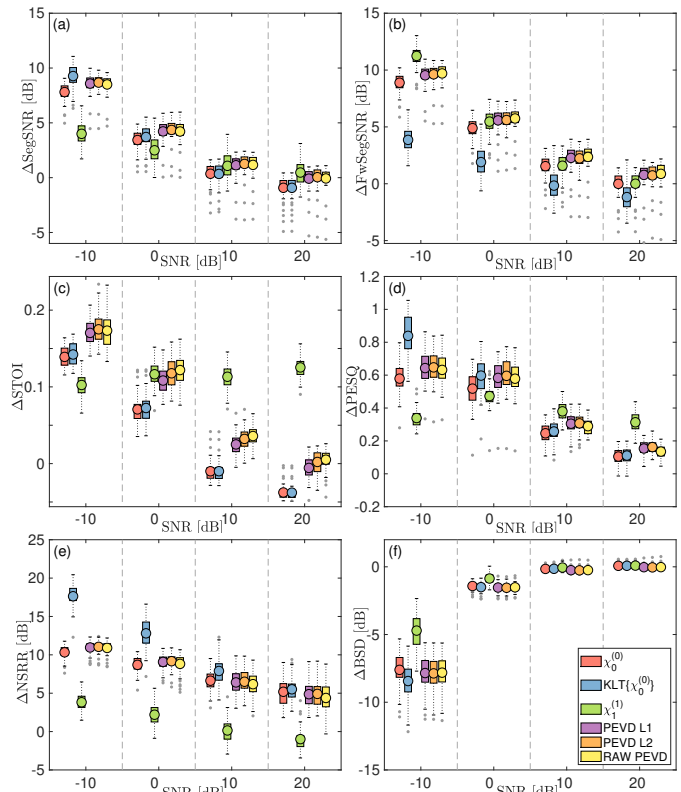

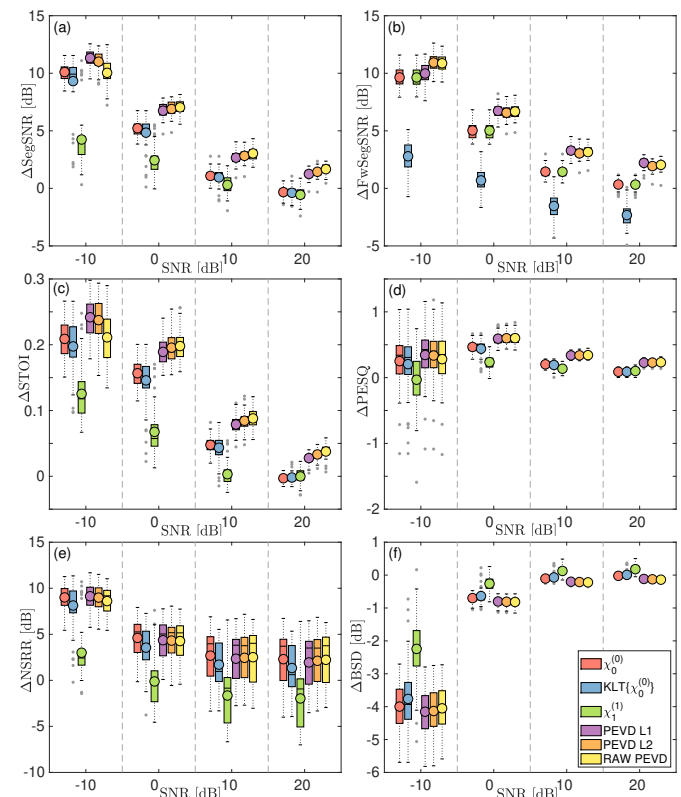Fig. 9: Speech enhancement results for white noise in a simulated room with $T_{60} = 0.3$ s.

Fig. 10: Speech enhancement results for babble noise in ACE Lecture Room 2 with $T_{60} = 1.22$ s.

TABLE III: Speech enhancement and source separation results for a single example using our approach.

(a) Speech enhancement of SMIRgen in 0 dB white noise.

