Politis, Archontis; Tervo, Sakari; Pulkki, Ville

COMPASS: Coding and Multidirectional Parameterization of Ambisonic Sound Scenes

Published in:
2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

DOI:
10.1109/ICASSP.2018.8462608

Published: 16/04/2018

Document Version
Peer reviewed version

COMPASS: CODING AND MULTIDIRECTIONAL PARAMETERIZATION OF AMBISONIC SOUND SCENES

Archontis Politis, Sakari Tervo, Ville Pulkki

Department of Signal Processing and Acoustics
School of Electrical Engineering
Aalto University, 02150 Espoo, Finland

ABSTRACT

Current methods for immersive playback of spatial sound content aim at flexibility in terms of encoding and decoding, abstracting the two from the recording or playback setup. Ambisonics constitutes such a method, that is however signal-independent, and at low spatial resolutions fails to provide appropriate spatialization cues to the listener, with potential severe colouration effects and localization ambiguity. We present a new signal-dependent method for parametric analysis and synthesis of ambisonic sound scenes that takes advantage of the flexibility of Ambisonics as a spatial audio format, while improving reproduction. The proposed approach considers a more general acoustic model than previous proposals, with multiple source signals and a non isotropic ambient component. According to a listening test using headphones, the method is perceived closer to binaural reference sound scenes than ambisonic playback.

Index Terms— spatial audio, acoustic scene analysis, audio coding, Ambisonics

1. INTRODUCTION

Modern spatial sound recording, processing and reproduction is moving away from playback-setup based channel formats, to systems that are flexible and able to distribute appropriately the spatial sound recording to arbitrary setups. A popular such approach is the one popularized under the name Ambisonics [1, 2, 3], which uses spherical harmonics (SH) as spatial basis functions to represent the sound scene. Ambisonics is essentially a signal-independent method, defining a linear encoding and decoding stage that takes into account only the properties of the capturing or reproduction setup. The perceptual performance of ambisonic decoding does not relate necessarily to the complexity of the sound scene, but to the number of ambisonic channels, related to a spatial resolution known as ambisonic order. For low-order reproduction, localization of directional sounds can be vague, audible coloration may occur at higher frequencies, and reverberant sound in the recording can have degraded spaciousness [4, 5, 6, 7]. Many of these issues can be alleviated by using higher-order Ambisonics (HOA). This comes however at a cost of increased bandwidth, since the number of channels rise quadratically with order, and in the case of recorded sound scenes, practical recording setups are still limited to between first- and third-order recording of ambisonic signals.

To improve upon the limitations of low-order Ambisonics, certain signal-dependent methods have been developed, all operating in the time-frequency domain and differing on their assumed soundfield model and estimation of parameters. The most prominent of them is Directional Audio Coding (DirAC) [8, 9], which in its basic incarnation improves first-order ambisonic (FOA) playback by extracting one direction-of-arrival (DoA) and one diffuseness parameter, splitting essentially the sound scene into a single source stream and an isotropic diffuse stream, reproduced then via loudspeakers or headphones. Another method limited to FOA signals is HARPEX [10], which estimates DoAs and amplitudes of two time-varying plane wave components, without diffuse sound, and hence renders two directional streams. These methods have been found effective in a variety of sound scenarios, allowing additionally useful flexible spatial modifications of the scene [11] and upmixing from FOA to HOA. In the case that HOA signals are available, the authors generalized DirAC to multiple source and diffuse streams, by segmenting the sound scene into spatially separated sectors and estimating the DirAC parameters for each one of them, improving parametric playback in cases where FOA analysis was challenging. An alternative approach is based on directional sparsity of source signals; using then iterative sparse recovery methods, multiple sharply localized source streams can be extracted from the ambisonic signals, and subsequently rendered or upmixed [12].

We present a new approach to analysis and synthesis of ambisonic signals, termed COding and Multidirectional Parameterization of Ambisonic Sound Scenes (COMPASS). Contrary to the previous methods, it uses a general acoustic model of the sound scene of multiple foreground source signals and a background ambient component that is not necessarily isotropic or diffuse. The method relies on the general subspace principle of array processing for estimation, and spatial filtering for synthesis. Contrary to [10] it is not limited only to FOA, and it is more appropriate for spatial modification of the sound scene components than [9], since the higher-order DirAC streams do not necessarily correspond to actual source components in the scene. Compared to the sparse recovery approach of [12], COMPASS does not require preprocessing in a non-sparse scenario, and it is computationally efficient, able to operate in real-time which is especially important for headphone playback where head-tracking may be employed. COMPASS is conceptually closer to array processing for speech enhancement [13, 14], however it operates on ambisonic signals instead of microphone signals, and aims to preserve all scene components without rejecting interferers, ambience and reverberation. Furthermore, due to the generality of the SH signal format, it can be applied both to ambisonic audio generated from mixing software, and to spatial sound recordings.

