Sparse Representations with Legendre Kernels for DOA Estimation and Acoustic Source Separation

Mert Burkay Çöteli, *Member, IEEE*, Hüseyin Hacıhabiboğlu, *Senior Member, IEEE*

***Abstract*—Recording multiple sound sources in a reverberant environment results in convolutive mixtures. Sound sources can be extracted from microphone array recordings of such mixtures using acoustic source separation techniques. Acoustic source separation using recordings obtained from rigid spherical microphone arrays (RSMA) benefit from the representation of sound fields as series of spherical harmonics. More specifically, RSMAs afford increased flexibility in acoustic beamforming and spatial filtering. We propose a data-driven DOA estimation and acoustic source separation method based on a dictionary-based sparse decomposition of sound fields. The proposed method involves identifying the time-frequency bins with contributions from a single source only and those with sensor noise or diffuse sound field components. The former set of bins is used in DOA estimation and beamforming in the sparse decomposition domain. The latter set is used to calculate the diffuse field covariance matrix used in Wiener post-filtering to improve the source separation performance further. We demonstrate the utility of the proposed method via extensive objective and subjective evaluations.**

***Index Terms*—rigid spherical microphone array, acoustic source separation, spherical harmonics, sparse recovery**

I. Introduction

Recordings of indoors acoustic scenes comprise the direct path components from sound sources as well as their reflections and late reverberation. Localization and separation of sound sources that are present in such scenes are two of the most important and computationally most demanding tasks for many acoustic applications such as speech recognition [1], robot audition [2] and smart home assistant devices [3].

Signals obtained using rigid spherical microphone arrays (RSMAs) can be processed to obtain order-limited spherical harmonic decompositions (SHD) of sound fields. SHD decouples directional and frequency-dependent components [4], and simplifies the associated microphone array processing algorithms. The development of practical RSMAs within the past two decades [5], [6] resulted in the proliferation of direction of arrival (DOA) estimation and beamforming algorithms.

Acoustic source separation from microphone array recordings typically involves two stages: DOA estimation and spatial filtering. These two components often rely on different signal models and processing frameworks, resulting in disjoint algorithms developed for DOA estimation or source separation alone.

The most straightforward DOA estimation algorithm for RSMAs involves steering a maximum directivity beam in all possible directions and finding the direction that maximizes the output power [7]. Subspace-based methods derived from classical spectrum estimation theory such as EB-MUSIC [8] and EB-ESPRIT [9], [10] provide high-resolution DOA estimation. Computationally less expensive DOA estimation approaches that rely on the energetic analysis of sound fields also exist. The earliest of these methods uses pseudo-intensity vectors (PIV), calculated from the first four spherical harmonic components of the sound field [11]. Related approaches such as subspace PIV (SS-PIV) [7] and augmented intensity vectors (AIV) [12] extend the idea that underlies the original PIV method. Another recent method, called hierarchical grid refinement (HiGRID), uses spatial entropy to adaptively refine a spherical search grid for identifying the local maxima of the steered response power density (SRPD) [13].

The accuracy of DOA estimation methods is typically reduced for acoustic scenes comprising multiple sources with overlapping spectra. Several methods were proposed to mitigate this problem. The direct path dominance (DPD) test uses the ratio of the two largest singular values of the spatial correlation matrix to identify the time-frequency bins where only a single source is active [14]. Another approach identifies single sound zones as contiguous regions in the time-frequency plane that predominantly contain a single dominant source [15]–[17]. Low-resolution HiGRID was used as a computationally effective source counting method together with EB-MUSIC [18]. The present authors have also proposed a sparsity-based approach to identify time-frequency bins with contributions substantially from a single source [19].

A commonly used acoustic source separation approach is spatial filtering via beamforming. Beamforming based approaches can be divided roughly into two categories: fixed [5] or signal dependent [20], [21]. The application of post-filtering in order to improve the performance of beamforming was also proposed [22]. However, such approaches typically do not perform well in highly reverberant environments.

More recently, sparse spectrotemporal or spatial representations of sound fields have been employed in acoustic source separation [23]. In one of the earliest applications of sparse recovery for sound field analysis in the spherical harmonic domain, dereverberation is achieved by estimating the direct sound and the early reflections followed by an approximate multichannel inversion of the room response [24]. Similarly,

Manuscript received June 4, 2021.

The work reported in this paper is supported by the Turkish Scientific and Technological Research Council (TÜBİTAK) Research Grant 119E254 "Audio Signal Processing for Six Degrees of Freedom (6DOF) Immersive Media". The authors are with METU Spatial Audio Research Group (SPARG), Middle East Technical University, Graduate School of Informatics (e-mail: mert.coteli@metu.edu.tr, hhuseyin@metu.edu.tr).

Digital Object Identifier 10.1109/TASL.2021.XYZXYZ

sparse recovery has been used in DOA estimation in conjunction with independent component analysis (ICA) [25] or by using dictionaries comprising spherical harmonic functions sampled at a discrete set of directions [26], [27]. It is also possible to jointly localize and separate sources using orthogonal dictionaries comprising spherical harmonic coefficients oriented in different directions [28], and the dictionaries employed for this purpose can be learned online [29]. Joint DOA estimation and dereverberation can also be achieved by sparse plane wave decomposition over a redundant dictionary of plane wave components defined in Cartesian coordinates [30].

The present authors proposed a dictionary-based approach operating in the time-frequency domain to obtain a sparse plane-wave decomposition of the sound field followed by spatial filtering over the dictionary weights for direction-informed source separation [31]. While this approach provides excellent source separation performance when the sources have similar direct-to-reverberant energy (D/R) ratios, increasing the D/R ratio differences reduces its performance. This degradation occurs due to the fixed beamwidth of the employed spatial filter that fails to capture the spread of the directional distribution that varies with the D/R ratio.

While some state-of-the-art methods perform well under anechoic conditions, their performance deteriorates with increasing reverberation and sensor noise, rendering their output mostly unusable in practical contexts, particularly in terms of interference suppression. Besides, the lack of a unified framework within which both DOA estimation, source separation and noise suppression can be carried out necessitates the use of algorithms that rely on essentially incompatible models, potentially increasing the computational cost.

In this paper, we unify the DOA estimation and source separation methods that we proposed in [31] and [19] under a single framework and extend this unified framework: 1) to make it adaptive to sources with different D/R ratios, and 2) to increase its robustness to reverberation and sensor noise. The approach we propose is based on a sparse decomposition of the steered response functional carried out in the time-frequency domain. DOA estimation relies on the identification of time-frequency bins that have a single dominant component. The distributions of the DOA estimates are then used to calculate and adapt the spatial filters used for directionally weighting the sparse plane wave decomposition of the sound field to separate sound sources from the mixture. The performance of the proposed approach is further improved by the application of Wiener post-filtering that relies on the identification of diffuse and noise-like time-frequency bins using the same sparse representation.

This article is organized as follows. Sec. II introduces the spherical harmonic decomposition framework used in the remainder of the article. The proposed DOA estimation and separation approaches are explained in Sec. III. Results of an objective evaluation of the method are presented in Sec. IV followed by Sec. V that presents a subjective evaluation, and by Sec. VI that presents a comparison with two other state-of-the-art methods. Sec. VII concludes the article.

II. Background

A. Spherical Harmonic Decomposition

An arbitrary function defined on a sphere can be represented as a weighted series of spherical harmonic functions, such that:

$$f(\Omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} f_{nm} Y_n^m(\Omega), \tag{1}$$

where $Y_n^m(\theta, \phi)$ is the spherical harmonic function of degree $n \in \mathbb{N}$ and order $m \in \mathbb{Z}$, in the direction $\Omega = (\theta, \phi)$ where $0 \leq \theta < \pi, 0 \leq \phi < 2\pi$ are the inclination and azimuth angles, respectively. The coefficients, f_{nm} can be calculated as:

$$f_{nm} = \int_S f(\Omega)[Y_n^m(\Omega)]^* d\Omega, \tag{2}$$

where $n \geq 0$ and $m \leq |n|$ and S is the unit sphere. The series expansion described above is called the spherical harmonic decomposition (SHD). The SHD coefficients of a unit amplitude plane wave, incident from the direction $\Omega_p = (\theta_p, \phi_p)$ in the acoustic free field are given as:

$$p_{nm} = 4\pi i^n j_n(kr)[Y_n^m(\Omega_p)]^* \tag{3}$$

where $j_n(\cdot)$ is the spherical Bessel function of the first kind. This representation naturally arises as a solution of the Helmholtz equation [32].

B. Rigid Spherical Microphone Arrays

A rigid spherical microphone array (RSMA) comprises a finite number, D, of pressure microphones positioned in the directions, $\{\Omega_d\}_{d=1\cdots D}$, on the surface of a sound-hard sphere of radius, r_a. The pressure signals captured using these microphones can be used to calculate the SHD coefficients of the recorded sound field. This is done by approximating the surface integral in (2) numerically such that:

$$p_{nm} = \sum_{d=1}^{D} w_d p_d(k)[Y_n^m(\Omega_d)]^* \tag{4}$$

where w_d are the quadrature weights that depend on the particular positions of microphones on the sphere, $p_d(k)$ is the pressure signal captured by the microphone, d, $k = 2\pi f/c$ is the wave number with f the frequency and c the speed of sound.

Note that sampling on the sphere results in spatial aliasing and the SHD obtained using RSMAs is therefore order-limited with the maximum order, $(N+1)^2 \leq M$. This means that only the first $N+1$ terms which are substantially alias-free would be useful in practice.

C. Plane Wave Decomposition (PWD)

An arbitrary sound field can be represented as a linear combination of plane waves in what is called the plane wave decomposition (PWD). PWD in the spherical harmonic domain allows utilizing the separation of frequency dependent and direction dependent components of sound fields.

The pressure around a rigid sphere of radius r_a due to a monochromatic plane wave propagating from the direction

$\Omega_q = (\theta_q, \phi_q)$ can be expressed in the spherical harmonic domain as [33]:

$$p_q(k, r, \Omega) = \alpha_q \sum_{n=0}^{\infty} \sum_{m=-n}^{m=n} p_{nm}^{(q)} Y_n^m(\Omega), \tag{5}$$

where $r \geq r_a$, $\alpha_q \in \mathbb{C}$ is the complex-valued amplitude of the plane wave, $p_{nm}^{(q)} = 4\pi i^n b_n(kr)[Y_n^m(\Omega_q)]^*$ are the SHD coefficients, and (r, Ω) is an observation point expressed in the spherical coordinates. Here $b_n(kr) = 4\pi i^n \left[j_n(kr) - j_n^{'}(kr_a)/h_n^{(2)'}(kr_a) h_n^{(2)}(kr) \right]$ represents the combined effect of the incident and scattered sound fields. After equalizing this frequency-dependent term, the sound field around the sphere can be expressed as a linear combination of multiple spherical harmonics evaluated at the directions of the individual plane waves, such that:

$$\tilde{p}_{nm} = \sum_{q=1}^{Q} \alpha_q [Y_n^m(\Omega_q)]^* \tag{6}$$

where $\{\alpha_q\}_{q=1\cdots Q}$ are the complex valued amplitudes of the plane waves that constitute the sound field.