Algorithm	ΔFwSegSNR	ΔSTOI	ΔPESQ	ΔNSRR	ΔBSD
$\chi_0^{(0)}$	4.86 dB	0.055	0.42	7.69 dB	-1.53 dB
KLT$\{\chi_0^{(0)}\}$	5.56 dB	0.054	**0.51**	**10.8 dB**	-1.65 dB
$\chi_1^{(1)}$	0.89 dB	0.122	0.44	1.08 dB	-0.65 dB
PEVD L1	5.72 dB	0.110	0.47	7.98 dB	-1.68 dB
PEVD L2	**5.92 dB**	**0.125**	**0.51**	7.78 dB	**-1.71 dB**
RAW PEVD	5.59 dB	0.119	0.49	8.13 dB	-1.62 dB

(b) Source separation of 1 female speaker for 2 female speakers in an anechoic scenario.

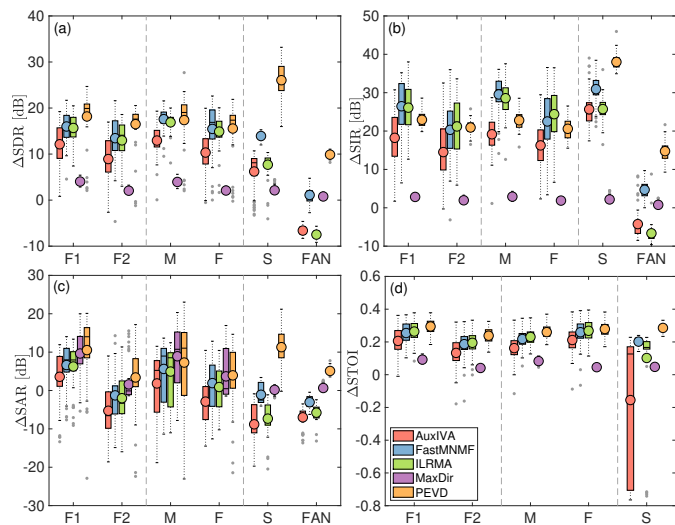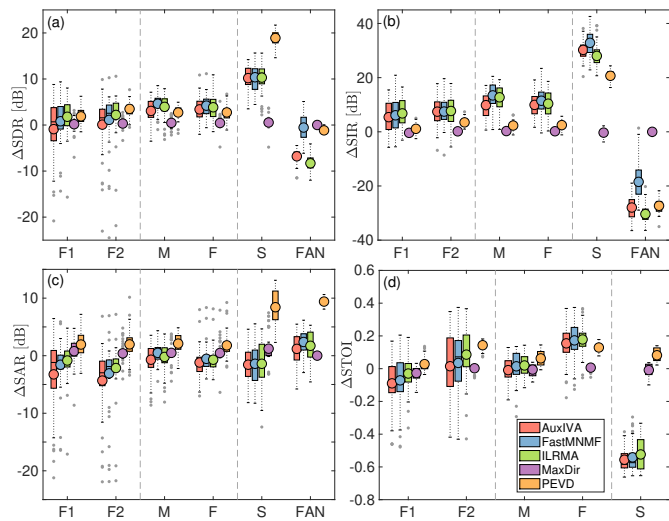| Algorithm	ΔSDR	ΔSIR	ΔSAR	ΔSTOI	ΔPESQ
AuxIVA	17.7 dB	25.3 dB	11.4 dB	0.21	1.05
FastMNMF	20.6 dB	**35.2 dB**	13.8 dB	0.21	1.28
ILRMA	19.5 dB	31.3 dB	12.8 dB	0.21	1.21
MHCARD	16.9 dB	17.8 dB	13.4 dB	0.21	0.93
PEVD	**21.8 dB**	25.3 dB	**16.4 dB**	**0.24**	**1.39**

Fig. 11: Source separation results in anechoic room.

Fig. 12: Source separation results in room with $T_{60} = 0.7$ s.

4) Separation of Multiple Sources: Table III(b) summarizes the results for a scenario involving the separation of 2 female speakers in an anechoic room. FastMNMF provides greatest ΔSIR of 35.2 dB and is closely followed by ILRMA, AuxIVA and PEVD. PEVD outperforms other algorithms in ΔSDR, ΔSAR, ΔSTOI and ΔPESQ by 21.75 dB, 16.38 dB, 0.237 and 1.39, respectively because it does not introduce processing artifacts. This is also observed in the listening

examples [57] which further indicate that non-target speech signals are attenuated but remain intelligible.

