Sector-Based Parametric Sound Field Reproduction in the Spherical Harmonic Domain

Archontis Politis, *Student Member, IEEE*, Juha Vilkamo, and Ville Pulkki

***Abstract*—This papers presents a parametric method for perceptual sound field recording and reproduction from a small-sized microphone array to arbitrary loudspeaker layouts. The applied parametric model has been found to be effective and well-correlated with perceptual attributes in the context of directional audio coding, and here it is generalized and extended to higher orders of spherical harmonic signals. Higher order recordings are used for estimation of the model parameters inside angular sectors that provide increased separation between simultaneous sources and reverberation. The perceptual synthesis according to the combined properties of these sector parameters is achieved with an adaptive least-squares mixing technique. Furthermore, considerations regarding practical microphone arrays are presented and a frequency-dependent scheme is proposed. A realization of the system is described for an existing spherical microphone array and for a target loudspeaker setup similar to NHK 22.2. It is demonstrated through listening tests that, compared to a reference scene, the perceived difference is greatly reduced with the proposed higher order analysis model. The results further indicate that, on the same task, the method outperforms linear reproduction with the same recordings available.**

***Index Terms*—Array processing, multichannel recording, sound reproduction, 3D audio.**

I. Introduction

SPATIAL sound techniques that consider jointly the capturing and reproduction problem are essential for high-quality natural and immersive reproduction of realistic sound scenes and acoustic environments. They commonly start from a small number of recordings of the original sound scene, obtained with a microphone array around the point of interest. By processing these recordings appropriately, they aim at reproducing that scene to a discrete loudspeaker setup or headphones perceptually as close to the original as possible.

The various methods can be categorized roughly in three approaches: direct, non-parametric and parametric. Direct techniques are the traditional sound engineering approaches of optimizing heuristically the array geometry and directional properties of the microphones to achieve the desired perceptual performance at reproduction [1], [2]. Such techniques map one microphone signal to one loudspeaker, with no intermediate signal processing, and target specific reproduction setups. A formal analysis of their principles based on observation of the reproduced acoustic intensity field is presented in [3].

Non-parametric and parametric methods, conversely, allow significant flexibility in terms of the target reproduction system and potentially support various microphone array types. Linear non-parametric methods refer to approaches that are using a static frequency-dependent or frequency-independent mixing matrix distributing the input signals to the output signals, including ambisonics [4]–[7] and various static beamforming approaches [8]–[12]. Linear methods are based on knowledge of the array properties and the reproduction setup, independently of the captured sound field properties.

On the contrary, parametric recording and reproduction methods [13]–[21], operating in a time-frequency transform domain, combine the directional response of the array with a sound field model and extract the model's spatial parameters from the recordings. Based on the parameters and the input signals or a mixture of them, the captured sound scene is reconstructed on the reproduction system. The parameterization allows an "objectification" of the acoustic scene into discrete spatial components, that can be flexibly manipulated and rendered to arbitrary loudspeaker layouts or headphones. Many of these techniques are closely related to spatial audio coding (SAC) methods for compression and up-mixing of multichannel content [22], [23] which, however, target specific content formats and are outside the scope of this paper.

The method presented in this work is based upon the principles of Directional Audio Coding (DirAC) [21], which performs the spatial analysis/synthesis in a perceptually motivated way. DirAC estimates directional parameters which are subsequently used in the synthesis stage to recreate correct inter-aural directional cues and coherence that would occur for a listener in the recording position. The parameters are extracted from a first-order spherical harmonic (SH) recording format, known as B-format, even though it has been additionally applied to various stereophonic arrangements [24], uniform linear arrays [25] and spaced surround music recording arrays [26]. Binaural reproduction is also possible [27]. Using B-format input, the approach of DirAC has proved both efficient and effective and able to provide high perceived quality in reproduction of complex scenes, exceeding that of a linear rendering method using the same microphone array [28].

Although the reproduction quality of DirAC is high in most realistic recording cases, there are acoustic scenarios that violate the sound field model of the method, causing perceivable

Manuscript received July 10, 2014; revised December 14, 2014 and March 03, 2015; accepted March 09, 2015. Date of publication March 23, 2015; date of current version July 14, 2015. This work was supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC under Grant [240453]. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Sascha Spors.

The authors are with the Department of Signal Processing and Acoustics, School of Electrical Engineering, Aalto University, 02150 Espoo, Finland (e-mail: archontis.politis@aalto.fi).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2015.2415762

artifacts. Such cases occur, for instance, when spectrally overlapping sounds arrive at the array simultaneously from opposite directions and typically cause an overestimation of diffuseness. Since diffuse sound is rendered in DirAC by decorrelation techniques, the result in such cases can be timbral artifacts and smearing of transients [29], or an added reverberation effect for complex non-reverberant scenes [30].

A potential solution is provided by using higher-order SH recordings of the sound-field. Higher-order recordings naturally encode the directional properties of the sound field with increased spatial resolution and allow estimation of additional parameters to describe the acoustic scene, compared to a first-order B-format representation. In this work such a solution is formulated with the following inter-related objectives: a) resolve the issues of first-order DirAC by appropriate use of higher-order signals, b) retain the energetic analysis/synthesis scheme of DirAC due to its robustness and perceptual effectiveness, c) preserve the energy of the recording at reproduction equally well for all directions.

The proposed method divides the captured sound field into angular sectors within which the single plane-wave plus diffuse-field model parameters of DirAC are estimated. The number of sectors depends on the order. The method resolves the issues of first-order DirAC by reducing the effect of simultaneous sources or reflections incident from directions outside the sector. This article presents the theoretical background and practical implementation of the method, which is optimized using a regularized least-squares mixing technique proposed in [31]. In addition, a novel parametric extension to frequencies above the spatial aliasing frequency of the array, where SH processing of any kind fails, is presented with a performance similar to first-order DirAC. The performance of the implementation is shown to exceed linear non-parametric techniques with higher-order microphone arrays currently available.

The rest of the article is organized as follows. In Section II we present an overview of related methods, as well as the basic theoretical concepts and quantities involved in the proposed method. The general analysis and synthesis formulation is given in Section III. A practical implementation and related issues for an existing spherical microphone array are presented in Section IV. Finally, in Section V, listening test results and their analysis are shown.

II. Background

A review of spatial sound techniques with principles that relate to the proposed method or parts of it is presented below. It is followed by a description of the signal model and notation conventions used throughout the article.

A. Spatial Sound Recording and Reproduction Methods

Regarding linear non-parametric techniques, a flexible framework for the complete chain of recording, storing and reproducing spatial sound is ambisonics, pioneered in the 70s by Gerzon [4]. Ambisonics are based on an expansion of both the captured sound field and the loudspeaker distribution into spherical harmonics. A theoretical summary of the method along with practical limitations can be found in [5], [6]. On the recording side, even though regularly arranged spherical microphone arrays are favored, the SH signals can be obtained from irregular or even random arrangements [5], [32]. On the reproduction side, depending on the formulation, a frequency-independent or frequency-dependent decoding mixing matrix is produced for the SH signals. However, perceptually meaningful decoding matrices are difficult to obtain for irregular loudspeaker setups without some form of numerical optimization. An efficient solution that circumvents this limitation and is utilized in this work is proposed in [7]. The reproduction quality of ambisonics is related to the maximum order of the SH expansion that can be achieved by the microphone array and the target loudspeaker setup in use [33]–[36].

Non-parametric methods that mix input signals to the outputs, captured with directional patterns or with different propagation delays, correspond invariably to some form of beamforming, one per loudspeaker, including ambisonics. Based on this view, beamformers that satisfy other criteria than the least-squares solution of ambisonics can be used for multichannel reproduction of array recordings. The method in [12] bypasses the ambisonic formulation and creates directly a set of beamformers of the maximum allowed order of the array, covering uniformly the sphere. Their output signals can then be spatialized to the target reproduction setup. A different approach is conceptualized in [8] and formulated for a linear microphone array in [9] where the generated beam patterns implement an amplitude panning law for the target setup. In [10] the beamformers are adjusted to follow optimally the recording principles outlined in [3]. Similar approaches on recording with a portable spherical array for the standardized NHK 22.2 loudspeaker setup are presented in [11].

With regards to parametric analysis and reproduction of microphone recordings, the various approaches differ in the generality of their array model, the assumed sound-field model and their target application. For example, certain methods are more disposed to separation and enhancement of directional components rather than perceptual reconstruction of the whole captured scene [15], [16], [20], [37]. Regarding the array support, many methods are based on coincident arrangements that allow directional analysis based on inter-channel level differences or intensity analysis, such as stereophonic pairs in [14], or the B-format in DirAC [21] and in [15], [17]. Other methods assume spaced arrays and use time-difference of arrival (TDoA) techniques for direction-of-arrival (DOA) estimation of the directional components, such as [13], [16], [18]–[20], [37]. In terms of the sound field model and the associated parameter estimation, a common assumption is that of a single plane wave plus a uniform diffuse field. That model is followed by DirAC and the methods in [13] and [14]. The method of [18] assumes a sparse model of a single active source per time-frequency block, while the Harpex method estimates the parameters of two plane waves from a matrix decomposition of the B-format signals [17]. The number and the DOAs of multiple simultaneous active sources per block are estimated in [37], with an assumption of fully diffuse sound above the aliasing frequency of the array. In terms of higher-order SH recording and reproduction, a recent proposal based on compressed sensing theory is presented in [38], performing a sparse plane-wave decomposition of the SH signals. If combined with the direct/diffuse separation that the authors

have demonstrated in [39] for acoustic analysis, the sparse approach could be an alternative method to reproducing multiple directional components with diffuse sound.

B. Signal Model and Definitions

Quantities are defined in both the spatial domain and the spherical harmonic domain (SHD). The basics of the discrete spherical harmonic transform (SHT) and its application to sound-field recording are outlined. Vectors and matrices are denoted with boldface symbols, with lowercase for vectors and uppercase for matrices. All operations are defined in a time-frequency transform domain such as the short-time Fourier transform (STFT), or the complex-modulated quadrature mirror filter (QMF) bank. The discrete frequency and time indices of each time-frequency block are denoted as (k, l) respectively.

Spherical coordinates are written as $\mathbf{r} = (r, \Omega)$, where $\Omega = (\theta, \varphi)$ with inclination from the north pole $\theta \in [0, \pi]$ and azimuth $\varphi \in [-\pi, \pi]$. The SH coefficients of degree n and order m of a square-integrable function $f(\Omega)$ on the unit sphere S^2 are given by

$$f_{nm} = \int\limits_{\Omega \in S^2} f(\Omega) Y_{nm}^*(\Omega)\, \mathrm{d}\Omega \tag{1}$$

where complex conjugation is denoted by $(^*)$, and with integration on the unit sphere denoted as $\int_{\Omega \in S^2} \mathrm{d}\Omega = \int_{-\pi}^{\pi} \int_{0}^{\pi} \sin\theta \mathrm{d}\theta \mathrm{d}\varphi$. The complex orthonormalized spherical harmonics Y_{nm} of order n and degree m are

$$Y_{nm}(\Omega) = \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{im\varphi} \tag{2}$$

where P_n^m are associated Legendre functions of degree n and order m, and $i^2 = -1$ the imaginary unit. The discrete SHT form of (1) can be computed by sampling $f(\Omega)$ at Q directions

$$f_{nm} \approx \sum_{q=1}^{Q} c_q f(\Omega_q) Y_{nm}^*(\Omega_q) \tag{3}$$

where c_q are integration weights that preserve the orthonormality of the SHs. For the above equality to hold exactly up to some order N, $f(\Omega)$ should be band-limited to N and the sampling points should follow $Q \geq (N+1)^2$. There are various uniform and non-uniform sampling arrangements that fulfill (3), for a review of various schemes and their properties in spherical acoustic processing the reader is referred to [40] and [41]. Uniform sampling arrangements are ones such that the integration weights become equal, with $c_q = 4\pi/Q$. A class of uniform arrangements that are utilized later in the presented method are the spherical t-designs detailed in [42].