In a reverberant environment with S sources, only $S \leq Q$ of the plane wave components will correspond to these sources and the remaining $Q - S$ plane waves will come from reflections and reverberation. DOA estimation aims to find $\{\Omega_q\}_{q=1\cdots S}$ and source separation aims to extract $\{\alpha_q\}_{q=1\cdots S}$ from the mixture. Notice that in the most general case, at a position close to a reflecting surface, there may be reflection components that may have higher power than the sources themselves, so the strongest components need not correspond to sound sources.

D. Steered Response Functional

An arbitrary shaped beam that can be steered symmetrically in three dimensions can be formed by combining spherical harmonic components [5]. A maximum directivity (MaxDF) beam can be obtained by weighting the equalized SHD coefficients with spherical harmonic functions evaluated at the steering direction Ω_s such that:

$$p_{\text{SRF}}(\Omega_s) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \tilde{p}_{nm} Y_n^m(\Omega_s) \tag{7}$$

$$= \sum_{q=1}^{Q} \alpha_q \sum_{n=0}^{\infty} \sum_{m=-n}^{n} [Y_n^m(\Omega_q)]^* Y_n^m(\Omega_s), \tag{8}$$

where the complex valued quantity, $p_{\text{SRF}}(\Omega_s)$, is called the *steered response functional* (SRF).

Truncating the series in (8) to its first $N+1$ terms and employing Christoffel's summation formula [34], SRF can be expressed as a linear combination of Q spatially bandlimited, real valued functions, $\Lambda_N(\Theta)$, such that:

$$p_{\text{SRF}}(\Omega_s) = \sum_{q=1}^{Q} \alpha_q \Lambda_N(\Omega_q | \Omega_s) \tag{9}$$

where $\Lambda_N(\Theta_{q,s}) \triangleq \Lambda_N(\Omega_q | \Omega_s)$, is called an N^{th}-order *Legendre kernel*, given as:

$$\Lambda_N(\Theta_{q,s}) = \sum_{n=0}^{N} \frac{2n+1}{4\pi} P_n(\cos \Theta_{q,s}). \tag{10}$$

Here, $P_n(\cdot)$ is the Legendre polynomial of order n and $\Theta_{q,s}$ is the angle between the direction of incidence (i.e. in the opposite direction to the DOA vector), Ω_q, and of the steering direction, Ω_s, respectively. Notice that $\Lambda_N(\Theta_{q,s})$ is rotationally symmetric with respect to the direction, Θ_s and that:

$$\lim_{N \to \infty} \Lambda_N(\Theta_{q,s}) = \frac{1}{2\pi} \delta(\cos \Theta_{q,s} - 1) \tag{11}$$

where $\delta(\cdot)$ is the Dirac delta function [33].

III. DOA Estimation and Source Separation using Sparse Representations

The method we propose in this paper uses the spatial distribution of source DOA estimates to adapt the characteristics of the spatial filter used to obtain a virtual beam. The method operates on the SHD coefficients obtained for each time-frequency bin of the individual channels of RSMA recordings and comprises three main components: i) a residual energy test to select bins that have a single source components for estimating source DOAs and bins that contain diffuse field and sensor noise for estimating the noise covariance matrix, ii) an adaptive spatial filtering stage based on the distributions of DOA estimates for source separation, and iii) Wiener post-filtering to improve the separation performance. All of these stages depend on a sparse plane wave decomposition of the recorded sound field carried out in the time-frequency domain. Fig. 1 shows the different stages of the proposed method.

A. Problem formulation

The signal model that we employ is based on the expression of the sound field at a given time-frequency bin (τ, κ) as a complex valued vector, $\mathbf{p}(\tau, \kappa) = [p_{\text{SRF}}(\tau, \kappa | \Omega_1), p_{\text{SRF}}(\tau, \kappa | \Omega_2), \cdots p_{\text{SRF}}(\tau, \kappa | \Omega_H)]^T$ comprising the values of the SRF at that time-frequency bin, sampled at H near-uniformly sampled directions, $\{\Omega_h\}_{h=1\cdots H}$, on the unit sphere. This vector can be expressed as a linear combination of R Legendre kernels, each representing a single plane wave, such that:

$$\mathbf{p}(\tau, \kappa) = \sum_{r=1}^{R} \alpha_r(\tau, \kappa) \mathbf{\Lambda}_r^{(N)} + \boldsymbol{\rho}(\tau, \kappa) \tag{12}$$

where $\{\alpha_r(\tau, \kappa)\}_{s=1\cdots R}$ are complex-valued weights that represent the amplitudes of each plane wave, and:

$$\mathbf{\Lambda}_r^{(N)} \triangleq [\Lambda_N(\Omega_r | \Omega_1) \ \Lambda_N(\Omega_r | \Omega_2) \ \cdots \Lambda_N(\Omega_r | \Omega_H)]^T \tag{13}$$

is an $H \times 1$ vector comprising the values of the Legendre kernel of order N localized at Ω_r sampled at the same directions as the SRF, and $\boldsymbol{\rho}(\tau, \kappa)$ is a residual term that contains the diffuse field components and distortions due to sensor noise given as:

$$\boldsymbol{\rho}(\tau, \kappa) = \sum_{q} \tilde{A}_q e^{-j\tilde{\chi}_q} \mathbf{\Lambda}_{\tilde{q}}^{(N)} + \mathcal{N}(\tau, \kappa) \tag{14}$$

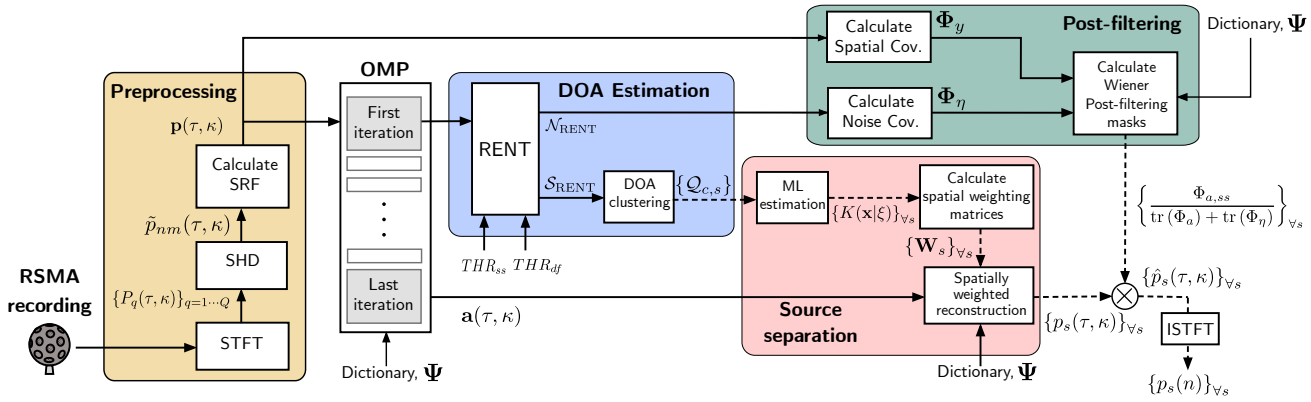
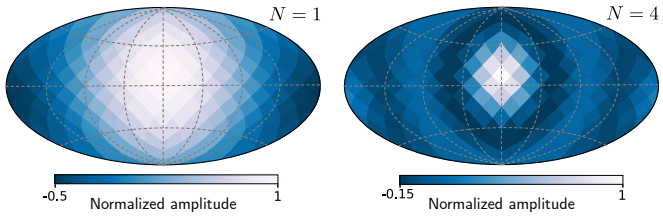


Fig. 1. Different stages of the proposed DOA estimation and source separation framework based on sparse plane wave decomposition.



(a) (b)

Fig. 2. Legendre kernels of orders (a) $N=1$ and (b) $N=4$ localized at $(\theta,\phi)=(\pi/2,0)$, calculated on a HEALPix grid with $H=192$ pixels and displayed using Mollweide projections.

where $\tilde{A}_q$ is a normally distributed amplitude term, $\tilde{\chi}_q$ is a phase term that is uniformly distributed in $[0,2\pi)$, $\tilde{\Omega}_q$ is uniformly distributed on the unit sphere, and $\mathcal{N}(\tau,\kappa)$ represents additive white Gaussian sensor noise.

Alternatively, S active sound sources can be separated from their respective reflections and reverberation and the latter would be lumped together into a compound term $\boldsymbol{\eta}(\tau,\kappa)$, such that:

$$\mathbf{p}(\tau,\kappa)=\sum_{s=1}^{S}\alpha_s(\tau,\kappa)\mathbf{\Lambda}_s^{(N)}+\boldsymbol{\eta}(\tau,\kappa) \tag{15}$$

This signal model can be expressed in matrix form as:

$$\mathbf{p}=\mathbf{L}\mathbf{a}(\tau,\kappa)+\boldsymbol{\eta}(\tau,\kappa) \tag{16}$$

where $\mathbf{L}\in\mathbb{R}^{H\times S}$ is a matrix comprising S Legendre kernels corresponding to sources in its columns, $\mathbf{a}\in\mathbb{C}^{S\times 1}$ is a vector comprising their amplitudes, and $\boldsymbol{\eta}\in\mathbb{C}^{H\times 1}$ is a noise vector. The models given in (12) and (15) are used for casting the source localization and separation problems into the same sparse framework.

B. Preprocessing Step

The preprocessing step of the proposed algorithm includes three stages. Firstly, an invertible time-frequency representation, $\{P_q(\tau,\kappa)\}_{q=1\cdots Q}$ of the signals from each individual microphone of the RSMA is computed. Secondly, the SHD for each time-frequency bin $\tilde{p}_{nm}(\tau,\kappa)$ where $0\leq n\leq N$ and $|m|\leq n$ is calculated using the appropriate quadrature formula for the employed RSMA. Finally, the SRF vector, $\mathbf{p}(\tau,\kappa)$ is calculated at near-uniformly sampled directions on the unit sphere. We use the short-time Fourier transform (STFT) as an invertible time-frequency representation, and the pixel centroids of a hierarchical equal area isolatitude pixelization (HEALPix) grid [35] for calculating the SRF vector.

C. Sparse Plane Wave Decomposition

The sparse plane wave decomposition that will be used in subsequent stages of the proposed method is obtained via orthogonal matching pursuit (OMP) [36], [37] by using a dictionary of Legendre kernels defined on the unit sphere. The fixed dictionary,

$$\mathbf{\Psi}=[\mathbf{\Lambda}_1^{(N)}\ \mathbf{\Lambda}_2^{(N)}\cdots\mathbf{\Lambda}_W^{(N)}], \tag{17}$$

used in OMP, comprises $H\times 1$ vectors $\{\mathbf{\Lambda}_w^{(N)}\}_{w=1\cdots W}$ in its columns as atoms that represent Legendre kernels sampled at H discrete directions on a near-uniform spherical grid with peaks corresponding to W different directions, $\{\Omega_w\}_{w=1\cdots W}$ covering the whole unit sphere where $W\geq H$. The maximum order, N, of the Legendre kernels depends on the maximum order afforded by the RSMA at a given frequency and will be different for different frequencies [33]. Fig. 2 shows Legendre kernels of order $N=1$ and $N=4$, respectively. The sharper spatial localization of the higher-order kernel is clearly visible.