These findings are also observed in Fig. 11 for 50 trials involving source separation with two female speakers: F1 and F2, male and female speakers: M and F, a single speaker and localized fan noise: S and FAN using the setup in Fig. 2. When different source types are used in an anechoic room, the source separation of the PEVD is comparable with FastMNMF and ILRMA and is better than AuxIVA in the majority of cases.

Results for the room with $T_{60} = 0.7$ s in Fig. 12 indicates that the PEVD is comparable to ICA-based and MNMF-based methods in ΔSDR, worse in ΔSIR but best in ΔSAR and ΔSTOI in most cases. For the S and FAN scenario, PEVD performs better in SDR and SAR for both sources as it does not rely on source density functions which may not model fan noise well. Listening examples in [57] further support the observations that the PEVD does not introduce distortions because SMD is a time-domain method which preserves spectral coherence. We have previously demonstrated that it achieves a good compaction, resulting in good performance in the above applications.

VII. Conclusion

For a spherical microphone array, signal representations using SHT, KLT and PEVD are compared. While the PEVD achieves close to optimum signal compaction, it is computationally expensive. In contrast, the SHT achieves sub-optimal data compaction, but is data-invariant and scales well with the number of microphones. Therefore, using the PEVD to spatially decorrelate signals over a range of time shifts while managing computational complexity, we propose to combine the SHT and PEVD approaches. We first apply the SHT and obtain a number of eigenbeams that is smaller than the number of input signals. If DoA information is available, further computational complexity saving is expected with even fewer beamformed signals for PEVD processing.

The proposed framework for signal representation demonstrates that the diagonality factor improves on average by 7 dB over microphone signal representations. When exploiting the framework for speech enhancement and source separation, the method improves STOI and SDR by up to 0.2 and 20 dB, respectively. Informal listening examples also indicate that the method does not introduce any audible artifacts [57].

References

[1] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding.* Upper Saddle River, NJ, USA: Prentice Hall, 1995.

[2] P. P. Vaidyanathan, "Theory of optimal orthonormal subband coders," *IEEE Trans. Signal Process.*, vol. 46, no. 6, pp. 1528–1543, Jun. 1998.

[3] S. Akkarakaran and P. P. Vaidyanathan, "On optimization of filter banks with denoising applications," in *Proc. Int. Symp. on Circuits and Syst.*, Jul. 1999, pp. 512–515.

[4] S. J. Campanella and G. S. Robinson, "A comparison of orthogonal transformations for digital speech processing," *IEEE Trans. Commun. Technol.*, vol. 19, no. 6, pp. 1045–1050, Dec. 1971.

[5] N. S. Jayant and P. Noll, *Digital Coding of Waveforms Principles and Applications to Speech and Video*. Prentice Hall, 1984.

[6] M. K. Tsatsanis and G. B. Giannakis, "Principal component filter banks for optimal multiresolution analysis," *IEEE Trans. Signal Process.*, vol. 43, no. 8, pp. 1766–1777, Aug. 1995.

[7] S. Redif, J. G. McWhirter, and S. Weiss, "Design of FIR paraunitary filter banks for subband coding using a polynomial eigenvalue decomposition," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5253–5264, Nov. 2011.

[8] P. P. Vaidyanathan, *Multirate Systems and Filters Banks*. New Jersey, USA: Prentice Hall, 1993.

[9] V. W. Neo, S. Redif, J. G. McWhirter, J. Pestana, I. K. Proudler, S. Weiss, and P. A. Naylor, "Polynomial eigenvalue decomposition for multichannel broadband signal processing," *IEEE Signal Process. Mag.*, to be published.

[10] J. G. McWhirter, P. D. Baxter, T. Cooper, S. Redif, and J. Foster, "An EVD algorithm for para-hermitian polynomial matrices," *IEEE Trans. Signal Process.*, vol. 55, no. 5, pp. 2158–2169, May 2007.

[11] S. Redif, S. Weiss, and J. G. McWhirter, "Sequential matrix diagonalisation algorithms for polynomial EVD of parahermitian matrices," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 81–89, Jan. 2015.

[12] V. W. Neo and P. A. Naylor, "Second order sequential best rotation algorithm with Householder transformation for polynomial matrix eigenvalue decomposition," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2019, pp. 8043–8047.

[13] S. Weiss, I. K. Proudler, and F. K. Coutts, "Eigenvalue decomposition of a parahermitian matrix: Extraction of analytic eigenvalues," *IEEE Trans. Signal Process.*, vol. 69, pp. 722–737, 2021.