Let us consider a generic incident plane wave field expressed by the complex amplitude density $a(k, l, \Omega)$, sampled by an array of Q microphones around the origin, at positions $\mathbf{r}_q = (r_q, \Omega_q)$. It is assumed that the properties of the array are known, meaning its geometry and the directional response of its sensors. The microphone array signal vector $\mathbf{x}_Q$ is then

$$\mathbf{x}_Q(k, l) = \mathbf{p}_Q(k, l) + \mathbf{e}_Q(k, l) \tag{4}$$

where $\mathbf{p}_Q$ is the acoustic signal vector, determined by the acoustic field and the sensor responses, and $\mathbf{e}_Q$ is a vector of additive sensor noise that is white and of power σ_e^2 across all channels. The noise signals are assumed uncorrelated with the acoustic signals and between them.

Focusing in the case of a spherical array of radius $r_q = R$ with similar microphones, which is usually the case of interest, the discrete SHT of the amplitude density can be computed as

$$a_{nm}(k, l) \approx \frac{1}{b_n\left(\frac{\omega R}{c}\right)} \sum_{q=1}^{Q} c_q p_q(k, l) Y_{nm}^*(\Omega_q) \tag{5}$$

where b_n are the modal (or radial) weights for order n, ω is the angular frequency and c the speed of sound. The modal weights depend on the type of the array and their expressions can be found for various spherical array types in [43]. Division by them cancels the effect of the array to the estimated SH signals. If the array consists of omnidirectional or directional microphones at different radii and with unequal directional characteristics, the modal weights have to be modeled separately for each sensor and they cannot be factored out as in (5). In this case the SHT can be formulated in a least-squares sense as has been shown in [5], [32].

Practical spherical arrays that aim to perform the SHT of (3) commonly employ uniform or nearly-uniform microphone arrangements, since they require the smaller number of microphones for a maximum operational order N, see the first-order design example of [44] and the fourth-order examples in [6], [45]. The SH coefficients of (5) up to order N, are collected in a vector $\mathbf{a}_N = [a_{00}, a_{1(-1)}, a_{10}, a_{11}, \ldots, a_{NN}]^T$ of length $(N+1)^2$, which is referred to as SH signals. We define similarly the vector of SH values $\mathbf{y}_N(\Omega) = [Y_{00}(\Omega), Y_{1(-1)}(\Omega), \ldots, Y_{NN}(\Omega)]^T$. By applying the SHT of (5) to the noisy microphone signals $\mathbf{x}_Q$ in matrix form, assuming uniform arrangement, we obtain the vector of noisy SH signals $\tilde{\mathbf{x}}_N$ as

$$\tilde{\mathbf{x}}_N(k, l) = \frac{4\pi}{Q} \mathbf{B}_N^{-1}(k) \mathbf{Y}_N^H \mathbf{x}_Q(k, l) = \mathbf{a}_N(k, l) + \tilde{\mathbf{e}}_N(k, l) \tag{6}$$

where $\mathbf{Y}_N = [\mathbf{y}_N(\Omega_1), \mathbf{y}_N(\Omega_2), \ldots, \mathbf{y}_N(\Omega_Q)]^T$ is a $Q \times (N+1)^2$ SH matrix and $\mathbf{B}_N = \mathrm{diag}\{b_0, b_1, b_1, b_1, \ldots, b_N\}$ is a $(N+1)^2 \times (N+1)^2$ diagonal matrix of the modal weights. The symbol $()^H$ denotes the Hermitian transpose. The noise signal vector $\tilde{\mathbf{e}}_N$ is the SHT of the sensor noise and since the transform is linear, the transformed signals are also uncorrelated in the SHD across different (n, m).

Beamforming in the SHD reduces to simple weight-and-sum of the SH signals with the SH coefficients of the beam pattern. A real beam pattern of order $\leq N$ is defined in the spatial domain as $\mathrm{w}(\Omega)$ and equally in the SHD by its SH coefficients vector $\mathbf{w}_N$ obtained by applying (1) to $\mathrm{w}(\Omega)$. Then, the beamformer's output is given by

$$y(k, l) = \mathbf{w}_N^T \tilde{\mathbf{x}}_N(k, l). \tag{7}$$

In case the beam pattern is of order $< N$ then the above notation implies that its coefficients in the vector $\mathbf{w}_N$ are padded with zeros up to length $(N+1)^2$.

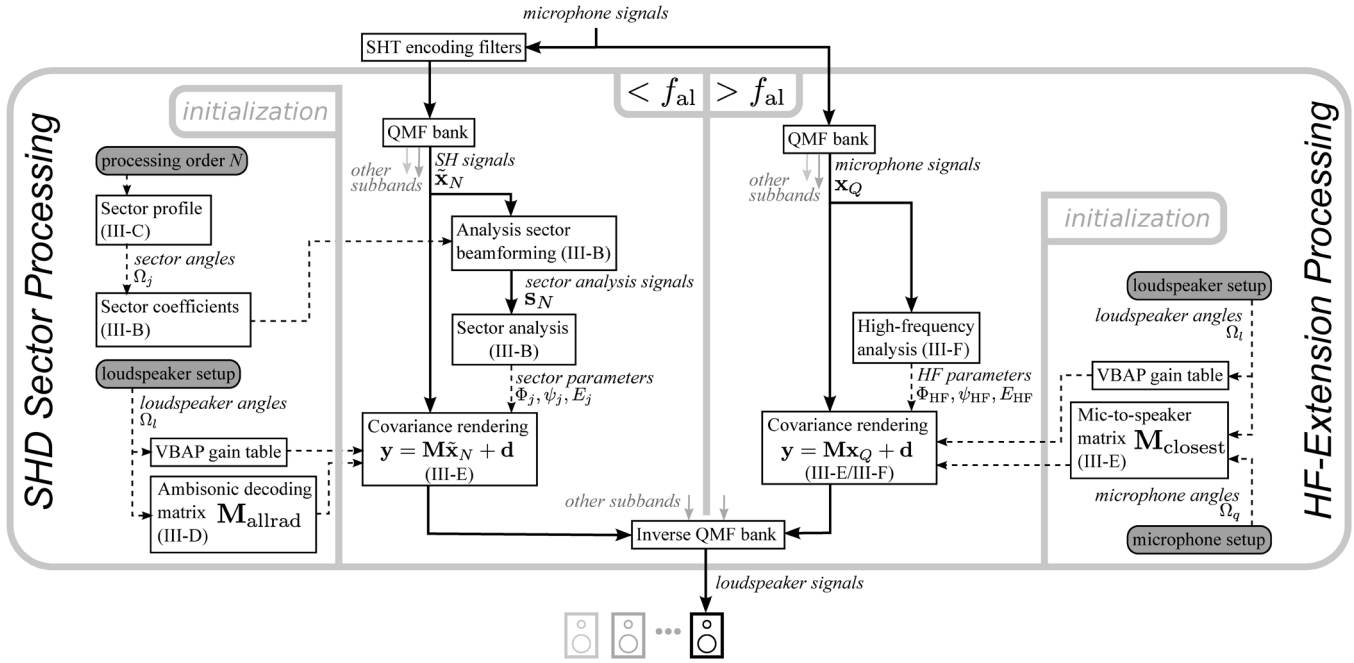

Fig. 1. Block diagram of the method, with specific blocks indicating inside parentheses the respective sections that describe them. The left half of the diagram applies to the SH signals and only at sub-bands below the aliasing frequency f_{al}. This part depicts the sector-based processing of Sections III-B, III-C, III-D & III-E. The initialization part refers to computations that are performed before the on-line processing, which depend only on the SH processing order for the current sub-band, and the loudspeaker setup. The right half of the diagram applies to microphone signals and only at sub-bands above the aliasing frequency. It depicts the high-frequency processing extension of Section III-F, applied when aliasing has an effect on reproduction quality.

The parametric sound field processing is performed based on signal statistics. We define the instantaneous covariance matrix expressing the inter-dependencies of the array signals in a single frame as $\hat{\mathbf{C}}_{\mathbf{x}_Q}(k,l) = \mathbf{x}_Q(k,l)\mathbf{x}_Q^H(k,l)$. Assuming stationarity of the signals, the true covariance matrix is $\mathbf{C}_{\mathbf{x}_Q}(k) = \mathbf{E}[\hat{\mathbf{C}}_{\mathbf{x}_Q}(k,l)] = \mathbf{E}[\mathbf{p}_Q(k,l)\mathbf{p}_Q^H(k,l)] + \sigma_e^2\mathbf{I}$, where $\mathbf{E}[\cdot]$ expresses statistical expectation. It can be estimated by averaging the instantaneous covariance matrices over multiple windows or by some recursive scheme. Similarly, the covariance matrix of the SH signals of (6) is

$$\mathbf{C}_{\tilde{\mathbf{x}}_N}(k) = \mathbf{E}\left[\hat{\mathbf{C}}_{\tilde{\mathbf{x}}_N}(k,l)\right] = \mathbf{E}\left[\hat{\mathbf{C}}_{\mathbf{a}_N}(k,l)\right] + \frac{4\pi\sigma_e^2}{Q}\mathbf{B}_N^{-2} \quad (8)$$

where $\hat{\mathbf{C}}_{\mathbf{a}_N}(k,l) = \mathbf{a}_N(k,l)\mathbf{a}_N^H(k,l)$ and $\mathbf{B}_N^{-2} = \mathrm{diag}\{1/|b_0|^2, 1/|b_1|^2, \ldots, 1/|b_N|^2\}$. The last term in (8) expresses the power spectral densities (PSDs) of $\tilde{\mathbf{e}}_N$ and it is derived by using (5) and the orthonormality of SHs. It is obvious from (8) that, contrary to the sensor noise, the noise power of $\tilde{\mathbf{e}}_N$ is not spectrally flat and is determined by the inverse of the modal weights, with severe amplification at lower frequencies and for increasing orders [43]. This fact limits in practice the usable range of the higher-order SH signals captured with a small array, and it is studied in more detail in the implementation description, in Section IV-B.

III. Method

The method consists of two main stages, the analysis stage, where energetic spatial sound field parameters are estimated in the SHD, and the synthesis stage, where these parameters for each time frame are used to adaptively mix the signals in a way that the spatial characteristics of the original sound scene are reconstructed in a perceptual sense. The parameters are the DOA of the net flow of sound energy and the diffuseness. Diffuseness expresses the ratio of non-directional sound energy to the total and is directly related to the direct-to-diffuse sound ratio (DDR), as shown in [46] for a single plane-wave plus diffuse field model. Based on the DOA and diffuseness and the loudspeaker setup, in the synthesis stage the diffuse part of the recording is reproduced surrounding the listener, ideally with zero coherence between loudspeakers. The non-diffuse sound is reproduced at the analyzed DOAs by means of vector-base amplitude panning (VBAP) [47].

Employing orders higher than one gives the possibility to estimate additional directional parameters in a single time-frequency tile. More specifically, depending on the SH order, the sphere can be subdivided by beamforming into sectors in which a local DOA and diffuseness can be estimated. Perfectly non-overlapping analysis sectors correspond to beamformers of infinite-order, hence in practice they are approximated with energy-preserving overlapping patterns of the maximum usable order in each frequency band. The largest expected benefit of the sector-based processing is in reproduction of sound scenes with multiple simultaneous sources or strong reflections. A further benefit is that the diffuse component has a directional distribution and therefore it covers the cases of non-uniform reverberation and sources with spatial extent.