The OMP algorithm comprises the iterative refinement of the sparse representation to find a good fit to the SRF vector, and the calculation of the residual error vector that is orthogonal to the approximation. More specifically, the residual error is initialized with the original SRF vector to be approximated (i.e. $\mathbf{r}_0(\tau,\kappa)=\mathbf{p}(\tau,\kappa)$) and each iteration involves the selection of the atom from the dictionary having the highest inner product with the residual. The residual vector for the next iteration is calculated as:

$$\mathbf{r}_{t+1}(\tau,\kappa)=\left[\mathbf{I}-\mathbf{L}_t\left(\mathbf{L}_t^T\mathbf{L}_t\right)^{-1}\mathbf{L}_t^T\right]\mathbf{p}(\tau,\kappa) \tag{18}$$

where $\mathbf{L}_t=[\mathbf{L}_{t-1}\,|\,\mathbf{\Lambda}_t]$ and $\mathbf{\Lambda}_t=\underset{\mathbf{\Lambda}_w\in\mathbf{\Psi}}{\text{argmax}}\,|\mathbf{\Lambda}_w^T\mathbf{r}_t|$.

The iterations can be stopped either after a prescribed number of repetitions or when the energy of the residual vector falls below a prescribed threshold. Upon the completion of the

iterations, a least-squares approximation to the complex valued $R \times 1$ complex amplitude vector, $\mathbf{a} = [\alpha_1,\ \alpha_2\ \cdots\ \alpha_R]^T$ is obtained using the Moore-Penrose pseudoinverse as:

$$\mathbf{a}(\tau,\kappa) = \left(\mathbf{L}^T\mathbf{L}\right)^{-1}\mathbf{L}^T\mathbf{p}(\tau,\kappa). \tag{19}$$

where $\mathbf{L} \triangleq \mathbf{L}(\tau,\kappa)$ comprises the R atoms $\{\mathbf{\Lambda}_r^{(N)}\}_{r=1\cdots R}$ used in the final sparse representation. For the case where the representation is obtained using a prescribed number of R iterations the representation uses R terms. For the case where a threshold is used as the stopping criterion, R depends on the choice of threshold. For the examples presented in this article, the iterations were stopped once the energy of the residual vector was more than 10 dB below the total energy. Note that $R \geq S$ in order for the final representation to accurately approximate the SRF vector.

D. Residual Energy Test (RENT)

Residual energy test (RENT) is used to select time-frequency bins that contain contributions substantially from a single source only as well as those that contain diffuse field or noise. RENT uses the output of the first iteration of the OMP algorithm to calculate the RENT measure:

$$\mathcal{R}(\tau,\kappa) = 1 - \frac{\|\mathbf{r}_1(\tau,\kappa)\|^2}{\|\mathbf{r}_0(\tau,\kappa)\|^2}, \tag{20}$$

where $0 < \mathcal{R}(\tau,k) \leq 1$. RENT allows classifying time frequency bins into three groups: bins with a single dominant source, bins with dominant diffuse field and noise, and bins that comprise one or more dominant sources, including their early reflections or diffuse field.

1) DOA Estimation: A time-frequency bin with a single plane wave can ideally be represented using a single atom resulting in $\mathcal{R}(\tau,k) \approx 1$. The time-frequency bins for which $\mathcal{R}(\tau,\kappa)$ is greater than a prescribed single source threshold, THR_{ss}, such that:

$$\mathcal{S}_{\text{RENT}} = \{(\tau,k) : \mathcal{R}(\tau,k) > THR_{ss}\} \tag{21}$$

are used in DOA estimation by registering the index, h, of the dictionary atom selected at the first iteration of OMP for each selected bin in $\mathcal{S}_{\text{RENT}}$. Since that atom is localized at a given direction, Ω_h, the DOA estimates are readily obtained for the corresponding bins.

The distribution of DOA estimates, $\mathcal{Q}_c$, is obtained by aggregating the individual DOA estimates in a given time-frequency region $\mathbf{\Upsilon}_a = \{(\tau,\kappa)|\tau \in [\tau_a - \tau_w/2, \tau_a + \tau_w/2],\ \kappa \in [\kappa_a - \kappa_w/2, \kappa_a + \kappa_w/2]\}$ where τ_a is the center and τ_w is the duration of the time interval, κ_a is the center and κ_w is the range of the frequency interval, respectively. Since RENT provides discrete DOA estimates, the resulting DOA histograms have bin resolutions defined by the resolution of the employed spherical grid.

The distribution of DOA estimates for the analysis region, $\mathbf{\Upsilon}_a$, is thresholded, binarized and then grouped into a set of clusters, $\{\mathcal{Q}_{c,s}\}_{s=1\cdots S}$, via connected components labeling (CCL) [38] such that $\mathcal{Q}_c = \bigcup_S \mathcal{Q}_{c,s}$. The source count, S and the DOAs $\{\Omega_s\}_{s=1\cdots S}$ for the given analysis interval are obtained from CCL as the number of clusters and their centroids

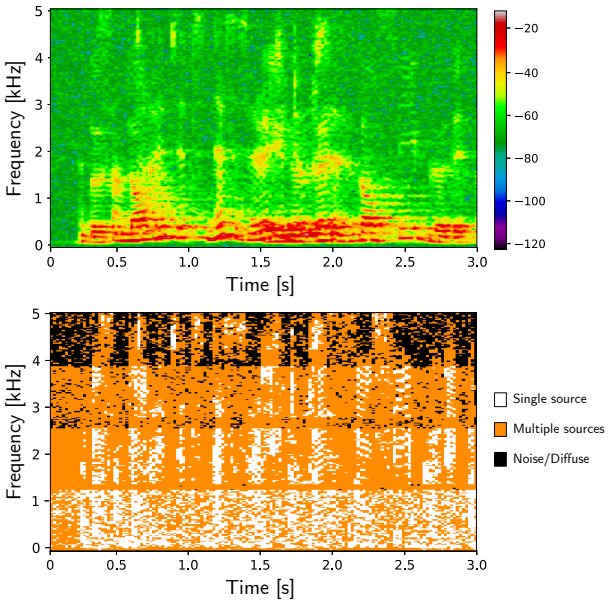


Fig. 3. Spectrogram of a two-source mixture with 20 dB SNR sensor noise (top) and time-frequency bins selected using RENT (bottom)

respectively. Note that for a reverberant sound field, the spread and orientation of the DOA distributions of individual sources depend substantially on the direct-to-reverberant (D/R) energy ratio, which covaries with the source distance, where a higher D/R ratio results in a lower spread of the DOA estimates, and vice versa.

2) Noise Covariance Matrix Estimation: The time-frequency bins without dominant sources are used for estimating the noise covariance matrix employed in Wiener post-filtering. In the ideal case, the energy of the SRF vector for a time-frequency bin that predominantly contains diffuse energy or noise is distributed uniformly in all directions. These time-frequency bins are selected as:

$$\mathcal{N}_{\text{RENT}} = \{(\tau,\kappa) : \mathcal{R}(\tau,\kappa) < THR_{df}\} \tag{22}$$

where for the ideally diffuse case $\mathcal{R}(\tau,\kappa) \approx 1/H$. These bins are considered to comprise substantially of $\boldsymbol{\eta}(\tau,\kappa)$.

The noise covariance matrix, $\mathbf{\Phi}_\eta = E[\boldsymbol{\eta}(\tau,\kappa)\boldsymbol{\eta}(\tau,\kappa)^H]$ can then be calculated from the identified time-frequency bins in the analysis interval as:

$$\mathbf{\Phi}_\eta \approx \frac{1}{\mathbf{card}(\mathcal{N}_{\text{RENT}})} \sum_{(\tau,\kappa)\in\mathcal{N}_{\text{RENT}}} \mathbf{p}(\tau,\kappa)\mathbf{p}(\tau,\kappa)^H \tag{23}$$

where $\mathbf{card}(\cdot)$ represents the cardinality of the set. The noise covariance matrix is updated at each analysis interval, $\mathbf{\Upsilon}_a$. Fig. 3 shows the spectrogram of a two-source mixture with additive sensor noise at 20 dB SNR alongside the single source and noise/diffuse time-frequency bins identified using RENT.

E. Beamforming via Adaptive Spatial Filtering

It is possible to weight the sparse approximation of the SRF vector given in (19) using a spatial weighting function defined on the sphere such that:

$$\mathbf{a}_s(\tau,\kappa) = \mathbf{W}_s(\tau,\kappa)^T\mathbf{a}(\tau,\kappa) \tag{24}$$

where $\mathbf{a}(\tau,\kappa)$ is the $R \times 1$ vector of complex amplitudes calculated in (19) and $\mathbf{W}_s(\tau,\kappa) \in \mathbb{R}^{R\times R}$ is a diagonal matrix comprising the weights to be applied on the amplitudes of the selected R atoms. It is possible to separate sources by applying an appropriate window function defined on the sphere to reconstruct a time-frequency representation for the selected source such that:

$$p_s(\tau,\kappa) = \mathbf{\Delta}_s^T \mathbf{L}(\tau,\kappa)\mathbf{a}_s(\tau,\kappa). \tag{25}$$

Here $\mathbf{\Delta}_s = [\delta_{1s}\ \delta_{2s}\cdots\delta_{Hs}]^T$ is an $H \times 1$ direction selector vector where δ_{ij} is the Kronecker delta function, $\mathbf{L} \in \mathbb{R}^{H\times R}$ contains the selected atoms, and s is the index of the grid position closest to the direction of the source of interest. The spatially filtered time-frequency representations are then inverted to obtain the separated signals.

A simple isotropic window that can be used for spatial filtering is the von Mises function defined as [39]:

$$w(\Omega_i|\Omega_j) = \frac{e^{\chi\cos\Omega_\Delta}}{2\pi I_0(\xi)} \tag{26}$$

where χ is the concentration parameter, Ω_Δ is the angular separation between the directions Ω_i and Ω_j and $I_0(\cdot)$ is the modified Bessel function of order 0. Prior to populating the weight matrix, $\mathbf{W}_s$, the calculated values are normalized. Note that, if $\chi = 0$, $\mathbf{W}_s$ would be an identity matrix and the resulting sound would be equivalent to the output of a maximally directive beamformer. The non-adaptive version of the algorithm which uses the von Mises function will henceforth be called as *sparse plane wave decomposition with fixed spatial filtering* (SPWD-FSF).

Since the spatial weighting operates on the sparse representation of the sound field, it bypasses the order limitations that would be encountered in acoustic mode beamforming. However, the source separation performance depends on the selected spatial weighting function, especially in terms of interference and reverberation [31]. When a narrow window function is used, sources at a closer distance would be separated well, but the distant sources have would have distortion and artefacts. In contrast, when a wide window is used, source separation performance diminishes for closer sources. This necessitates a spatial weighting function that adapts to individual sources.