[14] S. Weiss, I. K. Proudler, F. K. Coutts, and F. A. Khattak, "Eigenvalue decomposition of a parathermitian matrix: Extraction of analytic eigenvectors," *IEEE Trans. Signal Process.*, vol. 71, pp. 1642–1656, April 2023.

[15] S. Redif, S. Weiss, and J. G. McWhirter, "Relevance of polynomial matrix decompositions to broadband blind signal separation," *Signal Process.*, vol. 134, pp. 76–86, May 2017.

[16] S. Weiss, N. J. Goddard, S. Somasundaram, I. K. Proudler, and P. A. Naylor, "Identification of broadband source-array responses from sensor second order statistics," in *Sensor Signal Process. for Defence Conf. (SSPD)*, 2017.

[17] S. Weiss, M. Alrmah, S. Lambotharan, J. G. McWhirter, and M. Kaveh, "Broadband angle of arrival estimation methods in a polynomial matrix decomposition framework," in *Proc. IEEE Int. Workshop on Computational Advances in Multi-Sensor Adaptive Process. (CAMSAP)*, 2013, pp. 109–112.

[18] S. Weiss, S. Bendoukha, A. Alzin, F. K. Coutts, I. K. Proudler, and J. Chambers, "MVDR broadband beamforming using polynomial matrix techniques," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2015, pp. 839–843.

[19] V. W. Neo, S. Weiss, and P. A. Naylor, "A polynomial subspace projection approach for the detection of weak voice activity," in *Sensor Signal Process. for Defence Conf. (SSPD)*, Sep. 2022, pp. 81–85.

[20] V. W. Neo, S. Weiss, S. W. McKnight, A. O. T. Hogg, and P. A. Naylor, "Polynomial eigenvalue decomposition-based target speaker voice activity detection in the presence of competing talkers," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, Sep. 2022, pp. 1–5.

[21] V. W. Neo, C. Evers, and P. A. Naylor, "Enhancement of noisy reverberant speech using polynomial matrix eigenvalue decomposition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3255–3266, Oct. 2021.

[22] F. K. Coutts, J. Corr, K. Thompson, S. Weiss, I. K. Proudler, and J. G. McWhirter, "Memory and complexity reduction in parahermitian matrix manipulations of PEVD algorithms," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2016, pp. 1633–1637.

[23] B. Rafaely, *Fundamentals of Spherical Array Processing*, ser. Springer Topics in Signal Processing. Springer, 2015.

[24] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*, ser. Springer Topics in Signal Processing. Springer, 2017.

[25] N. Epain and C. T. Jin, "Spherical harmonic signal covariance and sound field diffuseness," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 10, pp. 1796–1807, Oct. 2016.

[26] M. Park and B. Rafaely, "Sound-field analysis by plane-wave decomposition using spherical microphone array," *J. Acoust. Soc. Am.*, vol. 118, no. 5, pp. 3094–3103, Nov. 2005.

[27] T. Betlehem, W. Zhang, M. A. Poletti, and T. D. Abhayapala, "Personal sound zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 81–91, 2015.

[28] C. Evers and P. A. Naylor, "Acoustic SLAM," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 9, pp. 1484–1498, Sep. 2018.

[29] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 2, May 2002, pp. 1781–1784.

[30] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 2, 2002, pp. 1949–1952.

[31] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan. 2005.

[32] Y. Peled and B. Rafaely, "Linearly-constrained minimum-variance method for spherical microphone arrays based on plane-wave decomposition of the sound field," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 21, no. 12, pp. 2532–2540, Dec. 2013.

[33] S. Braun, D. P. Jarrett, J. Fischer, and E. A. P. Habets, "An informed spatial filter for dereverberation in the spherical harmonic domain," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2013, pp. 669–673.

[34] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2010, pp. 442–446.

[35] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "Eigenbeam-based acoustic source tracking in noisy reverberant environments," in *Proc. Asilomar Conf. on Signals, Syst. & Comput.*, Nov. 2010, pp. 576–580.

[36] A. H. Moore, C. Evers, and P. A. Naylor, "Direction of arrival estimation in the spherical harmonic domain using subspace psuedointensity vectors," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 2016. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2016.2613280

[37] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Springer-Verlag, 2010.