A general block diagram of the method is shown in Fig. 1. The principles of the method can be summarized as follows:

- It is assumed that in a certain frequency band below the spatial aliasing frequency of the array f_{al}, the SH signals $\mathbf{a}_N$ of the amplitude density are order-limited to order $N_k \leq N$, where N is the maximum order supported by the

array configuration. A minimum of first-order approximation $N_k = 1$ is assumed for all frequencies below f_{al}.
- Based on the order N_k of a certain frequency band, a number of perceptually meaningful spatial parameters can be extracted in spatially separated angular sectors, where the formation of the sectors and the analysis signals is performed with beamforming operations. The number of sectors/parameters and the beamforming weights are determined solely by the current operational order N_k. This part is presented in Sections III-B & III-C.
- The analysis signals are detached from the synthesis, and are constructed only to obtain the spatial parameters.
- The synthesis can be formulated as an adaptive mixing problem for each time-frequency tile. Its least-squares solution produces a mixing matrix so that the input SH signals result in loudspeaker signals that match optimally the energetic spatial parameters of the analysis stage. This synthesis approach is discussed in Section III-E.
- Apart from matching the spatial parameters, the mixing solution of Section III-E is constrained so that the loudspeaker signals resemble temporally as much as possible a linear decoding of the SH signals. This approach incorporates higher-order static beamforming, or an ambisonic solution, into the synthesis method, and such a decoding example used in the current implementation is presented in Section III-D.

Finally, a parametric approach for analysis and synthesis above the aliasing limit is presented in Section III-F, which can be applied if aliasing occurs at frequencies that compromise the quality of reproduction. For the rest of the section the time index l of the time-frequency transform is omitted for brevity.

A. Analysis: Energetic Sound Field Analysis

For the following presentation of the spatial analysis and the estimation of the model parameters, the following assumptions are made. Firstly, it is assumed that in the kth analysis sub-band the array captures the sound field SH coefficients up to order N_k, and that the sound field is directionally band-limited to that order $\mathbf{a}_{N'>N_k} = 0$. This assumption holds in practice since higher-order components captured with a practical small-sized array decay rapidly at low frequencies. Secondly, it is assumed that in the same sub-band the transformed sensor noise $\tilde{\mathbf{e}}_{N_k}$ of (6) permits an adequate signal-to-noise ratio (SNR), so that the SHT of the microphone signals of (6) approximates the noiseless SH signals $\tilde{\mathbf{x}}_{N_k} = \mathbf{a}_{N_k}$, for practical purposes. Both conditions are met in the implementation by assigning an appropriate analysis/synthesis order for each sub-band, based on an analysis of the array noise amplification at different orders and frequencies, and by an alternative analysis/synthesis module for the high-frequency range above the spatial aliasing limit.

The pressure and the acoustic particle velocity at the origin due to the measured amplitude density are given by

$$p(k) = \int_{\Omega} a(k,\Omega)\, \mathrm{d}\Omega \tag{9}$$

$$\mathbf{u}(k) = -\frac{1}{Z_0}\int_{\Omega} a(k,\Omega)\mathbf{n}(\Omega)\, \mathrm{d}\Omega = -\frac{1}{Z_0}\mathbf{v}(k) \tag{10}$$

with $\mathbf{n}(\Omega) = [\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta]^T$ the unit vector pointing to the direction of incidence. The signal vector $\mathbf{v}(k) = [\mathrm{v_x}(k), \mathrm{v_y}(k), \mathrm{v_z}(k)]^T$ corresponds to the negative unnormalized Cartesian components of the particle velocity and $Z_0 = c\rho_0$ is the characteristic impedance of air. The pressure signal can be captured with an omnidirectional pattern, while the velocity, as it is obvious from (10), can be captured with three dipole patterns $\mathrm{x}(\Omega), \mathrm{y}(\Omega), \mathrm{z}(\Omega)$ corresponding to the components of $\mathbf{n}(\Omega)$ as

$$\mathbf{n}(\Omega) = \begin{bmatrix} \mathrm{x}(\Omega) \\ \mathrm{y}(\Omega) \\ \mathrm{z}(\Omega) \end{bmatrix} = \begin{bmatrix} \sin\theta\cos\varphi \\ \sin\theta\sin\varphi \\ \cos\theta \end{bmatrix}. \tag{11}$$

The signals captured by these dipoles form the signal vector $\mathbf{v}(k)$ and together with the omnidirectional pressure signal $p(k)$ they form the B-format signal set. Furthermore, the omnidirectional signal is related to the zeroth-order SH and the dipole signals to the first-order SH by the following relations

$$\mathbf{s}_1(k) = \begin{bmatrix} p(k) \\ \mathbf{v}(k) \end{bmatrix} = [\mathbf{w}_1, \mathbf{x}_1,\ \mathbf{y}_1,\ \mathbf{z}_1]^T \mathbf{a}_1(k) \tag{12}$$

where $\mathbf{w}_1 = [\sqrt{4\pi}, 0, 0, 0]^T$ are the SH coefficient of the omnidirectional component and

$$[\mathbf{x}_1,\ \mathbf{y}_1,\ \mathbf{z}_1] = \begin{bmatrix} 0 & 0 & 0 \\ \sqrt{\frac{2\pi}{3}} & \sqrt{\frac{2\pi}{3}}i & 0 \\ 0 & 0 & \sqrt{\frac{4\pi}{3}} \\ -\sqrt{\frac{2\pi}{3}} & \sqrt{\frac{2\pi}{3}}i & 0 \end{bmatrix} \tag{13}$$

are the SH coefficients of the dipole patterns.

To estimate the energetic quantities of interest we define the following quantities. The instantaneous power spectrum of the pressure signal is denoted as $\hat{S}_{\mathrm{pp}}(k) = |p(k)|^2$, the combined power spectrums of the velocity signals as $\hat{S}_{\mathrm{vv}}(k) = \mathbf{v}^H(k)\mathbf{v}(k)$, and the cross-spectrum between pressure and velocity is denoted as the vector $\hat{\mathbf{s}}_{\mathrm{pv}}(k) = p^*(k)\mathbf{v}(k)$. From these quantities, the active intensity vector $\mathbf{i}_a(k)$, the energy density $E(k)$ and the diffuseness $\psi(k)$ for a sub-band can be conveniently measured as [46]

$$\mathbf{i}_{\mathrm{a}}(k) = -\Re\{\hat{\mathbf{s}}_{\mathrm{pv}}(k)\} \tag{14}$$

$$E(k) = \frac{1}{2}\left[\hat{S}_{\mathrm{vv}}(k) + \hat{S}_{\mathrm{pp}}(k)\right] \tag{15}$$

$$\psi(k) = 1 - \frac{2\,\|\Re\{\hat{\mathbf{s}}_{\mathrm{pv}}(k)\}\|}{\hat{S}_{\mathrm{vv}}(k) + \hat{S}_{\mathrm{pp}}(k)} \tag{16}$$

where a constant factor of $1/(2Z_0)$ has been omitted from (14) and a factor of $1/(2cZ_0)$ from (15). These factors do not affect the use of the parameters and (15) ensures that the energy density is normalized to be equal to the power of the omnidirectional signal for a single plane wave or uniform diffuse field. The diffuseness of (16) is derived from the relation $\psi = 1 - \|\Re\{\mathbf{i}_a\}\|/(cE)$, where the intensity and energy density relations are used in the derivation including the omitted factors. Diffuseness is bounded between $\psi \in [0,1]$ with $\psi = 0$ for a single plane wave. Its maximum value $\psi = 1$ is obtained only in a standing wave field, in which case the intensity of (14)

vanishes, and in a purely diffuse field after adequate time-averaging between successive time-frames.

The DOA of the mean energy flow $\Phi = (\theta_{\mathrm{DOA}}, \varphi_{\mathrm{DOA}})$ is extracted as the opposite direction to the intensity vector as

$$\mathbf{n}(\Phi(k)) = \frac{\Re\{\hat{\mathbf{s}}_{\mathrm{pv}}(k)\}}{\|\Re\{\hat{\mathbf{s}}_{\mathrm{pv}}(k)\}\|}. \tag{17}$$

The energy density E, diffuseness ψ and energetic DOA Φ are the sound field parameters that are required for the perceptual reconstruction in the synthesis stage.

B. Analysis: Sector-Based Higher-Order Energetic Analysis

Let us consider that in a certain sub-band aliasing-free SH signals up to order N are supported. Assuming a directional weighting of the amplitude density function expressed by the beam pattern $\mathrm{w}(\Omega)$ of order $N-1$ and respectively its SH coefficients $\mathbf{w}_{N-1}$, the weighted pressure signal captured is given, similar to (7), by $p_{\mathrm{w}}(k) = \mathbf{w}_{N-1}^T \mathbf{a}_{N-1}(k)$. We refer to the beam pattern $\mathbf{w}_{N-1}$ as a sector pattern. The velocity due to the same weighted distribution is given by

$$\mathbf{u}_{\mathrm{w}}(k) = -\frac{1}{Z_0}\int_{\Omega} \mathrm{w}(\Omega)\mathbf{n}(\Omega)\, a(k,\Omega)\, \mathrm{d}\Omega = -\frac{1}{Z_0}\mathbf{v}_{\mathrm{w}}(k), \tag{18}$$

where the signal vector $\mathbf{v}_w$ corresponds to the signals captured with the directional patterns $\mathrm{w}(\Omega)\mathbf{n}(\Omega)$. It is evident from (18) that it is possible to measure the velocity components $\mathbf{u}_{\mathrm{w}}$ of the weighted field if we are able to generate beam patterns that are products of the original pattern and the three orthogonal dipoles as in

$$\mathrm{w}(\Omega)\mathbf{n}(\Omega) = \begin{bmatrix} \mathrm{w}_{\mathrm{x}}(\Omega) \\ \mathrm{w}_{\mathrm{y}}(\Omega) \\ \mathrm{w}_{\mathrm{z}}(\Omega) \end{bmatrix} = \begin{bmatrix} \mathrm{w}(\theta,\varphi)\sin\theta\cos\varphi \\ \mathrm{w}(\theta,\varphi)\sin\theta\sin\varphi \\ \mathrm{w}(\theta,\varphi)\cos\theta \end{bmatrix}. \tag{19}$$

These velocity beam patterns are of order N, since they are products of an $(N-1)$-order sector pattern and the first-order components of $\mathbf{n}(\Omega)$. Their coefficients $[\mathbf{w}_N^{\mathrm{x}}, \mathbf{w}_N^{\mathrm{y}}, \mathbf{w}_N^{\mathrm{z}}]$ are linearly connected to the sector coefficients $\mathbf{w}_{N-1}$ and can be found analytically as

$$\mathbf{w}_N^{\mathrm{i}} = \mathbf{A}_N^{\mathrm{i}} \mathbf{w}_{N-1}, \quad \text{with } i = \{x, y, z\}. \tag{20}$$

The matrices $\mathbf{A}_N^{\mathrm{x}}, \mathbf{A}_N^{\mathrm{y}}, \mathbf{A}_N^{\mathrm{z}}$ are $(N+1)^2 \times N^2$ sparse deterministic matrices that depend only on the SH coefficients of the dipoles of (13) and the order N. They can be pre-computed for some maximum order and then be applied to any beam pattern up to that order, as in (20). The exact derivation is lengthy and is omitted in this work, however their structure and the steps to derive them are presented at http://research.spa.aalto.fi/publications/papers/ho_dirac/, along with pre-computed matrices up to order $N = 21$.

After the sector and velocity patterns are determined, the Nth-order analysis signals $\mathbf{s}_N(k)$ can be obtained from the SH signals similar to the first-order case

$$\mathbf{s}_N(k) = \begin{bmatrix} p_{\mathrm{w}}(k) \\ \mathbf{v}_{\mathrm{w}}(k) \end{bmatrix} = [\mathbf{w}_N, \mathbf{w}_N^{\mathrm{x}}, \mathbf{w}_N^{\mathrm{y}}, \mathbf{w}_N^{\mathrm{z}}]^T\, \mathbf{a}_N(k) \tag{21}$$

and their instantaneous power and cross-spectra can be similarly defined as $\hat{S}_{\mathrm{pp,w}}(k)$, $\hat{S}_{\mathrm{vv,w}}(k)$ and $\hat{\mathbf{s}}_{\mathrm{pv,w}}(k)$ respectively.