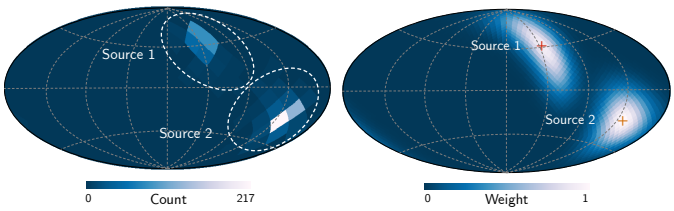
The second version of the algorithm is based on the selection of an anisotropic window on the sphere that matches the distribution of the DOA estimates pertaining to a given source. This approach is based on the observation that the DOA distributions obtained with RENT are not only different for different individual source D/R levels, but are also rotationally asymmetric. The distributions on the unit sphere without rotational symmetry can be modelled using the Kent distribution [40]. Defining an arbitrary point, $\mathbf{x}$, on the unit sphere, the probability density function of the Kent distribution:

$$K(\mathbf{x}|\boldsymbol{\xi}) = \frac{1}{c(\kappa,\beta)}\exp\{\chi\gamma_1^T\mathbf{x} + \beta[(\gamma_2^T\mathbf{x})^2 - (\gamma_3^T\mathbf{x})^2]\}, \tag{27}$$

is parameterized by $\boldsymbol{\xi} = \{\chi,\beta,\gamma_1,\gamma_2,\gamma_3\}$ where the χ represents the spread, β represents the ellipticity of the distribution. γ_1 represents the central tendency, γ_2 and γ_3 represent the principal axes of the distribution, respectively. Here $c(\chi,\beta)$ is a normalizing term given as:

$$c(\kappa,\beta) = 2\pi\sum_{j=0}^{\infty}\frac{\Gamma(j+\frac{1}{2})}{\Gamma(j+1)}\beta^{2j}(\frac{1}{2}\kappa)^{-2j-\frac{1}{2}}I_{2j+\frac{1}{2}}(\kappa), \tag{28}$$

where $I_s(\kappa)$ denotes the Bessel function of the first kind and $\Gamma(.)$ is the gamma function.



(a) (b)

Fig. 4. a) DOA distribution for two acoustic sources from the directions $(50°, 314°)$ and $(112°, 224°)$ obtained using RENT, b) Kent distributions fitted to data.

For each analysis interval, $\mathbf{\Upsilon}_a$, the DOA estimates $\{\Omega_{c,q}\}$ in each cluster representing a single source s, are converted to unit vectors on the sphere and a Kent distribution, $K_{c,s}(\mathbf{x}|\xi_s)$ is fitted to this data using maximum likelihood estimation (MLE) as described in [40] and implemented in [41]. This distribution is then normalized and sampled on the employed spherical grid to obtain $\mathbf{W}_{c,s} = \text{diag}\{K_{c,s}(\mathbf{x}_s|\boldsymbol{\xi}_s)\}_{s=1\cdots S}$ where $\mathbf{x}_s = [\cos\phi_s\sin\theta_s\ \ \sin\phi_s\sin\theta_s\ \ \cos\theta_s]^T$. Fig. 4 shows DOA distributions and the spatial weighting functions calculated for two sources. The anisotropic distributions of source DOA estimates and the fitted Kent distributions are clearly visible. The version of the proposed method that uses adaptive spatial filtering will be called as *sparse plane wave decomposition with adaptive spatial filtering* (SPWD-ASF).

Notice that while the OMP algorithm is applied on all time-frequency bins, the spatial weighting matrix $\mathbf{W}_{c,s}$ is updated only once for the corresponding analysis interval, Υ_c. If a previously identified source is not detected within a given time interval, the existing Kent distribution for the respective source is used for spatial filtering in order to maintain continuity across analysis intervals.

F. Wiener Post-filtering

Wiener post-filtering is applied as a final stage to reduce the effects of noise and reverberation. This stage involves the calculation of a time-frequency mask for each separated source based on an estimate of the power spectral density of the corresponding source. Assuming that $\boldsymbol{\eta}$ and $\mathbf{a}$ are uncorrelated, the spatial covariance matrix of SRF can be expressed as:

$$\begin{aligned}\mathbf{\Phi}_y(\tau,\kappa) &= E[\mathbf{p}\mathbf{p}^H] = E[(\widehat{\mathbf{L}}\mathbf{a}+\boldsymbol{\eta})(\widehat{\mathbf{L}}\mathbf{a}+\boldsymbol{\eta})^H] \\ &= \widehat{\mathbf{L}}E[\mathbf{a}\mathbf{a}^H]\widehat{\mathbf{L}}^T + E[\boldsymbol{\eta}\boldsymbol{\eta}^H] \\ &= \widehat{\mathbf{L}}\mathbf{\Phi}_a\widehat{\mathbf{L}}^H + \mathbf{\Phi}_{\boldsymbol{\eta}}\end{aligned} \tag{29}$$

where $\mathbf{\Phi}_\mathbf{a} \triangleq \mathbf{\Phi}_\mathbf{a}(\tau,\kappa)$ is the source covariance matrix which contains the source power spectral densities at the given time-frequency bin, $\widehat{\mathbf{L}} \in \mathbb{R}^{H\times S}$ is the matrix containing the dictionary atoms associated with each of the identified sources,

and $\mathbf{\Phi}_{\boldsymbol{\eta}}$ is estimated using (23). The least squares solution for the signal covariance matrix can be obtained using:

$$\mathbf{\Phi}_a \approx \widehat{\mathbf{L}}^{+}(\mathbf{\Phi}_y - \mathbf{\Phi}_{\boldsymbol{\eta}})(\widehat{\mathbf{L}}^{+})^H \tag{30}$$

where $\widehat{\mathbf{L}}^{+} = (\widehat{\mathbf{L}}^H\widehat{\mathbf{L}})^{-1}\widehat{\mathbf{L}}^H$ and spatial covariance matrix is obtained by averaging the output time-frequency bins over a time-frequency region centered at (τ, κ).

The estimates of power spectral densities corresponding to each source can be obtained directly from the terms on the diagonal of the signal covariance matrix. Wiener post-filtering applied on the source, s, involves the application of the following time-frequency mask:

$$\hat{p}_s(\tau,\kappa) = p_s(\tau,\kappa)\left[\frac{\mathbf{\Phi}_{a,ss}}{\text{tr}\,(\mathbf{\Phi}_a) + \text{tr}\,(\mathbf{\Phi}_{\boldsymbol{\eta}})}\right]^{\nu} \tag{31}$$

where $\mathbf{\Phi}_{a,ss}$ is the s-th element on the diagonal of the source covariance matrix, and $\text{tr}(\cdot)$ is the matrix trace operator and the exponent $\nu \geq 0$ determines the strength of the post-filtering operation.

IV. Objective Evaluation

We compared different versions of the proposed method with maximum directivity factor (MaxDF) beamforming which we employed as a baseline. We present two sets of evaluations: The first set of evaluations uses several different emulations to assess the effects of the complexity of the acoustic scene under noise-free conditions. The robustness to sensor noise is assessed in the second set of evaluations.

A. Acoustic scene emulations

The microphone array signals used in the objective evaluation were obtained by convolving anechoic speech signals with randomly selected acoustic impulse responses from the METU SPARG Eigenmike em32 Acoustic Impulse Response Dataset [42]. The dataset comprises 240 multichannel acoustic impulse responses (AIR) measured in a highly reverberant, empty classroom ($T_{60} \approx 1.12$ s) using an Eigenmike em32 microphone array.

The employed test signals are anechoic male and female speech signals from the European Broadcasting Union Sound Quality Assessment Material (EBU SQAM) Audio CD [43]. We used two-, three- and four-source mixtures obtained by summing the signals convolved with the AIRs for the noise-free scenarios and a two-source mixture only for assessing the robustness to noise. For the second set of evaluations, additive white Gaussian noise was added to each emulated microphone signal at 0, 10, 20 and 30 dB SNR.

For each of the cases to be evaluated 10 scenarios were randomly generated, resulting in 540 separated sources for the noise-free case and 480 separated sources for the noisy case. This way, a variety of different scenarios with different source DOAs and D/R ratios were obtained. Maximum angular separation between sources was $\pi/4$. All of the tested signals were 4 s long.

B. Compared methods

The methods compared in the evaluations are maximum directivity factor (MaxDF) beamforming, sparse plane wave decomposition with fixed spatial filtering (SPWD-FSF), sparse plane wave decomposition with adaptive spatial filtering (SPWD-ASF) as proposed in this article. SPWD-FSF uses a fixed von Mises function with a concentration parameter of $\chi = 10$. SPWD-ASF uses the normalized Kent distribution fitted to DOA data. Two different versions, with and without Wiener post-filtering (WPF) are evaluated. In all cases DOA estimation was carried out using RENT as described with $THR_{ss} = 0.5$ and $THR_{df} = 0.1$[1].

A 2048-sample Hann window was used with 50% overlap in the STFT. At each time-frequency bin OMP is iterated until the residual energy is at most 10 dB below the total energy. The employed spherical grid was a HEALPix grid [35] with $W = H = 192$ pixels. The peak directions of the dictionary atoms used for comparing different methods was calculated at the pixel centroids and each atom was sampled on the same grid. The frequency range used for RENT was $0 - 9$ kHz. Source separation via sparse plane wave decomposition uses the full signal bandwidth at the sampling rate of 48 kHz. The source DOAs were estimated every 250 ms over intervals of length $\tau_w = 500$ ms with 50% overlap. The reference signals for the computation of objective metrics were obtained by convolving only the direct path components of the AIRs with the anechoic speech signals and calculating an omnidirectional response from the lowest-order spherical harmonic components. For the computational efficiency, Wiener-post filtering is only applied in the valid time-frequency bins identified by RENT and with the exponent, $\nu = 1$.

C. Objective metrics

We used four commonly used objective metrics to evaluate the source separation performance. These are signal-to-interference ratio (SIR), signal-to-artifact ratio (SAR), signal-to-distortion ratio (SDR), and perceptual evaluation of speech quality (PESQ) [45] [46].

After these metrics were calculated, it was observed that PESQ was positively correlated with SDR and SAR. For the noise-free scenarios, the correlations of PESQ with SAR [$r(538) = 0.674, p < 0.001$] and with SDR [$r(538) = 0.617, p < 0.001$] were statistically significant. Similarly for the noisy scenarios, the correlations of PESQ with SDR [$r(478) = 0.649, p < 0.001$] and with SAR [$r(478) = 0.664, p < 0.001$] were statistically significant. In other words, PESQ scores increased with increasing SDR and SAR. Therefore, while the descriptive statistics for all metrics are presented, more detailed statistical analyses are provided only for SIR and PESQ. Readers who are not interested in the specific details of the statistical analyses presented in the following sections can skip to Sec. IV-F where the obtained results are summarized.

[1]Note that the two thresholds we used in the evaluations were selected heuristically. The RENT measure very likely depends on the D/R ratio in a similar way to thresholds used in the DPD test [44]. While it is possible that a more principled selection can be made also for the RENT thresholds, such optimization is left for future work.

D. Noise-free conditions

The convolution of anechoic sound signals with AIRs result in noise-free emulated recordings. Therefore, the corresponding evaluations provide useful information about the performance of the compared methods under ideal conditions.