[38] V. W. Neo, C. Evers, and P. A. Naylor, "Polynomial matrix eigenvalue decomposition of spherical harmonics for speech enhancement," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Jun. 2021, pp. 786–790.

[39] V. W. Neo, C. Evers, and P. A. Naylor, "Polynomial matrix eigenvalue decomposition-based source separation using informed spherical microphone arrays," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, New Paltz, NY, Oct. 2021, pp. 201–205.

[40] H. Teutsch, "Wavefield decomposition using microphone arrays and its application to acoustic scene analysis," Ph.D. dissertation, Friedrich-Alexander University, Germany, 2005.

[41] J. Meyer and T. Agnello, "Spherical microphone array for spatial sound recording," in *Proc. Audio Eng. Soc. (AES) Conv.*, New York, NY, USA, Oct. 2003, pp. 1–9.

[42] J. Meyer and G. W. Elko, "Spherical microphone arrays for 3D sound recording," in *Audio Signal Processing For Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds. Kluwer Academic Publisher, 2004, pp. 67–89.

[43] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 240–254, May 2000.

[44] S. Weiss and I. Proudler, "Comparing efficient broadband beamforming architectures and their performance trade-offs," in *Proc. IEEE Int. Conf. Digital Signal Process. (DSP)*, A. N. Skodras and A. G. Constantinides, Eds., Jul. 2002, pp. 417–424.

[45] S. Weiss, J. Pestana, and I. K. Proudler, "On the existence and uniqueness of the eigenvalue decomposition of a parahermitian matrix," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2659–2672, May 2018.

[46] S. Icart and P. Comon, "Some properties of Laurent polynomial matrices," in *IMA Int. Conf. on Math. in Signal Process.*, Dec. 2012.

[47] P. Vaidyanathan, "Theory of optimal orthonormal subband coders," *IEEE Trans. Signal Process.*, vol. 46, no. 6, pp. 1528–1543, 1998.

[48] J. G. McWhirter and Z. Wang, "A novel insight to the SBR2 algorithm for diagonalising para-hermitian matrices," in *11th IMA Conference on Mathematics in Signal Processing*, Birmingham, UK, December 2016.

[49] F. K. Coutts, I. K. Proudler, and S. Weiss, "Efficient implementation of iterative polynomial matrix EVD algorithms exploiting structural redundancy and parallelisation," *IEEE Trans. Circuits Syst. I*, vol. 66, no. 12, pp. 4753–4766, Dec. 2019.

[50] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 6, pp. 697–707, September 2001.

[51] C. Hold, S. J. Schlecht, A. Politis, and V. Pulkki, "Spatial filter bank in the spherical harmonic: Reconstruction and application," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, New Paltz, NY, Oct. 2021.

[52] T. Deppisch, J. Ahrens, S. V. A. Garí, and P. Calamia, "Spatial subtraction of reflections from room impulse responses measured with a spherical microphone array," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, New Paltz, NY, Oct. 2021.

[53] C. Delaosa, J. Pestana, N. J. Goddard, S. Somasundaram, and S. Weiss, "Sample space-time covariance matrix estimation," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2019, pp. 8033–8037.

[54] C. Delaosa, J. Pestana, N. J. Goddard, S. Somasundaram, and S. Weiss, "Support estimation of a sample space-time covariance matrix," in *Sensor Signal Process. for Defence Conf. (SSPD)*, 2019.

[55] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1620–1643, Apr. 2020.

[56] A. O. T. Hogg, V. W. Neo, S. Weiss, C. Evers, and P. A. Naylor, "A polynomial eigenvalue decomposition MUSIC approach for broadband sound source localization," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, New Paltz, NY, Oct. 2021, pp. 326–330.

[57] V. W. Neo, "PEVD Exploiting Spherical Microphone Array Processing," Apr. 2021. [Online]. Available: https://vwn09.github.io/pevd-smap

[58] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium (LDC), Philadelphia, USA, Corpus LDC93S1, 1993.

[59] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: Algorithm and applications," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1462–1472, Sep. 2012.

[60] mh acoustics, "EM32 Eigenmike microphone array release notes (v17.0)," M. H. Acoust., NJ USA, Hardware, Oct. 2013. [Online]. Available: http://www.mhacoustics.com/sites/default/files/ReleaseNotes.pdf

[61] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 10, pp. 1681–1693, Oct. 2016.