2. AMBISONIC PROCESSING

Assuming that all sound sources are on the far-field, a general sound scene can be described as a continuous distribution of plane waves with spatio-temporal amplitude $a(t, \boldsymbol{\gamma})$ for a plane wave incident

from DoA $\boldsymbol{\gamma} = [\cos\phi\cos\theta, \sin\phi\cos\theta, \sin\theta]^{\mathrm{T}}$, with (ϕ, θ) being azimuth and elevation angle respectively. Applying the spherical harmonic transform (SHT) on the amplitude density, we get the SH coefficients of the sound field $\mathbf{a}$, or equivalently, ambisonic signals

$$\mathbf{a}(t) = \mathcal{SHT}\{a(t, \boldsymbol{\gamma})\} = \int_{\boldsymbol{\gamma}} a(t, \boldsymbol{\gamma})\mathbf{y}(\boldsymbol{\gamma})\,\mathrm{d}\boldsymbol{\gamma}, \tag{1}$$

where $\int_{\boldsymbol{\gamma}}\,\mathrm{d}\boldsymbol{\gamma}$ denotes integration over the surface of the unit sphere, and $\mathrm{d}\boldsymbol{\gamma} = \cos\theta\mathrm{d}\theta\mathrm{d}\phi$. In practice, the basis vector $\mathbf{y}(\boldsymbol{\gamma})$ and signal vector $\mathbf{a}$ contain all the SHs and signals up to a specified maximum order N. For a SHT of order N, there are $M = (N+1)^2$ SHs and ambisonic signals. Following established ambisonic conventions, real orthonormal SHs are used herein, with the 0th order ambisonic signal $[\mathbf{a}]_1$ being equivalent to an omnidirectional (pressure) signal at the origin.

Conventional signal-independent ambisonic processing can be described by three linear matrices

$$\mathbf{z}(t) = \mathbf{DTE}\mathbf{s}(t) = \mathbf{DT}\mathbf{a}(t), \tag{2}$$

where $\mathbf{s}(t) = [s_1(t), ..., s_K(t)]^{\mathrm{T}}$ describes either a set of K source signals to be spatialized and encoded, or microphone signals to be encoded, $\mathbf{E}$ describes the $M \times K$ encoding matrix producing the SH signals, $\mathbf{T}$ is an $M \times M$ optional linear transformation matrix for spatial modifications of the sound scene. Finally, $\mathbf{D}$ is the $L \times M$ decoding matrix producing the L headphone or loudspeaker signals $\mathbf{z}$. It should be noted that in the simplest case of synthetic sound scene encoding, and loudspeaker decoding, the mixing matrices $\mathbf{E}, \mathbf{D}$ are frequency-independent, while in case of microphone encoding, or headphone playback, they are frequency-dependent in order to accommodate microphone array and head-related transfer function (HRTF) information respectively, hence matrix multiplications in (2) should be replaced with filter-and-sum operations. For a concise description of common ambisonic definitions for the above matrices, refer to [15]. An ambisonic operation fundamental to this work is encoding of a set of K plane wave source signals $\mathbf{s}$, incident from $\boldsymbol{\Gamma}_{\mathrm{s}} = [\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_K]$

$$\mathbf{a}(t) = \sum_{k=1}^{K} s_k(t)\mathbf{y}(\boldsymbol{\gamma}_k) = \mathbf{Y}_{\mathrm{s}}\mathbf{s}(t), \tag{3}$$

with $\mathbf{Y}_{\mathrm{s}} = [\mathbf{y}(\boldsymbol{\gamma}_1), ..., \mathbf{y}(\boldsymbol{\gamma}_K)]^{\mathrm{T}}$. Furthermore, ambisonic decoding matrices, possibly frequency-dependent, are used herein as

$$\mathbf{D}(f) = (1/V)\mathbf{G}_{\mathrm{vls}}(f)\mathbf{Y}_{\mathrm{vls}}^{\mathrm{T}}, \tag{4}$$

where $\mathbf{Y}_{\mathrm{vls}}$ is the $M \times V$ SH matrix for V directions of a uniformly-distributed set of virtual loudspeakers, such as a spherical t-design of $t > 2N + 1$ [16], and $\mathbf{G}_{\mathrm{vls}} = [\mathbf{g}(\boldsymbol{\gamma}_1), ..., \mathbf{g}(\boldsymbol{\gamma}_V)]$ is an $L \times V$ matrix of spatialization gains, such as amplitude panning gains for loudspeakers, or HRTFs $\mathbf{g}(\boldsymbol{\gamma}, f) = [h_{\mathrm{L}}(\boldsymbol{\gamma}, f),\, h_{\mathrm{R}}(\boldsymbol{\gamma}, f)]^{\mathrm{T}}$ for headphones. The above decoding approach corresponds to the ALLRAD method proposed in [3].

3. METHOD

The COMPASS method, depicted in Fig. 1, is based on the general model of the sound scene

$$\mathbf{a}(t, f) = \mathbf{a}_{\mathrm{s}}(t, f) + \mathbf{a}_{\mathrm{d}}(t, f) = \mathbf{Y}_{\mathrm{s}}(t, f)\mathbf{s}(t, f) + \mathbf{a}_{\mathrm{d}}(t, f), \tag{5}$$

as a combination of multiple $K < M$ directional source signals, captured in $\mathbf{a}_{\mathrm{s}}$, and an additional component without clear directionality including ambient sound, incoherently distributed sources, and late reverberation, captured in $\mathbf{a}_{\mathrm{d}}$. COMPASS aims at estimating the parameters of these two components, and exploiting them during synthesis/reproduction. Similar to most spatial audio coding methods, it operates on time-frequency transformed signals, e.g. with an appropriate short-time Fourier transform (STFT) or filterbank [17].