The average DOA estimation error was $5.76^\circ \pm 3.60^\circ$ which is less than half of the grid resolution (i.e. 14.7°). Multiple comparisons revealed that DOA error did not depend on the number of sources in the scenario. This level of accuracy has proven to be sufficient for the source separation task that relies on it. Notice that using a higher resolution grid would have resulted in more accurate DOA estimations. However, this would have proportionally increased the computational cost.

Fig. 5 summarizes the SIR, SAR, SDR and PESQ results for the noise-free case for scenarios comprising different numbers of sources. The highest average SIR was consistently achieved by SPWD-ASF with Wiener post-filtering. SPWD-ASF with Wiener post-filtering provided the mean SIRs of 22.75, 18.63, and 15.69 dB for 2, 3, and 4 sources, respectively. SPWD-FSF with Wiener post-filtering provided the highest mean SDRs of 3.42, 1.26 and 0.54 dB for 2, 3, and 4 sources, respectively. SPWD-FSF without Wiener post-filtering provided the highest mean SARs of 3.97, 2.11, and 1.37 dB for 2, 3, and 4 sources, respectively. Finally, SPWD-FSF without Wiener post-filtering provided the highest mean PESQ scores of 2.21, 1.99, and 1.91 for 2, 3, and 4 sources, respectively.

While the scenario based objective evaluation involved a random selection of the source positions, some of the acoustical properties associated with the emulation scenarios had statistically significant correlations with the objective metrics at $p < 0.05$ or lower. More specifically, the objective metrics improve with increasing angular separation with the nearest interference and increasing D/R ratio at the source position, but worsen with increasing DOA estimation error. All of these covariates were included in the statistical models employed.

Detailed statistical assessments of SIR and PESQ via analyses of covariance (ANCOVA) are presented next. In the employed statistical models, the fixed factors were *Separation method* (METHOD), *Wiener post-filtering* (WPF), and *Source count* (COUNT). The statistical models we employed included all of the main terms, the two-way interactions of the independent variables and the identified covariates described above. Notice that we have included the Wiener post-filtering as a factor to investigate its effects separately.

1) SIR results: The ANCOVA model used with the SIR as the dependent variable revealed that METHOD [$F(2, 523) = 285.019,\ p < 0.001,\ \eta^2 = 0.320$], WPF [$F(1, 523) = 190.329,\ p < 0.001,\ \eta^2 = 0.107$] and COUNT [$F(2, 523) = 47.232,\ p < 0.001,\ \eta^2 = 0.053$] were all statistically significant. Among the two-way interactions METHOD$\times$WPF [$F(2, 523) = 32.199,\ p < 0.001,\ \eta^2 = 0.036$] and WPF$\times$COUNT [$F(2, 523) = 4.651,\ p = 0.01,\ \eta^2 = 0.005$] were statistically significant, suggesting that different methods benefit differently from Wiener post-filtering and that the benefits of Wiener post-filtering also depend on the number of sources.

Post-hoc comparisons with Tukey correction revealed that pairwise differences between all of the tested methods are statistically significant at $p < 0.001$ level. The highest SIR performance was provided by SPWD-ASF, followed by SPWD-FSF and MaxDF beamforming. SPWD-ASF provided 9.58 dB and 3.88 dB mean SIR improvements over MaxDF and SPWD-FSF, respectively. Similarly, SPWD-FSF provided 5.70 dB mean SIR improvement over MaxDF.

The difference between the mean SIR performances with and without post-filtering was also statistically significant at $p < 0.001$ level. Wiener post-filtering provided an additional 4.55 dB SIR improvement on average.

The pairwise comparisons revealed that the decrease in SIR was on average 1.78 dB between two- and three-source scenarios, and 2.48 dB between three- and four-source scenarios.

The post-hoc comparisons also revealed that the effect of the post-filtering varied across different methods. The additional SIR improvement that can be achieved by applying post-filtering to the output of the MaxDF beamformer resulted in an additional 7.84 dB SIR improvement, whereas SIR improvements for SPWD-FSF and SPWD-ASF were 4.16 dB and 1.65 dB, respectively.

2) PESQ results: The ANCOVA model with the PESQ as the dependent variable revealed that METHOD [$F(2, 523) = 93.504,\ p < 0.001,\ \eta^2 = 0.121$], WPF [$F(1, 523) = 96.663,\ p < 0.001,\ \eta^2 = 0.063$] and COUNT [$F(2, 523) = 38.867,\ p < 0.001,\ \eta^2 = 0.050$] were statistically significant. Among the two-way interactions, only METHOD$\times$WPF [$F(2, 523) = 25.160,\ p < 0.001,\ \eta^2 = 0.033$] was statistically significant.

Post-hoc comparisons with Tukey correction revealed that the difference between the mean PESQ scores obtained for MaxDF and SPWD-FSF methods is not statistically significant. SPWD-ASF performed worse than both MaxDF ($MD = -0.252$) and SPWD-FSF ($MD = -0.3$).

Applying Wiener post-filtering reduced the mean PESQ scores by 0.189 ($p < 0.001$). Increasing the source count also reduced the mean PESQ scores by 0.11 ($p < 0.001$) per each additional source.

While the application of Wiener post-filtering did not significantly affect the PESQ score for MaxDF, it did decrease the PESQ scores ($p < 0.001$) for SPWD-FSF and SPWD-ASF, by 0.238 and 0.321, respectively, in comparison with their versions not using Wiener post-filtering.

E. Robustness to Sensor Noise

The robustness of the different algorithms to additive sensor noise was evaluated by adding uncorrelated white Gaussian noise to each one of the emulated microphone signals at 0, 10, 20, and 30 dB SNR for 10 randomly generated two-source scenarios. Fig. 6 shows the spectrograms of the pressure signal due to noise-free mixture, the noisy mixture with 30 dB SNR, the reference signals and the signals separated with SPWD-ASF with Wiener post-filtering.

The average DOA estimation errors were $7.13^\circ \pm 4.81^\circ$, $5.49^\circ \pm 3.48^\circ$, $5.12^\circ \pm 3.49^\circ$ and $3.74^\circ \pm 3.17^\circ$, for 0, 10, 20, and 30 dB SNR, respectively. A better DOA estimation performance was achieved at higher SNR levels in general.

Fig. 7 summarizes the SIR, SAR, SDR and PESQ results for the different SNR levels. Averaged across all cases, the highest

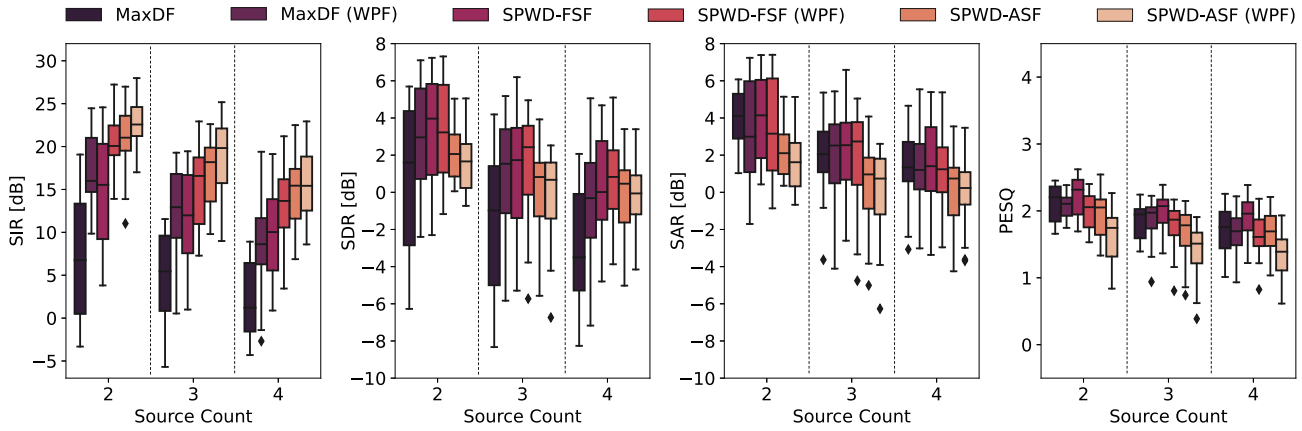


Fig. 5. SIR, SDR, SAR and PESQ scores for the tested algorithms for scenarios with different number of randomly selected source positions under ideal conditions with no sensor noise.

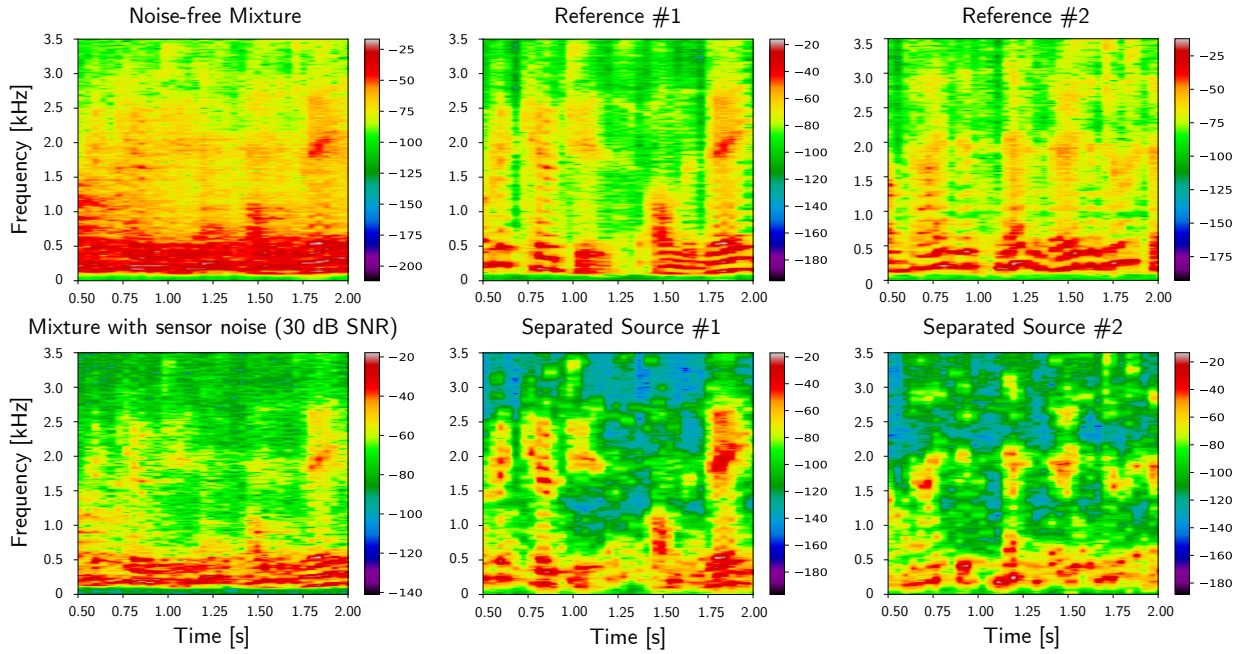


Fig. 6. Spectrograms of the pressure signals of the noise-free mixture, the mixture with additive sensor noise at 30 dB SNR, the two reference signals and the two separated sources obtained using SPWD-ASF with Wiener post-filtering.

mean SIR of 21.49 dB was achieved by SPWD-ASF with Wiener post-filtering. SPWD-ASF with Wiener post-filtering provided mean SIRs of 16.65 dB, 21.38 dB, 23.95 dB, and 24.00 dB for 0, 10, 20, and 30 dB SNR cases, respectively. The differences between the mean SDR performances of the tested methods were not statistically significant: SDR performance decreased similarly across different methods with increasing additive sensor noise. SPWD-FSF with Wiener post-filtering provided the highest mean SARs of -0.73 dB, 2.80 dB, and 3.40 dB for for 10, 20, and 30 dB SNR cases, respectively. However, SPWD-ASF with Wiener post-filtering provided a better performance with a mean SAR of -8.46 dB, only for the 0 dB SNR case. Finally, the highest mean PESQ score of 2.02 was achieved with SPWD-FSF without Wiener post-filtering. However, the differences between marginal means of PESQ scores across different methods were small.