With the generation of the patterns of (19) and the capture of the sector and velocity signals p_{w} and $\mathbf{v}_{\mathrm{w}}$, it is possible to estimate the energetic quantities of the previous section in a non-global manner but instead with a specific directional selectivity. More specifically, a local active intensity $\mathbf{i}_{\mathrm{w}}$, energy density E_{w} and local diffuseness ψ_{w} can be estimated with the exact same formulas of (14)–(16), if the total pressure and velocity power and cross-spectra are replaced by their spatially filtered versions.

C. Analysis: Sector Profiles

Based on the analysis scheme of the previous section, it is obvious that with higher-order SH signals it is possible to generate multiple sets of spatial parameters that can be utilized in the synthesis stage, one for each sector beamformer. If the array cannot support orders higher than one, then the method reduces to the first-order analysis of B-format DirAC for frequencies below the spatial aliasing limit. When higher-order signals are available, then the sphere is divided into sector patterns. The number, shape and orientation of the sector beams depend on the order and the application. For the generalized analysis/synthesis scheme presented in this work the following conditions should be met. Firstly, it is desired that the analysis performance is equal in all directions. This condition imposes similar axisymmetric sector patterns covering uniformly the sphere. Secondly, since the sector energy densities are used in the synthesis to distribute spatially the sound to the loudspeakers while preserving the energy of the recording, there should be no loss of energy at any direction for all sectors. This condition can be formulated as

$$\sum_{j=1}^{J} \beta \mathrm{w}_j^2(\Omega) = 1, \quad \forall\, \Omega \tag{22}$$

where J is the number of sectors, w_j is the sector pattern and β is a normalization constant that depends on the sector scheme. A final desired condition is that the number of the sector patterns is the minimum one that meets the previous conditions, as additional sectors increase the computational load without additional benefits.

There are only certain designs that fulfill these conditions. For an analysis order N, a direct solution is provided by minimal spherical t-designs of $t = 2N-2$. A spherical t-design defines a set of points on the sphere for which the integral of all spherical polynomials of degree $N \leq t$ is equal to their discrete sum across these points

$$\int_{\Omega} \mathrm{w}(\Omega)\, \mathrm{d}\Omega = \frac{4\pi}{J}\sum_{j=1}^{J} \mathrm{w}(\Omega_j) \tag{23}$$

where Ω_j are the directions of the points of the t-design. The term minimal refers to the t-design with the lowest number of points. Considering any axisymmetric analysis sector of order $N-1$, the condition of (22) is met if J sectors are oriented at angles Ω_j, with only the normalization constant β dependent on the shape of the sectors. The normalization constant β in this case reduces to

$$\beta = \frac{Q_{\mathrm{w}}}{J} \tag{24}$$

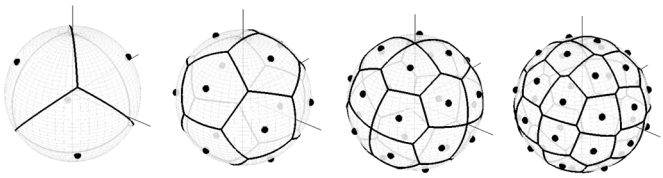

Fig. 2. Sector profiles for analysis orders N = 2–5, with the sector centers highlighted and the voronoi areas around them.

TABLE I
SECTOR SCHEMES FOR DIFFERENT ORDERS OF ANALYSIS AND RESPECTIVE NORMALIZATION FOR HIGHER-ORDER CARDIOID PATTERNS

Sector Order $N-1$	**Geometry**	**Num. of Sectors** J	**Normalization** β
1	reg. tetrahedron	4	3/4
2	reg. icosahedron	12	5/12
3	improved snub cube [42]	24	7/24
4	snub tetrahedra [42]	36	1/4

where $Q_w = (4\pi)/\mathbf{w}_{N-1}^H \mathbf{w}_{N-1}$ is the directivity factor of the sector pattern. A proof of the condition (22) and relation (24) is given in the Appendix.

Spherical t-designs can be found tabulated in [42]. A summary of them for the first four orders is given in Table I and a visual presentation in Fig. 2. The shape of the sectors in the present implementation is chosen to be that of higher-order cardioids, which are conceptually simple with only a single positive lobe and a single null in the opposite direction of the sector's orientation. The higher-order cardioid pattern of order N is described by the formula

$$\mathrm{w}_{\mathrm{card}}^N(\alpha) = \left(\frac{1}{2}\right)^N (1+\cos\alpha)^N \tag{25}$$

where $\cos\alpha = \mathbf{n}^T(\Omega)\cdot\mathbf{n}(\Omega_j)$ is the cosine of the angle between the DOA and the sector's orientation. Their directivity factor is also readily available as $Q_{\mathrm{card}}(N) = 2N+1$.

It is demonstrated how the sector-based scheme analyzes correctly the DOA and energy in the fundamental case of a single plane wave. Let us assume that the plane wave is incident to the array from the DOA Ω_0 and it carries a signal with power P_{pw}. Following the beamforming operations of Section III-B for J uniformly-arranged sectors, the PSDs and CSD of the analysis signals for a single sector $\mathrm{w}_j(\Omega)$ become $\hat{S}_{\mathrm{pp,w}} = \hat{S}_{\mathrm{vv,w}} = \beta \mathrm{w}_j^2(\Omega_0) P_{\mathrm{pw}}$ and $\hat{\mathbf{s}}_{\mathrm{pv,w}} = \beta \mathrm{w}_j^2(\Omega_0) P_{\mathrm{pw}} \mathbf{n}(\Omega_0)$. The property of the sector patterns $\mathrm{w}_j^2(\Omega) = \mathrm{w}_{j\mathrm{x}}^2(\Omega) + \mathrm{w}_{j\mathrm{y}}^2(\Omega) + \mathrm{w}_{j\mathrm{x}}^2(\Omega)$ is used in their derivation. Finally, the sector-based intensity vector, energy density, difuseness and analyzed DOA from (14)–(17) are

$$\begin{aligned} \mathbf{i}_j &= -\beta \mathrm{w}_j^2(\Omega_0) P_{\mathrm{pw}} \mathbf{n}(\Omega_0) \\ E_j &= \beta \mathrm{w}_j^2(\Omega_0) P_{\mathrm{pw}} \\ \psi_j &= 0 \\ \Phi_j &= \Omega_0. \end{aligned} \tag{26}$$

It is clear from (26) that in all sectors the estimated DOA points to the plane wave direction. Consequently, the signal energy contributed by each sector at that direction at synthesis is E_j, which based on the energy preserving property of the sectors (22) results in the correct plane wave power

$$\sum_{j=1}^{J} E_j = \sum_{j=1}^{J} \beta \mathrm{w}_j^2(\Omega_0) P_{\mathrm{pw}} = P_{\mathrm{pw}}, \quad \forall\, \Omega_0. \tag{27}$$

D. Synthesis: Non-Parametric Beamforming Stage

The high-quality variant of first-order DirAC, as presented in [28], employs B-format signals to generate first-order beams, termed in literature as virtual microphones, as an intermediate stage for distribution of the input signals to the directions of the loudspeakers. This approach combines the advantage of a non-parametric linear decoding, with high single-channel quality, with the perceptual reproduction of the parametric approach, and it reduces musical noise that can result from incorrect estimation of the model parameters in the parametric approach. Furthermore, the effort of spatially distributing the direct sound and producing decorrelated outputs for the diffuse sound is handled partially by the static beamformers. In the present method, instead of higher-order virtual microphones, the beams are formed by an ambisonic decoding matrix optimized for irregular layouts. An even directional distribution of energy that takes into account the density of the loudspeaker setup is achieved in this manner. The decoding matrix is computed according to the efficient solution for ambisonic rendering on irregular layouts presented in [7], termed All-round Ambisonic Decoding (ALLRAD). For a detailed description the reader is referred to [7], however for sake of completeness a summary of the basic steps is reproduced here. Starting from a description of the loudspeaker setup, given in angles Ω_l,

(a) compute an equivalent ambisonic order of decoding N_{amb}, based on the number of loudspeakers L and their average angular density,
(b) select a minimal t-design of $t = 2N_{\mathrm{amb}}+1$ with P vertices at directions Ω_p,
(c) generate a $P\mathrm{x}(N_{\mathrm{amb}}+1)^2$ ambisonic decoding matrix $\mathbf{M}_{\mathrm{amb}}$, for P virtual loudspeakers at the uniform angles Ω_p,
(d) generate an $L\mathrm{x}P$ VBAP gain matrix $\mathbf{M}_{\mathrm{vbap}}$, for rendering the P virtual loudspeakers signals at the L real ones.

The final mixing matrix to generate the loudspeaker signals is

$$\mathbf{M}_{\mathrm{allrad}} = \mathbf{M}_{\mathrm{vbap}}\, \mathbf{M}_{\mathrm{amb}}. \tag{28}$$

Finally, the loudspeaker signals for the linear non-parametric decoding can be obtained by

$$\mathbf{y}_{\mathrm{lin}}(k) = \mathbf{M}_{\mathrm{allrad}} \tilde{\mathbf{x}}_N(k), \tag{29}$$

however, instead of using these signals directly, their properties are combined in a single step with the parametric processing as described in the following section.

E. Synthesis: Parametric Sound Scene Rendering Based on the Model Parameters

The analysis stage provides a set of parameters describing the sound field as a function of time and frequency. At the synthesis stage, loudspeaker signals are processed from the microphone

signals such that the result corresponds to the analyzed parameters. In detail, the non-diffuse portions of the sound energy are reproduced at their estimated directions. The diffuse portions of the sound energy are reproduced with all loudspeakers with mutual incoherence. Although the sound field parameters are known, the audio signals corresponding to the non-diffuse or diffuse parameters are not available as independent signals.

The early synthesis techniques in first-order DirAC divided the frequency band signals into separate non-diffuse and diffuse signals using amplitude weighting according to the analyzed parameters [21]. The non-diffuse part was amplitude gated to the analyzed direction-of-arrival and the diffuse part was produced with all loudspeakers using decorrelators. More recently, in [48] it was shown that in a perceptual sense higher quality of reproduction can be obtained by first building the target loudspeaker characteristics in the parametric domain, and then processing the non-diffuse and diffuse sounds in a single combined least-squares optimized mixing step. Furthermore, as shown in the following, this approach simplifies the present task of synthesizing a combined signal from all sector profiles into the simple additive process of defining a target covariance matrix.