[62] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[63] S. Brown and D. Sen, "Error analysis of spherical harmonic soundfield representations in terms of truncation and aliasing errors," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2013, pp. 360–364.

[64] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[65] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, Oct. 2011, pp. 189–192.

[66] S. Makino, *Audio Source Separation*, ser. Signals and Communication Technology. Springer-Verlag, 2018.

[67] K. Sekiguchi, A. A. Nugraha, Y. Bando, and K. Yoshii, "Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2019.

[68] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Apr. 2018, pp. 351–355.

[69] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2006, pp. 1447–1450.

[70] P. A. Naylor, N. D. Gaubitch, and E. A. P. Habets, "Signal-based performance evaluation of dereverberation algorithms," *J. of Elect. and Comput. Eng.*, vol. 2010, pp. 1–5, 2010.

[71] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[72] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Int. Telecommun. Union (ITU-T), Recommendation P.862, Nov. 2003.

[73] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[74] T. I. Laakso, V. Valimaki, M. Karjalainen, and U. K. Laine, "Splitting the unit delay [FIR/all pass filters design]," *IEEE Signal Process. Mag.*, vol. 13, no. 1, pp. 30–60, Jan. 1996.

[75] J. Selva, "An efficient structure for the design of variable fractional delay filters based on the windowing method," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3770–3775, Aug. 2008.

Vincent W. Neo (Member, IEEE) received the M.Eng. degree in 2014 and Ph.D. degree in 2022, both in electrical and electronic engineering from Imperial College London, UK, under the support of the Defence Science and Technology Agency (DSTA) Scholarship from Singapore. Dr Neo is currently a principal engineer in DSTA. Before his current role, he was a visiting postdoctoral researcher with Imperial College London, UK and has worked in several engineering roles in DSTA and Nuance Communications. His research focuses on multichannel signal processing and polynomial matrix decomposition with applications to speech, audio and acoustics.

Christine Evers (Senior Member, IEEE) is a lecturer in the School of Electronics and Computer Science at the University of Southampton. She was the recipient of an EPSRC Fellowship, hosted at Imperial College London, between 2017-2019. She worked as a research associate at Imperial College London between 2014-2017; as a senior systems engineer at Selex Electronic Systems between 2010-2014; and as a research fellow at the University of Edinburgh between 2009-2010. She received her PhD from the University of Edinburgh in 2010; her MSc degree in Signal Processing and Communications from the University of Edinburgh in 2006; and her BSc degree in Electrical Engineering and Computer Science from Jacobs University, Germany, in 2005. Her research focuses on Bayesian learning for machine listening. She is currently member of the IEEE SPS Technical Committee on Audio and Acoustic Signal Processing and serves as an associate editor for IEEE/ACM Transactions on Audio, Speech, and Language Processing, and the EURASIP Journal on Audio, Speech, and Music Processing.

Stephan Weiss (Senior Member, IEEE) received a Dipl.-Ing. degree from the University of Erlangen-Nürnberg, Erlangen, Germany, in 1995, and a Ph.D. degree from the University of Strathclyde, Glasgow, Scotland, in 1998, both in electronic and electrical engineering. He is professor for signal processing at the University of Strathclyde, Glasgow, following previous academic appointments at both the Universities of Strathclyde and Southampton. His research interests lie in adaptive, multirate, and array signal processing with applications in acoustics, communications, audio, and biomedical signal processing, where he has published more than 300 technical papers. Dr Weiss is a member of EURASIP and a senior member of the IEEE. He was the technical co-chair for EUSIPCO 2009 and general chair of IEEE ISPLC 2014, both organised in Glasgow, and special session co-chair for ICASSP 2019.

Patrick A. Naylor (Fellow, IEEE) received the B.Eng. degree in electronic and electrical engineering from the University of Sheffield, Sheffield, U.K., and the Ph.D. degree from Imperial College London, London, U.K. He is currently Professor of speech and acoustic signal processing with Imperial College London. His current research addresses microphone array signal processing, speaker diarization, and multichannel speech enhancement for applications including binaural hearing aids and augmented reality. In addition to his academic research, he enjoys several collaborative links with industry. He is currently a member of the Board of Governors of the IEEE Signal Processing Society and Past President of the European Association for Signal Processing.