The dependence of SIR and PESQ on different experimental parameters were assessed via analyses of covariance (ANCOVA). The fixed factors were *Separation method* (METHOD), *Wiener post-filtering* (WPF), and *Signal-to-noise ratio* (SNR). The covariates were the same as in the noise-free evaluation. The employed models include all of the main factors and two-way interactions as well as the covariates.

1) SIR results: The ANCOVA model used with the SIR as the dependent variable revealed that METHOD [$F(2,459) = 230.874$, $p < 0.001$, $\eta^2 = 0.266$], WPF [$F(1,459) = 318.381$, $p < 0.001$, $\eta^2 = 0.184$] and SNR [$F(3,459) = 22.190$, $p < 0.001$, $\eta^2 = 0.038$] were all statistically significant. Among the interactions, METHOD×WPF [$F(2,459) = 46.346$, $p < 0.001$, $\eta^2 = 0.053$], WPF×SNR [$F(3,459) = 5.615$, $p < 0.001$, $\eta^2 = 0.010$] and METHOD×SNR [$F(6,459) = 4.244$, $p < 0.001$, $\eta^2 = 0.015$] were all

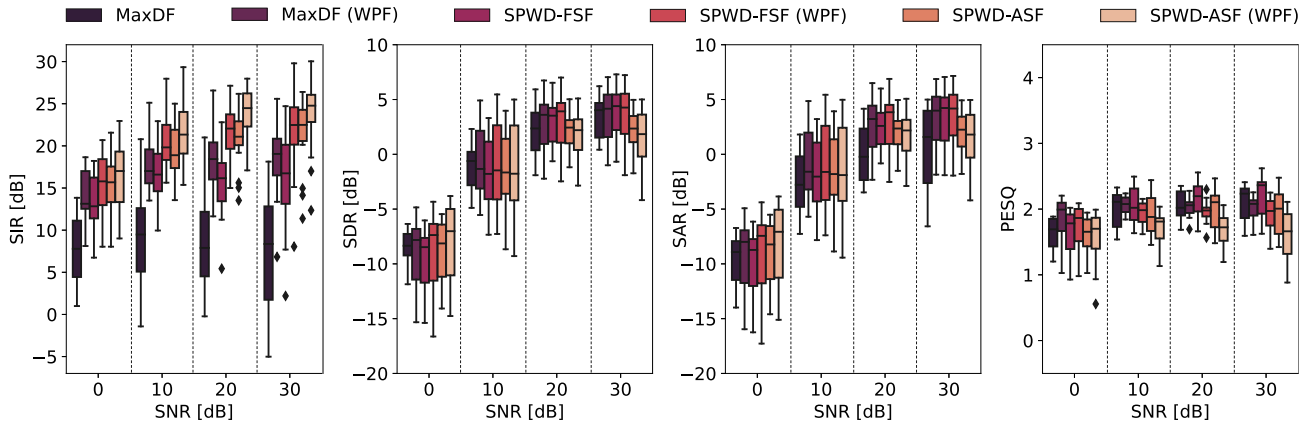


Fig. 7. SIR, SDR, SAR and PESQ scores for the tested algorithms for scenarios with comprising two randomly selected source positions with 0 dB, 10 dB, 20 dB, and 30 dB SNR, respectively.

statistically significant.

Post-hoc comparisons with Tukey correction revealed that all pairwise comparisons between different methods were statistically significant at $p < 0.001$ level. As with the noise-free case, SPWD-ASF provided the highest average SIR performance and achieved SIR improvements of 7.70 dB over MaxDF and 2.77 dB over SPWD-FSF. Wiener post-filtering improved the average SIR by 5.23 dB.

The mean SIR performance for the case with 0 dB SNR is significantly lower ($p < 0.001$) than those of all other SNR levels. The differences between the case with 0 dB SNR and 10, 20, and 30 dB SNR cases are -2.78, -3.13, and -3.26 dB, respectively. Different methods in the evaluation are affected differently: the differences between the mean SIR performance of MaxDF for different noise levels are not statistically significant. The SIR performances of SPWD-FSF and SPWD-ASF were different only for the 0 dB SNR case ($p < 0.001$), but not between other SNR levels. This indicates that the proposed approach is robust to noise in its SIR performance except at very high sensor noise levels.

The application of Wiener post-filtering significantly improved the SIR performances at $p < 0.001$ level. The highest SIR improvement with post-filtering is observed for MaxDF at 9.12 dB, followed by SPWD-FSF at 4.47 dB and SPWD-ASF at 2.27 dB, respectively. The mean SIR improvements as a result of post-filtering are statistically significant at $p < 0.001$ level for all noise levels. The mean SIR improvements achieved by post-filtering are 3.40, 5.06, 6.48, and 6.22 dB for 0, 10, 20, and 30 dB SNR, respectively.

2) PESQ results: The ANCOVA model used with the SIR as the dependent variable revealed that METHOD [$F(2, 459) = 85.503$, $p < 0.001$, $\eta^2 = 0.098$], WPF [$F(1, 459) = 35.429$, $p < 0.001$, $\eta^2 = 0.021$], SNR [$F(3, 459) = 49.881$, $p < 0.001$, $\eta^2 = 0.088$], METHOD×WPF [$F(2, 459) = 27.583$, $p < 0.001$, $\eta^2 = 0.033$], METHOD×SNR [$F(6, 459) = 2.156$, $p < 0.046$, $\eta^2 = 0.008$], WPF×SNR [$F(3, 459) = 27.583$, $p < 0.001$, $\eta^2 = 0.04$] were statistically significant.

Post-hoc comparisons with Tukey correction indicate that PESQ scores for SPWD-ASF were significantly lower than both MaxDF and SPWD-FSF with mean differences of 0.212 and 0.2, respectively. The differences in the mean PESQ scores between MaxDF and SPWD-FSF were not significant. The difference between mean PESQ scores for cases with and without Wiener post-filtering was -0.09 which was significant at $p < 0.001$ level, albeit not a meaningful difference. Similarly to the SIR, PESQ scores for the 0 dB SNR case were significantly lower than other cases at $p < 0.001$ level. No statistically significant differences exist between the other SNR levels.

Post-filtering does not significantly reduce the PESQ scores for MaxDF. While post-filtering significantly reduces PESQ scores ($p < 0.001$) for both SPWD-FSF and SPWD-ASF, the difference is only practically meaningful for SPWD-ASF where post-filtering reduces the mean PESQ score by 0.21.

Finally, post-filtering improves the mean PESQ scores at lower SNR levels while it impairs them for higher SNR levels. More specifically, while the difference in mean PESQ score of -0.109 statistically significant at $p = 0.008$ for 0 dB SNR, the difference is not statistically significant for 10 dB SNR, and is 0.172 and 0.214 for 20 and 30 dB SNR levels, respectively.

F. Discussion of results

The evaluation we presented shows that a higher source separation performance in terms of the SIR is achievable at the cost of reducing the PESQ scores that correlate with the perceived quality of the separated sources.

The evaluations confirmed our initial expectations that a higher D/R ratio at the source position, would result in better the objective metrics. Also, an increasing source count or decreasing SNR negatively impacts all objective metrics.

The highest SIR performance in the noise-free case was achieved using the SPWD-ASF method with Wiener post-filtering. The highest PESQ score was achieved using the SPWD-FSF without Wiener post-filtering. It should be noted that the additional SIR improvement that the post-filtering provides for SPWD-ASF in the noise-free case is not sufficiently high to justify the additional computational cost incurred. The results for the noise-free case are relevant in the context of demixing artificial higher-order Ambisonics (HOA) mixes.

The proposed source separation approach using sparse plane wave decomposition is robust to noise where it can provide mean SIR improvements in excess of 15 dB even in the considerably unrealistic case of 0 dB SNR. The proposed Wiener post-filtering approach also improves the SIR performance for beamforming based source separation for all SNR levels.

Note that the combination of a high level of reverberation and the additive sensor noise especially at 0 and 10 dB SNR, causes SAR and SDR to be very low. While these conditions act as a sanity check while specifying the performances the proposed methods, they also represent very unrealistic situations, especially since the nominal SNR even for low-cost MEMS microphones are much higher than 30 dB (see for example [47], [48]). In this sense, the more realistic scenario among those evaluated is the one that emulates 30 dB SNR. Based on the results from that case, the highest mean SIR performance is provided by SPWD-ASF with Wiener post-filtering, and the highest PESQ score was achieved using SPWD-FSF without post-filtering.

G. Computational aspects

The proposed approach is data-driven, meaning that the computational cost depends on the acoustic scene complexity and the acoustical properties of the space in which the recordings are made. However we have carried out a preliminary comparison of the computational costs of the methods that we evaluated. The comparison involved executing all of the compared algorithms on six acoustic scenes containing two speech sources each randomly generated using the AIR dataset used in the previous section. The comparison was carried out on a desktop computer with an Intel Core i7-7700K CPU at a clock frequency of 4.20 GHz, having 32 GB RAM, running Ubuntu Linux 20.04 LTS. The implementation of the algorithms was in Python 3.6. Table I contains the computational time per time-frequency bin for each of the tested methods. The presented computation times are averaged over a total of 73728 time-frequency bins. It may be observed that SPWD-ASF with Wiener post-filtering, that is the most complex version of the proposed algorithm, has a computational cost that is less than an order of magnitude higher than the baseline MaxDF algorithm. This result indicates the proposed method's suitability for real-time operation.

Note that our implementation is not optimized for computational efficiency in terms of the employed interpreted language, single threaded processing on a single core, and the use of standard matrix and scientific computing libraries. Therefore, the ratios of the total runtimes of each method to that of the baseline MaxDF approach as given in the last column of the table are more informative than the absolute runtime measurements. The minor localization runtime differences between methods are due to measurement noise.

TABLE I
AVERAGE COMPUTATION TIMES FOR A SINGLE BIN IN MILLISECONDS.

Method	Localization	MLE	Separation	Total
MaxDF	0.67	-	0.33	1.00
MaxDF (WPF)	0.65	-	1.01	1.66
SPWD-FSF	0.69	-	5.40	6.09
SPWD-FSF (WPF)	0.66	-	6.25	6.91
SPWD-ASF	0.67	0.93	6.47	8.07
SPWD-ASF (WPF)	0.64	0.95	7.59	9.18

V. SUBJECTIVE EVALUATION

Different versions of the proposed approach were compared with the MaxDF beamformer in a multiple stimulus hidden reference and anchor (MUSHRA) test [49]. Three acoustic scenes comprising two speech sources with random positions were emulated using the AIR dataset used in the previous sections. The reference in each case was obtained by convolving the windowed direct path component of the respective AIR with the anechoic signal. The anchor was obtained as the pressure component which represents the omnidirectional convolutive mixture without any directional filtering.