The target covariance matrix $\mathbf{C_y}(k)$ for a time-frequency block is built as a function of the model parameters Φ_j, ψ_j, and E_j, where $j = 1, \ldots, J$, and J is the number of sound field sectors applied in the analysis. For a sector j, the estimated energy of the diffuse part is $\psi_j E_j$, and the energy of the non-diffuse part is $(1-\psi_j)E_j$. Firstly, the non-diffuse energy is steered at the estimated direction Φ_j using energy normalized vector base amplitude panning (VBAP) [47] gains $\mathbf{g}(\Phi_j)$. For sector j, the target covariance matrix for the non-diffuse sound is $E_j \mathbf{g}(\Phi_j)\mathbf{g}^H(\Phi_j) \cdot (1 - \psi_j)$. Secondly, the diffuse sound energy is distributed incoherently to the loudspeakers. The target covariance matrix for the diffuse energy of the sound is $E_j \mathbf{D}_j \cdot \psi_j$, where $\mathbf{D}_j$ is a diagonal diffuse energy distributor matrix with a property $\mathrm{tr}(\mathbf{D}_j) = 1$, where $\mathrm{tr}(\cdot)$ is the matrix trace. Concluding the above, the target covariance matrix combining the parametric information from all sector profiles is

$$\mathbf{C_y} = \sum_{j=1}^{J} E_j \left[\mathbf{g}(\Phi_j)\mathbf{g}^H(\Phi_j) \cdot (1 - \psi_j) + \mathbf{D}_j \cdot \psi_j \right]. \quad (30)$$

Having the target covariance matrix defined, an optimized mixing solution can be formulated following the steps proposed in [31], which is now reviewed applying the present notation. The technique assumes an input-output relation

$$\mathbf{y}(k) = \mathbf{M}\tilde{\mathbf{x}}_N(k) + \mathbf{d}(k), \quad (31)$$

where M is a mixing matrix to process the input signal $\tilde{\mathbf{x}}_N(k)$ such that the output signal $\mathrm{y}(k)$ has the defined target covariance matrix $\mathbf{C_y}$. However, the solution is subject to regularization, which in turn limits the ability of the system to obtain the target. Signal $\mathbf{d}(k)$ is a synthesized residual signal that compensates for the effect of the regularization in a stochastic sense, and is defined by its covariance matrix

$$\mathbf{C_d} = \mathbf{C_y} - \mathbf{M}\mathbf{C_x}\mathbf{M}^H. \quad (32)$$

The requirement for (32) is that $\mathbf{d}(k)$ is incoherent with respect to $\tilde{\mathbf{x}}_N(k)$. As derived in [31], a set of mixing solutions is first formulated assuming an idealized condition that no residual signal $\mathbf{d}(k)$ is necessary. With this assumption the set of solutions providing $\mathbf{C_y}$ is

$$\mathbf{M} = \mathbf{K_y}\mathbf{P}\mathbf{K_x}^{-1}, \quad (33)$$

where $\mathbf{K_y}$ and $\mathbf{K_x}$ are matrix decompositions $\mathbf{C_y} = \mathbf{K_y}\mathbf{K_y}^H$ and $\mathbf{C_x} = \mathbf{K_x}\mathbf{K_x}^H$, and $\mathbf{P}$ is any unitary matrix. An error measure is then defined as

$$e(k) = \left\| \mathbf{y}(k) - \mathbf{G}\mathbf{y}_{\mathrm{lin}}(k) \right\|^2, \quad (34)$$

where $\mathbf{y}_{\mathrm{lin}}(k)$ are the linearly-decoded signals of (29) using ambisonics. The diagonal matrix $\mathbf{G}$ adapts the energies of $\mathbf{y}_{\mathrm{lin}}(k)$ to those of $\mathrm{y}(k)$, to ensure that the error measure is weighted with respect to the target channel energies. In this context, the ambisonic signals $\mathbf{G}\mathbf{y}_{\mathrm{lin}}(k)$ constitute a constraint for the least-squares mixing solution in terms of a desirable signal waveform for each output channel. Note that contrary to normal ambisonic decoding, the output channel energies, coherences and the spatial energy distribution are completely determined by the model parameters through the target covariance matrix $\mathbf{C_y}$. Minimizing $e(k)$ in (34) leads to

$$\mathbf{P} = \mathbf{V}\mathbf{U}^H, \quad (35)$$

where $\mathbf{V}$ and $\mathbf{U}$ are unitary matrices from a singular value decomposition $\mathbf{U}\mathbf{S}\mathbf{V}^H = \mathbf{K_x}^H \mathbf{M}_{\mathrm{allrad}}^H \mathbf{G}^H \mathbf{K_y}$.

Unless regularized, the inverse of $\mathbf{K_x}$ in (33) poses a problem for sound quality. A robust means for regularization is to formulate a singular value decomposition $\mathbf{U_x}\mathrm{S_x}\mathbf{V_x}^H = \mathbf{K_x}$, to lower limit the diagonal values of $\mathrm{S_x}$ to obtain a regularized diagonal matrix $\mathrm{S_x'}$, and to formulate the inverse as $\mathbf{K_x'}^{-1} = \mathbf{V_x}^H \mathrm{S_x'}^{-1} \mathbf{U_x}^H$. In the present test implementation, the diagonal values of $\mathrm{S_x'}$ were limited such that they were in minimum 0.2 times the largest diagonal value in $\mathrm{S_x}$. Replacing $\mathbf{K_x'}^{-1}$ in (33) causes that the target covariance matrix is no longer reached, which is compensated for by the additive residual signal with a covariance matrix $\mathbf{C_d}$ in (32). Additionally to its covariance matrix, the spectro-temporal content of the residual signal also matters. Therefore in the test implementation it is generated by applying decorrelating processes to the ambisonic signals $\mathbf{y}_{\mathrm{lin}}(k)$, and then applying the above mixing procedure such that the result obtains the residual covariance matrix $\mathbf{C_d}$. In design of the diagonal distributor matrix $\mathbf{D}_j$, the particular approach applied in the test implementation was to diagonalize the ambisonic signal covariance matrix $\mathbf{M}_{\mathrm{allrad}}\mathbf{C}_{\mathbf{x}_N}(k)\mathbf{M}_{\mathrm{allrad}}^H$, and normalize it so that its trace is unity to form $\mathbf{D}_j$. Such an approach provides a spatial distribution of the diffuse sound energy that has similarities to the distribution of the sound in the analyzed sound scene.

F. High-Frequency Processing Extension

Above the spatial aliasing limit the higher-order analysis/synthesis approach based on the SH signals cannot be used anymore, as both the analysis beam patterns and the synthesis decoding matrices become erratic. In the case of an array with directional microphones or microphones mounted on a rigid

baffle, an ad-hoc approach that preserves some of the directionality of the incident field is to use the microphone signal directly for the loudspeaker that its direction is closer to the microphone's orientation. This fixed input-output signal mapping can be expressed as

$$\mathbf{y}_{\mathrm{mic}}(k) = \mathbf{M}_{\mathrm{closest}}\mathbf{x}_Q(k) \tag{36}$$

where $\mathbf{M}_{\mathrm{closest}}$ is a LxQ matrix of zeros and ones mapping the microphone signals to the closest loudspeakers.

Whether the above approach is of sufficient quality depends on the application and the frequency at which spatial aliasing appears. For example, compact arrays intended for a maximum of fourth-order recording, with a radius in the range of 4∼8 cm, will have aliased components at around $6 \sim 3$ kHz, see also the design example in Section IV-B. Failing to reproduce correctly the spatial and energetic properties of the recording at this range can have a detrimental effect on the overall quality of reproduction [49]. We propose herein a parametric approach for the aliased region that can offer a performance close to the first-order parametric rendering of the method. The method assumes a uniform spherical or circular array of directional microphones or microphones mounted on a rigid sphere. A single plane wave model is assumed incident from direction Φ with no diffuse field present. It is known that the continuous pressure distribution on the array due to the plane wave exhibits an axisymmetric shape which is oriented towards the DOA of the plane wave. This pattern is frequency-dependent for the baffled array or frequency-independent for an array of ideal directional microphones. Let us denote this distribution as $|p_0|d(f,\Phi,\Omega)$ where $|p_0|$ is a quantity proportional to the plane wave magnitude. By taking the magnitude of this pattern $|p_0||d(f,\Phi,\Omega)|$ we neglect directional phase effects and interference which contribute to the aliasing. The magnitude of the distribution is also axisymmetric and oriented towards Φ. Finally, integration of the unit vector $\mathbf{n}(\Omega)$ on the unit sphere, weighted with the directional pressure magnitude, results in a vector oriented also at Φ due to the symmetry of the magnitude distribution

$$\alpha(f)\mathbf{n}(\Phi) = |p_0| \int_{\Omega} |d(f,\Phi,\Omega)|\, \mathbf{n}(\Omega)\, \mathrm{d}\Omega \tag{37}$$

where α is the magnitude of the vector, proportional to the plane wave magnitude. In practice, for a discrete array this DOA vector can be estimated in the kth sub-band by approximating the integration of (37) by the discrete summation of

$$\mathbf{r}_{\mathrm{HF}}(k) = \alpha(k)\, \mathbf{n}\left(\Phi_{\mathrm{HF}}(k)\right) = \sum_{q=1}^{Q} |x_q(k)|\, \mathbf{n}(\Omega_q) \tag{38}$$

from which the high-frequency DOA $\Phi_{\mathrm{HF}}(k)$ can be extracted. An analysis window longer than the temporal dimensions of the array is required for correct estimation. This DOA vector has been used before for a first-order array of cardioid microphones in [49] and for widely spaced multichannel arrays for music recording in [26].

Since pressure and velocity cannot be estimated correctly above the aliasing frequency, the energetic diffuseness estimation of (16) cannot be used. Instead an alternative diffuseness estimator is used based on the temporal variation of the intensity vector [50]. Herein, the intensity vector is replaced by the DOA vector of (37) and the diffuseness is given by

$$\psi_{\mathrm{HF}}(k) = \sqrt{1 - \frac{\|\mathbf{E}\left[\mathbf{r}_{\mathrm{HF}}(k)\right]\|}{\mathbf{E}\left[\|\mathbf{r}_{\mathrm{HF}}(k)\|\right]}}. \tag{39}$$

In the case of an array of omnidirectional microphones that do not provide any DOA-dependent directionality, the previous method cannot be used. Alternatively, the approach of [51] can be used, based on TDOA of the envelope of the signal in the high-frequency region.

Finally, after the high-frequency parameters have been extracted, we can apply the least-squares mixing technique of Section III-E for rendering. Since SH signals are not usable, the method is formulated directly for the microphone signals, with the input covariance matrix $\mathbf{C}_{\mathbf{x}_Q}$, and by replacing the ambisonic signals $\mathbf{y}_{\mathrm{lin}}$ with the microphone signals $\mathbf{y}_{\mathrm{mic}}$. The target covariance matrix is constructed as in (30) for a single sector, with E_{HF} the mean energy of the microphone signals.

IV. Implementation

An offline implementation of the proposed method was realized in Matlab. The inputs to the method, following Fig. 1, are the spherical harmonic signals, the SH processing order for each sub-band for the SHD analysis/synthesis, and the loudspeaker directions. The microphone signals and the array geometry are provided if rendering above aliasing is needed. The following sections explain in more detail the implementation choices regarding the configuration of the method.

A. Time-Frequency Transform and Interpolation of Parameters

A 64-band uniform, complex-modulated QMF bank is applied for the time-frequency analysis, equipped with cascaded sub-subband filters in the lowest three frequency bands to obtain a 71 band resolution. The purpose of the secondary filters is to achieve at the lower range a frequency resolution similar to that of human spatial hearing. The filterbank has been earlier applied, for example in [52]–[54].

The processing is performed independently in each frequency band. The instantaneous covariance matrix is estimated first, based on which the directional analysis corresponding to the applied order of spherical harmonics is performed. The division to sectors, and the within sectors parametric analysis is performed as described in Sections III-B & III-C. The target covariance matrices are thus formulated also based on the instantaneous covariance data.

The instantaneous covariance matrices of the input signal, and the instantaneous target covariance matrices are consecutively averaged over a window of 64 QMF time indices, in half-overlapping frames every 32 QMF time-indices. These average covariance matrices are applied to formulate the least-squares mixing matrices as described in Section III-E. Such averaging is necessary to provide, in a stochastic sense, meaningful spatial sound processing, while preserving the sound quality by avoiding excessively fast changes in the resulting mixing coefficients. The mixing matrices are linearly interpolated between

the frames so that the center of the frame receives a non-interpolated mixing matrix.

The decorrelated signals are processed from the linearly-decoded signals of (29) by applying pseudo-random delays at each frequency band, ranging between approximately 20 and 80 milliseconds in the low range, transitioning to an interval of approximately 5 and 15 milliseconds in the high range. These intervals are a result of manual adjustment between sufficiently large delays, to avoid coloration when decorrelated and non-decorrelated sounds are mixed, and sufficiently short delays to avoid the artifact of added reverberation effect. A recursive onset suppressor is implemented prior to feeding the frequency band signals to the decorrelators, since typically the result of applying decorrelating operations to transients is detrimental for the perceived sound quality [29], [55].

B. Microphone Array Analysis and Order Selection

As mentioned earlier, for a real microphone array the method requires the SH signals $\tilde{\mathbf{x}}_N$, obtained from the microphone signals, and a specification of the SH processing order N_k for each sub-band, according to the array capabilities. Furthermore, if the high-frequency extension is employed, the method requires the microphone signals $\mathbf{x}_Q$ and an approximate spatial aliasing frequency limit to assign the two different processing modes at the respective ranges. Note that for the high-frequency extension, a description of the orientations of the microphones Ω_q is also needed. If the array is nearly-uniform, an approximate rule can be derived that determines a lower frequency limit for each order, based on a maximum tolerated noise amplification. Based on the assumption of an ideal uniform array and according to (8), the noise power amplification for order n is

$$G_n^2(k) = \frac{4\pi}{Q} \frac{1}{\left|b_n\left(\frac{\omega R}{c}\right)\right|^2}. \tag{40}$$

It is known that $20\log_{10}|b_n|$ for successive orders has at low frequencies a roll-off of $6n$ dB per octave [56], up to approximately $\omega R/c = n$. Since the response of (40) is linear in a log-log axis, it is approximated as a line corresponding to a power function of the form $y = \alpha x^p$, as can be seen in Fig. 3. The parameters α, p can be found by the logarithmic line and finally the lower frequency limit f_{low}^n for a certain order n with an array of Q microphones that results in a G_{db} noise amplification is given by

$$f_{\text{low}}^n(R, Q, G_{\text{db}}) = \frac{c}{2\pi R}\left(\frac{10^{\frac{G_{\text{db}}}{10}}\, Q\,|b_n(1)|^2}{4\pi}\right)^{\frac{-10\log_{10}(2)}{6n}}. \tag{41}$$

This practical rule requires only knowledge of the number of microphones and the array radius. If the array is not uniform or spherical, a more thorough theoretical or numerical analysis is needed to determine the frequency limits in which it supports each order without excessive noise amplification.

The implementation was tested with the Eigenmike spherical microphone array[1], consisting of $Q = 32$ microphones mounted on a rigid sphere with a radius of $R = 4.2$ cm. The sampling geometry is a truncated icosahedron one, with the microphones mounted on the faces. This sampling geometry can obtain up to fourth-order SH signals. Based on the approximate formula of $f_{\text{al}} = cN/(2\pi R)$ [41], spatial aliasing occurs at around $f_{\text{al}} \approx 5.5$ kHz for such an array. Based on these specifications, the approximate rule of (41) and a maximum noise amplification of $G_{\text{db}} = 10$ dB, the processing order for each QMF sub-band was set as presented in Table II. The noise amplification curves and the selected lower frequency limits per order are shown in Fig. 3. The range for the high-frequency extension was set slightly above the approximate aliasing frequency, at 6 kHz. QMF bands below 6 kHz of the SH signals $\tilde{\mathbf{x}}_N$ are processed according to the sector-based method of Sections III-B–III-E. The QMF bands above 6 kHz applied to the microphone signals $\mathbf{x}_Q$, are processed according to the high-frequency extension of Section III-F.

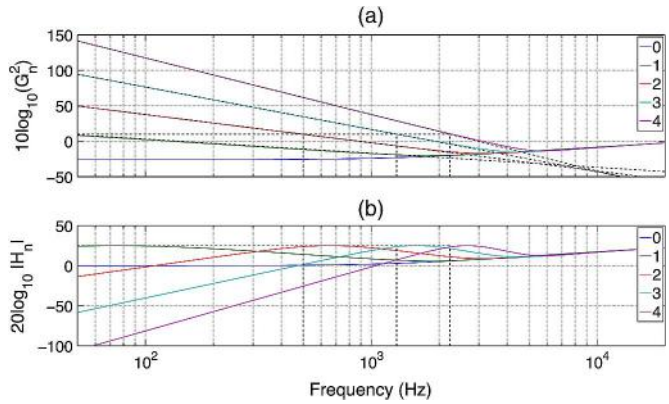

Fig. 3. (a) Noise amplification curves of simulated Eigenmike for orders n = 0–4, with the linear approximation curves in the $\omega R/c \leq n$ region, and the frequency limits for maximum amplification of 10 dB. (b) Regularized inverse filters for the realization of the SHT for orders n = 0–4, with the regularization set for maximum amplification of 10 dB.

TABLE II
LOW FREQUENCY LIMITS FOR HIGH-ORDER PROCESSING FOR THE EIGENMIKE AND RESPECTIVE QMF BANDS

Analysis order N	1	2	3	4	HF-ext.
Frequency limits	45Hz	500Hz	1300Hz	2230Hz	>6kHz
QMF bands	1-7	8-10	11-13	14-23	24-71

In order to evaluate the proposed order selection scheme, a rigid spherical microphone array simulator was implemented. The simulator was additionally used to simulate noisy array recordings from reference sound scenes for the listening tests of Section V. The directional response of ideal omnidirectional microphones on a rigid sphere were approximated as a series of Legendre polynomials P_n, frequency-weighted with the modal weights b_n [56]. For a simulation of the Eigenmike, the series was truncated at 30 terms without loss of accuracy. Array impulse responses for any direction of incidence were obtained in this way. Simulated microphone signals were then further transformed to SH signals by applying the summation of (5), followed by an inverse filter per order, realizing the inversion of the modal weights $1/b_n$. In order to avoid excessive amplification of the higher-order signals at low frequencies, Tikhonov regularization was used, set according to the same maximum noise amplification as in the order-selection scheme. The regularization value with respect to noise amplification was obtained as described in [6]. Example inverse filter responses are plotted

[1]http://www.mhacoustics.com/products#eigenmike1

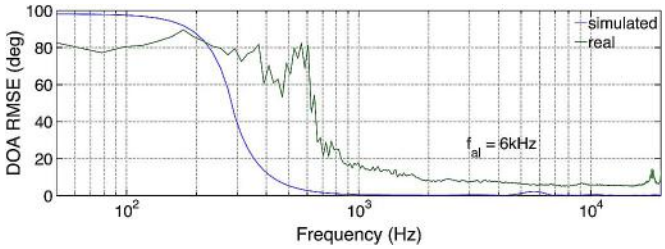

Fig. 4. Mean directional error for HF-extension.

in Fig. 3(b) for $G_{\mathrm{db}} = 10$ dB. It is clear that the filters are in accordance with the selected lower frequency limits of each order of Fig. 3(a) and Table II, since they equalize the responses down to that frequency and then they decay again. Note that the true amplification of the filters reaches $10\log_{10} Q + G_{\mathrm{db}} \approx 25$ dB due to the regularization taking into account the improvement of SNR by the number of microphones.

In the case of the real Eigenmike array, instead of using a theoretical model, the SHT was realized by FIR filters obtained by a least-squares regularized inversion of free-field measurements around the array, similar to [6], [57]. A measurement-based SHT compensates for any unidealitites and deviations of the real array from the model. It has been shown that performance in the SHT close to the theoretical one can be reached in this manner [6], [57].

C. High-Frequency Extension Evaluation

The HF analysis scheme presented in Section III-F was evaluated for the Eigenmike, both for a simulated ideal array and for a real device, based on a dense grid of free-field response measurements. Firstly, the performance of the DOA estimator $\mathbf{r}_{\mathrm{HF}}$ was evaluated for a single plane wave incident from Φ_{HF}. Since estimation errors can be direction-dependent, a spherical grid of 162 DOAs was generated by subdividing the faces of an icosahedron and the error between the true and estimated DOA $\arccos(\mathbf{n}(\Phi_{\mathrm{HF}})\cdot\mathbf{n}(\hat{\Phi}_{\mathrm{HF}}))$ was computed. The root-mean square error (RMSE) across all directions is plotted in Fig. 4. It is evident that in this basic case, the error above aliasing is close to zero for the ideal array, and for the real one is less than 10° which is deemed acceptable for the high-frequency range.

Secondly, the performance of the diffuseness estimator of (39) was evaluated. A diffuse field was simulated as 162 incoherent plane waves of gaussian noise of 1 second. A direct component was simulated as an additional plane wave of noise incident from the front. By adjusting their relative power, an ideal diffuseness value $\psi = S_{\mathrm{diff}}/(S_{\mathrm{dir}} + S_{\mathrm{diff}})$ could be set. The error between true and estimated diffuseness $\psi - \hat{\psi}$ was computed and averaged for 50 realizations. The results for 5 diffuseness values are plotted in Fig. 5. The estimator performs well above aliasing for both the simulated and the real case, with the error being less than 0.2 for all cases.

V. Listening Experiments

Two listening experiments were organized to evaluate the perceptual benefits provided by the proposed method. The first test assumed ideal SH signals up to fourth order at all frequencies. This scenario is practical only in the case that the method is applied to synthetic and virtual sound scenes, where encoding

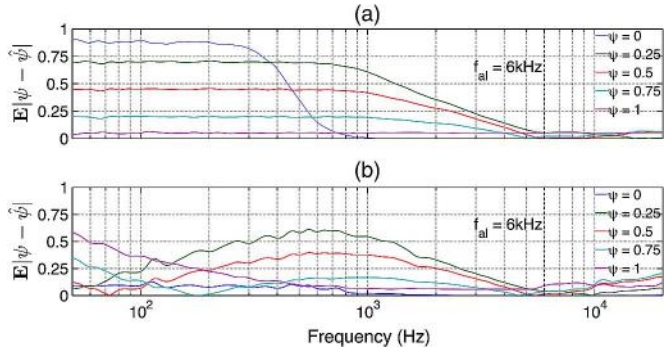

Fig. 5. Diffuseness error for HF-extension: (a) simulated ideal eigenmike, (b) real measured eigenmike.

TABLE III
Layout of the Applied 29-Channel Loudspeaker Configuration

Azimuth	Elevation	Azimuth	Elevation
0°	0°	0°	22°
±15°	0°	±30°	±22°
±30°	0°	0°	±45°
±45°	0°	±90°	±45°
±90°	0°	180°	±45°
±105°	0°	0°	90°
±135°	0°		
180°	0°		

of sound material into ideal components of an arbitrary order is possible. The main condition tested in this case was the effect of utilizing the proposed higher-order processing compared to a similarly ideal first-order processing. A second condition tested was if there exists any advantage of using a parametric technique over direct linear reproduction such as the ambisonic decoding of Section III-D, even at orders as high as four.

The second test considered a practical scenario of a spherical microphone array subject to microphone self noise and spatial aliasing. This condition tested the performance of applying the method to a recording with a real-world array, in which case the parametric processing with frequency-dependent order was employed. Furthermore, the high-frequency extension for frequencies above the aliasing limit of the array was used. The processing order for each frequency band for the sector-based method and the band limits for switching to the high-frequency extension were set according to the array analysis of Section IV-B. Comparisons were both with respect to the first-order DirAC technique and with respect to linear reproduction using the same signals. The listening tests were conducted in an anechoic chamber using 28 loudspeakers in a sphere, with positions as listed in Table III. Such a loudspeaker configuration resembles the recent 22.2 surround layout described in [2]. A more detailed description of how the different reproduction modes were generated is given in Section V-B.

A. Reference Sound Scenes

Five synthetic sound scenes were generated, listed in Table IV, for the 28-loudspeaker configuration. The *mix* scenes contained three sources in the horizontal plane, and one source above the listener. The *music* scenes contained four instruments in the horizontal plane in the frontal arc between ±60°. The

TABLE IV
REFERENCE SOUND SCENES

Identifier	Description
Speech_hall	Female speech in front in a large hall
Music_freefield	Four instruments in a free field
Music_room	Four instruments in a venue-sized room
Mix_freefield	Female speech, fountain, piano and claps/ free field
Mix_room	Female speech, fountain, piano and claps/ small room

TABLE V
REPRODUCTION MODES UNDER TEST

Identifier	Description
Ref.	Original 29-channel reference sound
P_full	Parametric higher order reproduction (proposed method)
P_1st	Parametric first order reproduction
L_full	Linear higher order reproduction
L_1st	Linear first order reproduction

original recordings in all cases were anechoic. The room effect was generated using the image source method. The angles of the image sources were quantized to the nearest loudspeaker in the applied layout of 28 loudspeakers. The design of the program material involving simultaneous sources and prominent specular reflections was known *a priori* to stress especially the low-order model-based parametric rendering systems. Simplified cases, such as fully diffuse sound fields, or a single source in a free field, were not included in the test, since these were known to be well reproduced with the parametric techniques regardless of the processing order.