The experiment was run remotely using `webMUSHRA` [50]. Ten participants (9 male and 1 female; 23-40 years old) with normal hearing took part in the experiment. The participants were asked to rate the *source separation performance* of the compared methods on a scale from 0 (Bad) to 100 (Excellent). They were also requested to identify the anchor and the reference and rate them as 0 and 100, respectively.

The mean scores and standard errors for the anchor and the reference were 11.96 ± 14.99 and 99.87 ± 1.12, respectively[2]. This indicates that the participants were able to easily discriminate these extreme cases. The mean scores and standard errors for each tested method are, 41.11 ± 23.89 (MaxDF), 60.58 ± 20.09 (MaxDF with Wiener post-filtering), 54.49 ± 20.97 (SPWD-FSF), 65.63 ± 22.05 (SPWD-ASF), 66.13 ± 17.07 (SPWD-FSF with Wiener post-filtering) and 71.40 ± 22.37 (SPWD-ASF with Wiener post-filtering).

The results were analyzed via a one-way ANOVA where the response was the dependent variable and METHOD was the independent variable which was found to have a significant effect $[F(7, 632) = 140.16,\ p < 0.001,\ \eta^2 = 0.608]$. The results of post-hoc comparisons revealed statistically significant differences between the different methods (see Table II). These results indicate that MaxDF was scored significantly lower than all the other tested methods. Adding Wiener post-filtering to MaxDF alone improved its mean MUSHRA score, making it comparable with the mean scores for SPWD-FSF and SPWD-ASF. Adaptive spatial filtering is also seen to improve the mean scores in comparison with fixed spatial filtering. The highest mean score was obtained for SPWD-ASF with Wiener post-filtering which received higher mean scores than all of the other evaluated methods. However, the differences were not statistically significant between the mean scores given to SPWD-ASF with Wiener post-filtering and SPWD-ASF without Wiener post-filtering and SPWD-FSF without Wiener post-filtering. In summary, the results of the MUSHRA test reveal that the proposed modular approach in any of its profiles can provide a good level of subjective

[2]All participants identified and scored the anchor and the reference at the two opposite ends of the scale, giving the highest score to the reference and the lowest score to the anchor. However, some participants gave a rating of 20 points for the anchor and a single participant gave a rating of 90 points to the reference in a single trial. These results were included in the final analysis.

TABLE II
MEAN MUSHRA SCORE DIFFERENCES (ROW - COL). STATISTICAL SIGNIFICANCES ARE INDICATED FOR $p < 0.01$ (*) AND $p < 0.001$ (**).

MaxDF (WPF)	SPWD-FSF	SPWD-ASF	SPWD-FSF (WPF)	SPWD-ASF (WPF)	Anchor	Reference	
-19.46**	-13.38**	-24.51**	-25.01**	-30.28**	29.15**	-58.76**	**MaxDF**
	6.09	-5.05	-5.55	-10.82*	48.61**	-39.3**	**MaxDF (WPF)**
		-11.13*	-11.64*	-16.91**	42.52**	-45.38**	**SPWD-FSF**
			-0.5	-5.78	53.66**	-34.25**	**SPWD-ASF**
				-5.28	54.16**	-33.75**	**SPWD-FSF (WPF)**
					59.44**	-28.48**	**SPWD-ASF (WPF)**
						-87.91**	**Anchor**

separation performance.

VI. COMPARISON WITH THE STATE OF THE ART

SPWD-ASF with Wiener post-filtering was compared with two other methods from the literature. The first method was proposed by Kalkur *et al.* [28] and uses a sparse decomposition using a dictionary comprising complex-valued spherical harmonic decomposition coefficients representing plane waves. The second approach proposed by Fahim *et al.* [22] comprises MaxDF beamforming followed by Wiener post-filtering. While the former is a joint DOA estimation and source separation method, the latter requires source DOAs to be a priori known. The comparisons also included MaxDF as the baseline. Three types of evaluations were carried out: 1) using simulated plane waves emulating anechoic conditions, 2) using reverberant recordings obtained as described in the previous sections, and 3) using anechoic speech sources with added ambient noise to assess the robustness of the evaluated algorithms to non-stationary noise. Speech mixtures comprising two sources were used in all comparisons.

The first set of evaluations used five anechoic simulations. While source DOAs were accurately estimated using the method described in [28], RENT was used to estimate source DOAs for the method described in [22] and also for the baseline method. Three scenarios generated as in the previous section were tested for the reverberant case in the second set of evaluations. Since the method described in [28] did not provide accurate DOA estimates under reverberant conditions and with ambient noise, DOAs estimated with RENT were used instead. The third set of evaluations used five simulations with randomly selected source directions with anechoic emulations of speech sources and additive recorded ambient noise at 10 and 20 dB SNR levels. The employed ambient noise was recorded in a busy street using an Eigenmike em32 microphone [51]. In all of the simulations, the duration of the audio signals was 2 s and all the other parameters were selected as described in the previous sections.

Table III shows the means and the standard deviations of the objective metrics calculated for the anechoic and reverberant cases. It may be observed that all of the methods except MaxDF beamforming perform similarly well in the anechoic case. The highest mean SIR is achieved by the proposed method while the highest PESQ score is achieved by the method proposed in [22]. It is interesting to note that the baseline MaxDF approach has the highest mean SAR performance while its PESQ score is the lowest due to the low SIR. The mean SIRs of the two other methods in the comparison decrease noticeably under reverberation and become comparable to the baseline MaxDF beamforming, whereas the proposed method continues to provide the highest mean SIR. However, what the proposed method gains in SIR, it loses in the mean PESQ score, which is the lowest among the compared methods. This is due to the decrease in the SAR and SDR that the proposed method suffers under reverberation. Regardless of its lower performance in terms of the metrics related to quality in comparison with the other methods, the substantial improvement in SIR alone can justify using the proposed method in applications such as automatic speech recognition. The speech quality could likely be further improved by optimizing the model parameters which is left for future work.

Table IV shows the source separation metrics for different levels of ambient noise. It may be observed that the proposed approach provides a high SIR and PESQ score irrespective of the level of the ambient noise. For the lower ambient noise level with 20 dB SNR, SAR and SDR for the proposed method are slightly lower than the other methods in the comparison. However, for the higher ambient noise level with 10 dB SNR, the proposed approach outperforms all the other methods in the comparison. It may also be observed that the PESQ score for the proposed approach is improved substantially in comparison with the reverberant case. We speculate that the diffuse characteristics of the ambient noise might have had a positive effect on this outcome. Note that the ambient noise employed for this evaluation is non-reverberant, and it is possible that the performance of the proposed method could slightly decrease under reverberant ambient noise.

VII. CONCLUSIONS

A modular framework for DOA estimation and source separation using RSMAs was proposed in this article. The proposed approach relies on a dictionary-based sparse plane wave decomposition that uses real-valued dictionary atoms arising naturally with plane wave decomposition in the SHD domain. The proposed sparse decomposition is obtained via orthogonal matching pursuit over complex-valued steered responses calculated for each time-frequency bin. The same process is used also to identify the time-frequency bins with contributions substantially from a single plane wave as well as those bins that have no dominant components. The former set is used to estimate the source DOAs while the latter is used to calculate the Wiener post-filter. The distribution of the DOA estimates are used to calculate the parameters of a spatial weighting filter, later to be used for virtual beamforming

TABLE III
MEANS AND STANDARD DEVIATIONS OF SIR, SAR, SDR AND PESQ FOR THE COMPARED METHODS FOR ANECHOIC AND REVERBERANT CONDITIONS

	Anechoic conditions				Reverberant conditions			
Method	**SIR (dB)**	**SAR (dB)**	**SDR (dB)**	**PESQ**	**SIR (dB)**	**SAR (dB)**	**SDR (dB)**	**PESQ**
MaxDF	14.11 ± 2.37	8.89 ± 2.12	7.52 ± 0.56	2.33 ± 0.35	10.56 ± 3.52	3.42 ± 0.65	2.17 ± 1.30	2.20 ± 0.41
Kalkur *et al.* [28]	23.22 ± 5.82	6.21 ± 2.95	6.04 ± 3.07	2.44 ± 0.28	11.75 ± 4.97	-0.54 ± 1.94	-1.40 ± 1.49	2.22 ± 0.47
Fahim *et al.* [22]	26.46 ± 1.69	8.61 ± 0.20	8.52 ± 0.21	3.22 ± 0.25	13.36 ± 2.21	3.71 ± 0.66	3.04 ± 0.87	2.29 ± 0.43
SPWD-ASF (WPF)	27.42 ± 1.84	6.88 ± 0.12	6.83 ± 1.16	2.89 ± 0.35	22.77 ± 0.83	2.54 ± 1.03	2.48 ± 1.03	1.87 ± 0.41

TABLE IV
MEANS AND STANDARD DEVIATIONS OF SIR, SAR, SDR AND PESQ FOR THE COMPARED METHODS FOR ADDITIVE AMBIENT NOISE

	Anechoic source + Ambient Noise (20 dB SNR)				Anechoic source + Ambient Noise (10 dB SNR)			
Method	**SIR (dB)**	**SAR (dB)**	**SDR (dB)**	**PESQ**	**SIR (dB)**	**SAR (dB)**	**SDR (dB)**	**PESQ**
MaxDF	16.29 ± 2.89	9.09 ± 1.08	8.15 ± 1.20	2.03 ± 0.15	15.90 ± 3.26	7.82 ± 1.10	6.98 ± 1.41	1.69 ± 0.03
Kalkur *et al.* [28]	24.30 ± 5.59	7.60 ± 3.45	7.43 ± 3.50	2.22 ± 0.18	22.99 ± 7.86	7.43 ± 2.93	7.12 ± 3.23	1.92 ± 2.11
Fahim *et al.* [22]	27.88 ± 2.15	8.94 ± 1.25	8.88 ± 1.26	2.50 ± 0.12	26.97 ± 3.34	8.03 ± 1.17	7.95 ± 1.20	1.92 ± 0.08
SPWD-ASF (WPF)	29.36 ± 2.05	8.60 ± 1.75	8.56 ± 1.75	2.78 ± 0.11	29.52 ± 1.69	8.24 ± 1.67	8.20 ± 1.67	2.40 ± 0.08

over the sparse decomposition to separate sources. Finally the calculated Wiener post-filter is applied on the resulting time-frequency representations to reduce the effects of noise and diffuse field components.

Different versions of the proposed approach were evaluated objectively and subjectively. It was shown that the proposed methods are robust to a high level of reverberation and sensor noise. A comparison with two other source separation methods showed that the proposed approach achieves a higher suppression of interference, especially at a high reverberation[3].

Since the proposed approach is modular, it can be adapted for different problems with different constraints. For example, if the approach is to be used as a speech recognition front-end, the suppression of interference components would take precedence and the proposed algorithm would be used with an adaptive spatial filter and post-filtering. In contrast, if an object-based audio (OBA) reproduction scenario is considered, a lower level of distortion and artefacts and a higher quality would be desired, meaning that the proposed approach would be used with a fixed spatial filter and without post-filtering.

Although no special parameter optimization was carried out, superior results were obtained using the proposed algorithm. It is possible that even better objective and subjective results can be obtained with parameter optimization. Also, outputs from the different stages of the algorithm can be linearly or non-linearly combined to achieve a better trade-off between interference suppression and separation quality. For example, Wiener post-filtering can be applied on the output time-frequency bins selectively or outputs from different versions of the method could be linearly combined. Such simple modifications may reduce the relative level of undesirable musical noise that is the major audible artefact occurring due to the employed time-frequency processing approach.

REFERENCES

[1] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel mmse-based framework for speech source separation and noise reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 9, pp. 1913–1928, 2013.

[3]Examples of separated sound sources accompany this article as supplementary material.

[2] J.-M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *2004 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, vol. 3, 2004, pp. 2123–2128.

[3] H. M. Do, M. Pham, W. Sheng, D. Yang, and M. Liu, "Rish: A robot-integrated smart home for elderly care," *Robotics and Autonom. Syst.*, vol. 101, pp. 74–92, 2018.

[4] B. Rafaely, *Fundamentals of Spherical Array Processing*. Berlin, Heidelberg: Springer-Verlag, Oct. 2015, vol. 8.

[5] G. W. Elko and J. Meyer, "Spherical microphone arrays for 3D sound recordings," in *Audio Signal Processing for Next Generation Multimedia Communication Systems*, 2004.

[6] J. Zamojski, P. Makaruk, L. Januszkiewicz, and T. Zernicki, "Recording, mixing and mastering of audio using a single microphone array and audio source separation algorithms," in *143th Conv. of Audio Eng. Soc.*, Oct 2017.

[7] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *18th European Signal Process. Conf. (EUSIPCO 2010)*, 2010, pp. 442–446.

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

[9] H. Teutsch and W. Kellermann, "Detection and localization of multiple wideband acoustic sources based on wavefield decomposition using spherical apertures," *IEEE Int. Conf. on Acoust. Speech and Signal Process.*, pp. 5276—5279, Apr. 2008.

[10] A. Herzog and E. A. Habets, "Eigenbeam-ESPRIT for DOA-vector estimation," *IEEE Signal Process. Lett.*, vol. 26, no. 4, pp. 572–576, 2019.

[11] A. Moore, C. Evers, P. A. Naylor, D. L. Alon, and B. Rafaely, "Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test," in *Proc. 23rd European Signal Process. Conf. (EUSIPCO-15)*, Nice, France, August 2015, pp. 2296–3000.

[12] S. Hafezi, A. H. Moore, and P. A. Naylor, "Augmented intensity vectors for direction of arrival estimation in the spherical harmonic domain," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 10, pp. 1956–1968, 2017.

[13] M. B. Çöteli, O. Olgun, and H. Hacıhabiboğlu, "Multiple Sound Source Localization With Steered Response Power Density and Hierarchical Grid Refinement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2215–2229, 2018.

[14] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1494–1509, 2014.

[15] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2193–2206, 2013.

[16] D. Pavlidi, S. Delikaris-Manias, V. Pulkki, and A. Mouchtaris, "3D localization of multiple sound sources with intensity vector estimates

in single source zones," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1556–1560.
[17] ——, "3D DOA estimation of multiple sound sources based on spatially constrained beamforming driven by intensity vectors," in *2016 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2016, pp. 96–100.
[18] O. Olgun and H. Hacihabiboglu, "Localization of multiple sound sources in the spherical harmonic domain with hierarchical grid refinement and EB-MUSIC," *2018 Int. Workshop on Acoust. Signal Enhancement (IWAENC-18)*, 2018.
[19] M. B. Çöteli and H. Hacıhabiboğlu, "Multiple sound source localization with rigid spherical microphone arrays via residual energy test," in *2019 IEEE Int. Conf. Acoust. Speech and Signal Process. (ICASSP-19)*, 2019, pp. 790–794.
[20] S. Yan, H. Sun, U. P. Svensson, X. Ma, and J. M. Hovem, "Optimal modal beamforming for spherical microphone arrays," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 2, pp. 361–371, 2010.
[21] N. R. Shabtai and B. Rafaely, "Generalized spherical array beamforming for binaural speech reproduction," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 238–247, 2013.
[22] A. Fahim, P. N. Samarasinghe, and T. D. Abhayapala, "PSD estimation of multiple sound sources in a reverberant room using a spherical microphone array," in *2017 IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA-17)*, 2017, pp. 76–80.
[23] C. T. Jin, N. Epain, T. Noohi, V. Pulkki, S. Delikaris-Manias, and A. Politis, "Sound field analysis using sparse recovery," in *Parametric Time-Frequency-Domain Spatial Audio*. John Wiley & Sons, 2017.
[24] P. K. T. Wu, N. Epain, and C. Jin, "A dereverberation algorithm for spherical microphone arrays using compressed sensing techniques," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP-2012)*, 2012, pp. 4053–4056.
[25] T. Noohi, N. Epain, and C. T. Jin, "Direction of arrival estimation for spherical microphone arrays by combination of independent component analysis and sparse recovery," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP-2013)*, 2013, pp. 346–349.
[26] K. Singhal and R. M. Hegde, "A sparse reconstruction method for speech source localization using partial dictionaries over a spherical microphone array," in *Proc. 15th Annual Conf. Int. Speech Comm. Assoc.*, 2014.
[27] T. Noohi, N. Epain, and C. T. Jin, "Super-resolution acoustic imaging using sparse recovery with spatial priming," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP-2015)*, 2015, pp. 2414–2418.
[28] S. N. Kalkur, S. Reddy C, and R. M. Hegde, "Joint source localization and separation in spherical harmonic domain using a sparsity based method," in *16th Ann. Conf. Int. Speech Comm. Assoc. (ISCA-15)*, 2015.
[29] V. Varanasi and R. Hegde, "Stochastic online dictionary learning for speech source localization and separation in spherical harmonic domain," in *2018 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP-18)*, 2018, pp. 66–70.
[30] N. Antonello, E. De Sena, M. Moonen, P. A. Naylor, and T. van Waterschoot, "Joint source localization and dereverberation by sound field interpolation using sparse regularization," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6892–6896.
[31] M. B. Çoteli and H. Hacıhabiboğlu, "Acoustic source separation using rigid spherical microphone arrays via spatially weighted orthogonal matching pursuit," *2018 Int. Workshop Acoust. Signal Enhancement (IWAENC-18)*, September 2018.
[32] N. A. Gumerov and R. Duraiswami, *Fast multipole methods for the Helmholtz equation in three dimensions*. Elsevier, 2005.
[33] B. Rafaely, *Fundamentals of Spherical Array Processing*, W. K. Jacob Benesty, Ed. Springer, 2015, vol. 8.
[34] B. Rafaely, B. Weiss, and E. Bachmat, "Spatial aliasing in spherical microphone arrays," *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 1003–1010, 2007.
[35] K. M. Gorski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann, "HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere," *Astrophys. J.*, vol. 622, no. 2, p. 759, 2005.
[36] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conf. on Signals, Systems and Computers*, Pacific Grove, CA, USA, November 1993, pp. 40–44.
[37] S. Mallat, *A wavelet tour of signal processing: the sparse way*. Burlington, MA, USA: Academic Press, 2008.
[38] P. Soille, *Morphological image analysis: principles and applications*. Springer Science & Business Media, 2013.
[39] K. V. Mardia, *Statistics of directional data*. London, UK: Academic Press, 2014.
[40] J. T. Kent, "The Fisher-Bingham distribution on the sphere," *J. Royal Stat. Soc.: Series B (Methodological)*, vol. 44, no. 1, pp. 71–80, 1982.
[41] D. Fraenkel, "Kent distribution," https://git.io/JYhuk, 2017.
[42] O. Olgun and H. Hacihabiboglu, "METU SPARG Eigenmike em32 Acoustic Impulse Response Dataset v0.1.0 (Version 0.1.0)," http://doi.org/10.5281/zenodo.2635758, 2019.
[43] European Broadcasting Union (EBU) Technical Centre, "Sound quality assessment material. recordings for subjective tests, cd and user's handbook for the EBU-SQAM Compact Disc, Tech. 3253-E." Brussels, Belgium, April 1988.
[44] O. Olgun and Hacıhabiboğlu, "Data-driven threshold selection for direct path dominance test," in *Proc. 23rd Int. Congr. on Acoust.*, Sep 2019, pp. 3313–3320.
[45] 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, 2006.
[46] International Telecommunications Union (ITU), "ITU-T P. 862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.
[47] S. Walser, C. Siegel, M. Winter, G. Feiertag, M. Loibl, and A. Leidl, "MEMS microphones with narrow sensitivity distribution," *Sensors and Actuators A: Phys.*, vol. 247, pp. 663 – 670, 2016.
[48] J. Citakovic, P. F. Hovesten, G. Rocca, A. van Halteren, P. Rombach, L. J. Stenberg, P. Andreani, and E. Bruun, "A compact CMOS MEMS microphone with 66 dB SNR," in *2009 IEEE Int. Solid-State Circ. Conf.*, 2009, pp. 350–351.
[49] International Telecommunications Union (ITU), "ITU-R BS. 1534-3: Method for the subjective assessment of intermediate quality level of audio systems," October 2015.
[50] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA—A comprehensive framework for web-based listening tests," *J Open Res. Soft.*, vol. 6, no. 1, 2018.
[51] M. Green and D. Murphy, "Eigenscape: A database of spatial acoustic scene recordings," *Applied Sciences*, vol. 7, no. 11, p. 1204, Nov 2017.

Mert Burkay Çöteli received his B.Sc. (honors) and M.S. in electrical and electronic engineering and Ph.D. in Information Systems from the Middle East Technical University (METU), Ankara, Turkey, in 2009, 2013, and 2021, respectively. He is working as a senior system engineer at Aselsan A.S. His research interests include microphone array signal processing and acoustic scene analysis.

Hüseyin Hacıhabiboğlu (S'96-M'00-SM'12) is an Associate Professor of Signal Processing at Graduate School of Informatics, Middle East Technical University, Ankara, Turkey. He received the B.Sc. (honors) degree from the Middle East Technical University (METU), Ankara, Turkey, in 2000, the M.Sc. degree from the University of Bristol, Bristol, U.K., in 2001, both in electrical and electronic engineering, and the Ph.D. degree in computer science from Queen's University Belfast, Belfast, U.K., in 2004. He held research positions at University of Surrey, Guildford, U.K. (2004-2008) and King's College London, London, U.K. (2008-2011). His research interests include audio signal processing, room acoustics, multichannel audio systems, psychoacoustics of spatial hearing, microphone arrays, and game audio. He is a member of the IEEE Signal Processing Society, a member of UKRI International Development Peer Review College, Audio Engineering Society (AES), Turkish Acoustics Society (TAD), and the European Acoustics Association (EAA). Between 2017-2021, he was an Associate Editor for the *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.