B. Reproduction Modes

Both listening tests had the same reference scenes of Table IV. In the first test, fourth-order noiseless ideal SH signals were generated directly by encoding the reference signals into up to fourth-order SHs. According to the formulation of Section II-B, these signals would correspond to a spatially band-limited sound scene captured in the vector $\mathbf{a}_4$. Ideal B-format signals were additionally generated by multiplying the first-order SH signals $\mathbf{a}_1$ with the coefficients of (12). The SH signal vector was processed with the proposed method set to fourth-order processing at all QMF bands (*P_full*), and by the ambisonic decoding outlined in Section III-D (*L_full*), using all 28 loudspeakers. B-format was processed with first-order DirAC (*P_1st*) and, additionally, with a first-order ambisonic (*L_1st*) decoding included as a low-quality anchor, using a quasi-regular subset of 12 loudspeakers. This subset was a compromise between the minimum of 4 loudspeakers, which is too sparse for consistent localization, and higher numbers, which are known to induce spectral coloration [33].

In the second test, microphone signals capturing the sound scenes were generated according to the array simulation described in Section IV-B. According to the specifications of the Eigenmike, with an equivalent input noise level of 15 dBA, and assuming a moderate level for the sound scenes of 60 dBA, we added gaussian noise to each microphone signal corresponding to an SNR of 45 dB. The signal power of an omnidirectional reference encoding of the sound scenes was used to set the noise power. The result corresponds to the noisy microphone signal vector $\mathbf{x}_Q$. The fourth-order noisy SH signals $\tilde{\mathbf{x}}_4$ were obtained by applying the regularized SHT described in Section IV-B. Note that the SH signals in this case were also spatially aliased above approximately $f_{\mathrm{al}} = 5.5$ kHz. Noisy and aliased B-format signals were generated in the same way as in the ideal case. Finally, the noisy SH and B-format signals were processed again with the proposed parametric method (*P_full*), fourth-order ambisonic decoding (*L_full*), first-order DirAC (*P_1st*) and first-order ambisonics (*L_1st*). However, in this case the sector-based processing was operating at different orders at each QMF band according to the specification of Table II. Furthermore, the microphone signals and the high-frequency extension of Section III-F was used for bands above 6 kHz. The original 28-channel signal was included to the test set as a hidden reference. The reproduction modes under test are summarized in Table V.

C. Test Setting

Ten subjects participated to both tests, all of which researchers in the field of audio and not authors of this paper. The listeners rated the similarity of the reproduction modes with respect to a known reference using a graphical interface through a touch screen display. The reproduction modes were presented in random order. The scores were given using sliders without intermediate labels. The top of the scale indicated that the item is indistinguishable from the reference, and scores towards the low denoted an increasing perceived difference. The subjects were instructed to use the scale broadly. The subjects were allowed to rotate in the chair but not to move away from it while listening.

D. Results

A two-way repeated measures analysis of variance (RM-ANOVA) was applied to the result data of both tests, with factors *Reference_scene* and *Reproduction_mode*. The analysis with both tests provided the following results: Significant effects were found with factor *Reproduction_mode* and with the interaction *Reproduction_mode*Reference_scene*. The means and the 95% confidence intervals of factor *Reproduction_mode* are shown for the two tests at Figs. 6 and 7. in both tests, all means differed significantly from each other, except the first order parametric reproduction with respect to the higher order linear reproduction.

E. Discussion

The following observations can be made on the results:

a) The higher-order SH signals can be exploited to clearly improve the performance regarding reproduction accuracy in the parametric rendering scheme.
b) Using ideal fourth-order SH signals the proposed method achieves almost perceptually transparent results for all scenarios.
c) For both low and high orders, and with both idealized conditions as well as using an actual microphone array, the parametric technique was perceived closer to the reference than the linear technique.

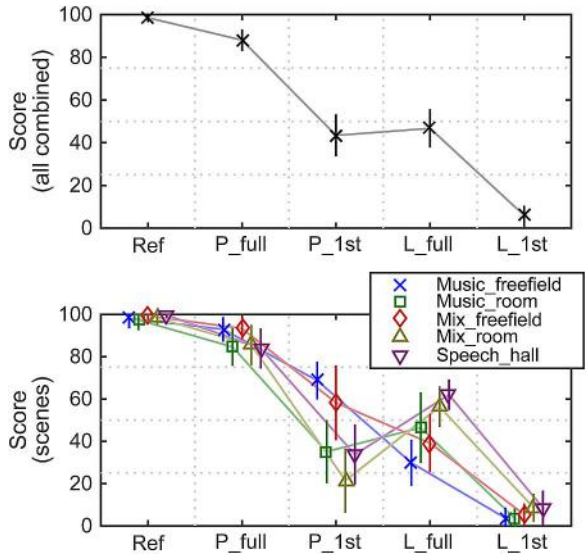

Fig. 6. Means and 95% confidence intervals of the first listening test with idealized reproduction assuming up to fourth order of available spherical harmonics. P denotes parametric reproduction and L denotes linear reproduction.

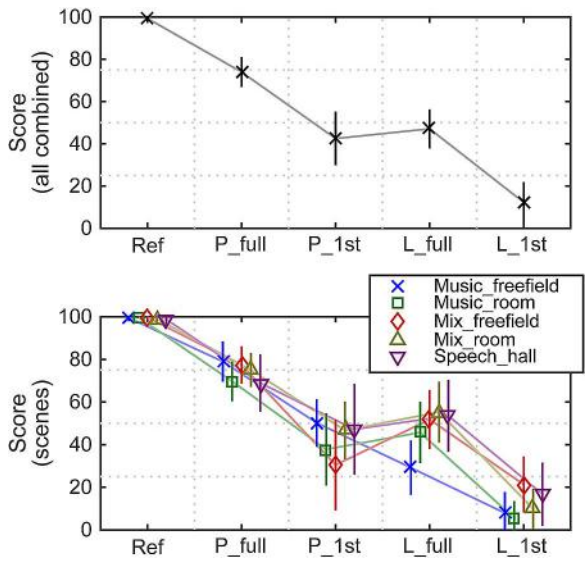

Fig. 7. Means and 95% confidence intervals of the second listening test assuming an actual microphone array. P denotes parametric reproduction and L denotes linear reproduction.

Concluding the results, at least for dense loudspeaker setups, it is perceptually beneficial to capture higher-order SH signals and to apply a parametric method such as the proposed method to process the loudspeaker signals.

VI. Conclusion

This work presents a novel method for high-quality parametric reproduction of sound scenes captured with a small-sized microphone array. The method is formulated in the spherical harmonic domain and is scalable in the sense of using the higher-order signals only in frequency bands that support them. A minimum performance of first-order rendering is guaranteed, similar to least-squares optimized DirAC rendering, which has previously been shown to offer good perceptual quality in most cases. In addition, the frequency range above the spatial aliasing limit of the array, which is usually neglected, is processed with a similar parametric approach offering a performance equal to first-order rendering.

The analysis of intensity and diffuseness, previously utilized in parametric spatial audio coding, is extended to higher-orders by segregating the sphere into an order-dependent number of sectors and then performing the analysis for each one of them. This approach extracts multiple local spatial parameters instead of single global ones. The multiple parameters permit panning of the directional content in the recording to multiple directions simultaneously, reduced use of decorrelation, and diffuse rendering with a directional distribution, offering improved perceptual reproduction of the spatial properties of the recording, compared to first-order processing.

Based on a listening test with reference sound scenes, the performance of the method was evaluated against two cases. The first was assessing the improvement of the higher-order model, compared to a first-order processing scheme similar to DirAC. The second was assessing the improvement against a state-of-the-art linear high-order rendering such as an optimized ambisonic decoding. In both cases, the present method was judged perceptually closer to the reference by the listeners. Furthermore, the method applied to idealized noiseless recordings achieved close to transparent results, while operating on realistic simulated array recordings the performance was only slightly degraded. Hence, it is concluded that if the array supports high-order recordings, the proposed parametric method is beneficial in all cases.

Appendix
Proof of Energy Preservation of Sector Patterns

Let us assume J axisymmetric real beam patterns $c_j(\Omega)$ of order N oriented at the vertices of a spherical t-design of $t = 2N$. Axisymmetric patterns can be described by $N+1$ SH coefficients $\mathbf{c}_N$ with their pattern given by

$$c(\alpha) = \sum_{n=0}^{N} \sqrt{\frac{2n+1}{4\pi}} c_n P_n(\cos\alpha) \tag{42}$$

where P_n is the Legendre polynomial of degree n and $\cos\alpha = \mathbf{n}^T(\Omega)\cdot\mathbf{n}(\Omega_j)$ is the cosine of the angle between the DOA and the beam's orientation. The squared pattern $d(\alpha) = c^2(\alpha)$ is also axisymmetric and a polynomial of degree $2N$, thus exactly integrated by a spherical $2N$-design as in (23). Furthermore, due to the orthogonality of SH, the integral of SH of order $n \le 2N$ is

$$\int_\Omega Y_{nm}(\Omega)\,\mathrm{d}\Omega = \frac{4\pi}{J}\sum_{j=0}^{J} Y_{nm}(\Omega_j) = \begin{cases} 0, & \text{for } n \neq 0 \\ \sqrt{4\pi}, & \text{for } n = 0 \end{cases}. \tag{43}$$

Based on the above relations, the condition of (22) can be written as

$$\begin{aligned}\sum_{j=0}^{J} c_j^2(\Omega) &= \sum_{j=0}^{J} c^2(\alpha_j) = \sum_{j=0}^{J} d(\alpha_j) \\ &= \sum_{j=0}^{J}\sum_{n=0}^{2N} \sqrt{\frac{2n+1}{4\pi}} d_n P_n(\cos\alpha_j) \\ &= \sum_{j=0}^{J}\sum_{n=0}^{2N} \sqrt{\frac{4\pi}{2n+1}} d_n \sum_{m=-n}^{n} Y_{nm}(\Omega_j) Y_{nm}^*(\Omega) \\ &= \sum_{n=0}^{2N} \sqrt{\frac{4\pi}{2n+1}} d_n \sum_{m=-n}^{n} Y_{nm}^*(\Omega) \sum_{j=0}^{J} Y_{nm}(\Omega_j) \\ &= d_0 \frac{J}{\sqrt{4\pi}} = \frac{J}{Q_{\mathrm{c}}} \end{aligned} \tag{44}$$

where the spherical harmonics addition theorem was used in the third line and relation (43) was used in the last line. The quantity Q_{c} corresponds to the directivity factor of the pattern $Q_{\mathrm{c}} = (4\pi)/\int_{\Omega} c^2(\Omega)\mathrm{d}\Omega = (\sqrt{4\pi})/d_0$ and can be directly computed from the coefficients of the pattern as

$$Q_{\mathrm{c}} = \frac{4\pi}{\mathbf{c}_N^H \mathbf{c}_N}. \tag{45}$$

REFERENCES

[1] G. Theile, "Multichannel natural recording based on psychoacoustic principles," *Audio Eng. Soc. Conv. 108*, 2000.

[2] K. Hamasaki, K. Hiyama, and R. Okumura, "The 22.2 multichannel sound system and its application," *Audio Eng. Soc. Conv. 118*, 2005.

[3] E. De Sena, H. Hacihabiboglu, and Z. Cvetkovic, "Analysis and design of multichannel systems for perceptual sound field reconstruction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 8, pp. 1653–1665, Aug. 2013.

[4] M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2–10, 1973.

[5] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *J. Audio Eng. Soc.*, vol. 53, no. 11, pp. 1004–1025, 2005.

[6] S. Moreau, S. Bertet, and J. Daniel, "3D sound field recording with higher order ambisonics—Objective measurements and validation of spherical microphone," *Audio Eng. Soc. Conv. 120*, 2006.

[7] F. Zotter and M. Frank, "All-round ambisonic panning and decoding," *J. Audio Eng. Soc.*, vol. 60, no. 10, pp. 807–820, 2012.

[8] J. Backman, "Microphone array beam forming for multichannel recording," *Audio Eng. Soc. Conv. 114*, 2003.

[9] J. S. Abel, Y. Hur, Y.-C. Park, and D. H. Youn, "A set of microphone array beamformers implementing a constant-amplitude panning law," *Audio Eng. Soc. Conv. 129*, 2010.

[10] H. Hacihabiboglu and Z. Cvetkovic, "Panoramic recording and reproduction of multichannel audio using a circular microphone array," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'09)*, 2009, pp. 117–120.

[11] K. Ono, T. Nishiguchi, K. Matsui, and K. Hamasaki, "Portable spherical microphone for Super Hi-Vision 22.2 multichannel audio," *Audio Eng. Soc. Conv. 135*, 2013.

[12] A. Farina, A. Amendola, L. Chiesi, A. Capra, and S. Campanini, "Spatial PCM sampling: A new method for sound recording and playback," in *Proc. 52nd Int. Conf. Audio Eng. Soc.*, 2013.

[13] E. Gallo and N. Tsingos, "Extracting and re-rendering structured auditory scenes from field recordings," in *Proc. 30th Int. Conf. Audio Eng. Soc.*, 2007.

[14] C. Faller, "Microphone front-ends for spatial audio coders," *Audio Eng. Soc. Conv. 125*, 2008.

[15] B. Gunel, H. Hacihabiboglu, and A. M. Kondoz, "Acoustic source separation of convolutive mixtures based on intensity vector statistics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 748–756, May 2008.

[16] K. Niwa, T. Nishino, and K. Takeda, "Encoding large array signals into a 3D sound field representation for selective listening point audio based on blind source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'08)*, 2008, pp. 181–184.

[17] S. Berge and N. Barrett, "A new method for B-format to binaural transcoding," in *Proc. 40th Int. Conf. Audio Eng. Soc.*, 2010.

[18] M. Cobos, J. J. Lopez, and S. Spors, "A sparsity-based approach to 3D binaural sound synthesis using time-frequency array processing," *EURASIP J. Adv. Signal Process.*, vol. 2010, pp. 2:1–2:13, 2010.

[19] C. Verron, P.-A. Gauthier, J. Langlois, and C. Guastavino, "Spectral and spatial multichannel analysis/synthesis of interior aircraft sounds," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1317–1329, Jul. 2013.

[20] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 3, pp. 727–739, Mar. 2014.

[21] V. Pulkki, "Spatial sound reproduction with Directional Audio Coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, 2007.

[22] J. Breebaart and C. Faller, *Spatial Audio Processing: MPEG Surround and Other Applications*. Chichester, U.K.: Wiley, 2007.

[23] I. Elfitri, B. Gunel, and A. M. Kondoz, "Multichannel audio coding based on analysis by synthesis," *Proc. IEEE*, vol. 99, no. 4, pp. 657–670, Apr. 2011.

[24] J. Ahonen, V. Pulkki, F. Küch, G. Del Galdo, M. Kallinger, and R. Schultz-Amling, "Directional audio coding with stereo microphone input," *Audio Eng. Soc. Conv. 126*, 2009.

[25] O. Thiergart, M. Kratschmer, M. Kallinger, and G. Del Galdo, "Parameter estimation in directional audio coding using linear microphone arrays," *Audio Eng. Soc. Conv. 130*, 2011.

[26] A. Politis, M.-V. Laitinen, J. Ahonen, and V. Pulkki, "Parametric spatial audio coding for spaced microphone array recordings," *Audio Eng. Soc. Conv. 134*, 2013.

[27] M.-V. Laitinen and V. Pulkki, "Binaural reproduction for directional audio coding," in *IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'09)*, 2009, pp. 337–340.

[28] J. Vilkamo, T. Lokki, and V. Pulkki, "Directional audio coding: Virtual microphone-based synthesis and subjective evaluation," *J. Audio Eng. Soc.*, vol. 57, no. 9, pp. 709–724, 2009.

[29] M.-V. Laitinen, F. Küch, S. Disch, and V. Pulkki, "Reproducing applause-type signals with directional audio coding," *J. Audio Eng. Soc.*, vol. 59, no. 1/2, pp. 29–43, 2011.

[30] M.-V. Laitinen and V. Pulkki, "Utilizing instantaneous direct-to-reverberant ratio in parametric spatial audio coding," *Audio Eng. Soc. Conv. 133*, 2012.

[31] J. Vilkamo, T. Bäckström, and A. Kuntz, "Optimized covariance domain framework for time-frequency processing of spatial audio," *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 403–411, 2013.

[32] A. Laborie, R. Bruno, and S. Montoya, "A new comprehensive approach of surround sound recording," *Audio Eng. Soc. Conv. 114*, 2003.

[33] A. Solvang, "Spectral impairment of two-dimensional higher order ambisonics," *J. Audio Eng. Soc.*, vol. 56, no. 4, pp. 267–279, 2008.

[34] S. Bertet, J. Daniel, L. Gros, E. Parizet, and O. Warusfel, "Investigation of the perceived spatial resolution of higher order ambisonics sound fields: A subjective evaluation involving virtual and real 3D microphones," in *Proc. 30th Int. Conf. Audio Eng. Soc.*, 2007.

[35] S. Braun and M. Frank, "Localization of 3D ambisonic recordings and ambisonic virtual sources," in *Proc. Int. Conf. Spatial Audio (ICSA)*, 2011.

[36] A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, and B. Rafaely, "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution," *J.. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. 2711–2721, 2013.

[37] A. Alexandridis, A. Griffin, and A. Mouchtaris, "Capturing and reproducing spatial audio based on a circular microphone array," *J. Elect. Comput. Eng.*, vol. 2013, pp. 1–16, 2013.

[38] A. Wabnitz, N. Epain, A. McEwan, and C. Jin, "Upscaling ambisonic sound scenes using compressed sensing techniques," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'11)*, 2011.

[39] N. Epain and C. T. Jin, "Super-resolution sound field imaging with sub-space pre-processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'13)*, 2013, pp. 350–354.

[40] F. Zotter, "Sampling strategies for acoustic holography/holophony on the sphere," in *Proc. NAG/DAGA Int. Conf. Acoust.*, Rotterdam, The Netherlands, 2009.
[41] B. Rafaely, B. Weiss, and E. Bachmat, "Spatial aliasing in spherical microphone arrays," *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 1003–1010, Mar. 2007.
[42] R. H. Hardin and N. J. Sloane, "McLaren's improved snub cube and other new spherical designs in three dimensions," *Discrete Comput. Geom.*, vol. 15, no. 4, pp. 429–441, 1996.
[43] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan. 2005.
[44] M. A. Gerzon, "The design of precisely coincident microphone arrays for stereo and surround sound," *Audio Eng. Soc. Conv. 50*, 1975.
[45] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'02)*, 2002, vol. 2, pp. II–1781.
[46] G. Del Galdo, M. Taseska, O. Thiergart, J. Ahonen, and V. Pulkki, "The diffuse sound field in energetic analysis," *J. Acoust. Soc. Amer.*, vol. 131, no. 3, pp. 2141–2151, 2012.
[47] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, 1997.
[48] J. Vilkamo and V. Pulkki, "Minimization of decorrelator artifacts in directional audio coding by covariance domain rendering," *J. Audio Eng. Soc.*, vol. 61, no. 9, pp. 637–646, 2013.
[49] A. Politis and V. Pulkki, "Broadband analysis and synthesis for directional audio coding using A-format input signals," *Audio Eng. Soc. Conv. 131*, 2011.
[50] J. Ahonen and V. Pulkki, "Diffuseness estimation using temporal variation of intensity vectors," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'09)*, 2009, pp. 285–288.
[51] M. Kratschmer, O. Thiergart, and V. Pulkki, "Envelope-based spatial parameter estimation in Directional Audio Coding," *Audio Eng. Soc. Conv. 133*, 2012.
[52] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegard, "Low complexity parametric stereo coding," *Audio Eng. Soc. Conv. 116*, 2004.
[53] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, and H. Purnhagen *et al.*, "MPEG Surround—The ISO/MPEG standard for efficient and compatible multichannel audio coding," *J. Audio Eng. Soc.*, vol. 56, no. 11, pp. 932–955, 2008.
[54] J. Breebaart, J. Engdegârd, C. Falch, O. Hellmuth, J. Hilpert, and A. Hoelzer *et al.*, "Spatial audio object coding (SAOC)-the upcoming MPEG standard on parametric object based audio coding," *Audio Eng. Soc. Conv. 124*, 2008.
[55] A. Kuntz, S. Disch, T. Bäckström, and J. Robilliard, "The transient steering decorrelator tool in the upcoming MPEG unified speech and audio coding standard," *Audio Eng. Soc. Conv. 131*, 2011.
[56] J. Meyer and G. W. Elko, "Spherical microphone arrays for 3D sound recording," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*. New York, NY, USA: Springer, 2004, pp. 67–89.
[57] C. T. Jin, N. Epain, and A. Parthy, "Design, optimization and evaluation of a dual-radius spherical microphone array," *IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP)*, vol. 22, no. 1, pp. 193–204, Jan. 2014.

Archontis Politis obtained his M.Eng. degree in civil engineering at Aristotle's University of Thessaloniki, Greece, and his M.Sc. degree in sound & vibration studies at ISVR, University of Southampton, UK, in 2006 and 2008 respectively. Currently, he is pursuing a doctoral degree in the field of parametric spatial sound recording and reproduction. From 2008 to 2010 he was a Graduate Acoustic Consultant at Arup Acoustics, Glasgow, UK, and as a Researcher in a joint collaboration between Arup Acoustics and the Glasgow School of Arts, on interactive auralization of architectural spaces using 3D sound techniques.

Juha Vilkamo received his M.Sc. degree in 2008 and D.Sc. degree in 2014 at the Aalto University (formerly Helsinki University of Technology), Finland. In between the academic studies from 2008 until 2011 he worked as a Researcher at Fraunhofer IIS, Germany, in the fields of binaural technologies and spatial sound reproduction. His research has lead to audio signal processing techniques that have been applied in several products and also in the MPEG-H 3D audio standard. He has contributed to the fields of parametric time–frequency audio signal processing, adaptive microphone techniques, artificial reverberation, multi-channel downmixing and upmixing, and spatial sound enhancement.

Ville Pulkki received his M.Sc. and D.Sc. (Tech) degrees from Helsinki University of Technology in 1994 and 2001, respectively. He majored in acoustics, audio signal processing and information sciences. Between 94 and 97 he was a full-time student at the Department of Musical Education in Sibelius Academy.

In his doctoral dissertation he developed vector base amplitude panning (VBAP), which is a method for positioning virtual sources to multi- channel loudspeaker configurations. In addition, he studied the performance of VBAP with psychoacoustic listening tests and with modeling of auditory localization mechanisms. The VBAP method is now widely used in multi-channel virtual auditory environments, and in computer music installations. Later, he developed with his group a non-linear time–frequency-domain method for spatial sound reproduction and coding, Directional Audio Coding (DirAC). DirAC takes coincident first-order microphone signals as input, and processes output to arbitrary loudspeaker layouts or to headphones. He also researches computational functional model of the brain organs devoted to binaural hearing. He is leading a research group in Aalto University (earlier: Helsinki University of Technology, TKK or HUT), which consists of 15 researchers. The group conducts research also on head-related acoustics measurements, and conducts psychoacoustical experiments to better understand spatial sound perception.