Assuming that the ambient and directional part are uncorrelated, the narrowband second-order statistics of (5) are given by the power spectral density (PSD) matrix

$$\mathbf{C}_{\mathbf{a}}(t, f) = \mathbb{E}\left[\mathbf{a}(t, f)\mathbf{a}^{\mathrm{H}}(t, f)\right] = \mathbf{C}_{\mathbf{a},\mathrm{s}}(t, f) + \mathbf{C}_{\mathbf{a},\mathrm{d}}(t, f), \tag{6}$$

where $\mathbb{E}[\cdot]$ denotes statistical expectation. Assuming additionally that the source signals are uncorrelated between them, based on (3) their PSD matrix is (dropping the (t, f) indices for compactness)

$$\mathbf{C}_{\mathbf{a},\mathrm{s}} = \mathbb{E}\left[\mathbf{a}_{\mathrm{s}}\mathbf{a}_{\mathrm{s}}^{\mathrm{H}}\right] = \mathbf{Y}_{\mathrm{s}}\mathbf{C}_{\mathrm{s}}\mathbf{Y}_{\mathrm{s}}^{\mathrm{T}} = \sum_{k=1}^{K} P_k\mathbf{y}(\boldsymbol{\gamma}_k)\mathbf{y}^{\mathrm{T}}(\boldsymbol{\gamma}_k), \tag{7}$$

where $\mathbf{C}_{\mathrm{s}} = \mathrm{diag}\,[\mathbf{p}_{\mathrm{s}}]$ contains the source powers $\mathbf{p}_{\mathrm{s}} = [P_1, ..., P_K]^{\mathrm{T}}$, with total source power $P_{\mathrm{s}} = \sum_k P_k$. Based on the property $\|\mathbf{y}(\boldsymbol{\gamma})\|^2 = M$, and (7), the power of the source component is

$$P_{\mathbf{a},\mathrm{s}} = \mathbb{E}\left[\|\mathbf{a}_{\mathrm{s}}\|^2\right] = \mathrm{trace}\,[\mathbf{C}_{\mathbf{a},\mathrm{s}}] = M\sum_{k=1}^{K} P_k = MP_{\mathrm{s}}, \tag{8}$$

Contrary to most analysis methods, we do not assume necessarily isotropic diffuse conditions on the ambient component, in which case the PSD in the SH domain reduces to $\mathbf{C}_{\mathbf{a},\mathrm{d}} = P_{\mathrm{d}}\mathbf{I}$, where P_{d} is the power of the diffuse signal. In the more general case that the ambient signal is non-isotropic but incoherently-distributed, it is straightforward to show that $\mathbf{C}_{\mathbf{a},\mathrm{d}}$ is non-diagonal but its power is still captured by

$$P_{\mathbf{a},\mathrm{d}} = \mathbb{E}\left[\|\mathbf{a}_{\mathrm{d}}\|^2\right] = \mathrm{trace}\,[\mathbf{C}_{\mathbf{a},\mathrm{d}}] = MP_{\mathrm{d}}. \tag{9}$$

3.1. Parameter analysis

Dominance of directional or ambient components in the sound scene is reflected in the structure of the second-order statistics of the ambisonic signals, as captured in the PSD matrix of (6). Detection of these conditions and parameter estimation in COMPASS is based on the subspace principle of sensor array processing. The eigenvalue decomposition (EVD) of the PSD has the form

$$\mathbf{C}_{\mathbf{a}} = \mathbf{VUV}^{\mathrm{H}} = \sum_{m=1}^{K} \lambda_m\mathbf{v}_m\mathbf{v}_m^{\mathrm{H}} + \sum_{m=K+1}^{M} \lambda_m\mathbf{v}_m\mathbf{v}_m^{\mathrm{H}}, \tag{10}$$

where $\lambda_1 > ... > \lambda_m > ... > \lambda_M \geq 0$ are the sorted eigenvalues of the EVD, $\mathbf{v}_m$ the respective eigenvectors, and $K < M$ the assumed number of sources. This decomposition is exploited to detect diffuse conditions and estimate the number of sources and source DOAs.

In practice, the PSD matrix $\mathbf{C}_{\mathbf{a}}$ is estimated by temporal and frequency averaging

$$\mathbf{C}_{\mathbf{a}}(t, j) = \alpha\mathbf{C}_{\mathbf{a}}(t-1, j) + \frac{(1-\alpha)}{\Delta f_j}\sum_{f_{j-1}+1}^{f_j} \mathbf{a}(t, f)\mathbf{a}^{\mathrm{H}}(t, f), \tag{11}$$

where $\alpha \in [0, 1]$ is a temporal averaging coefficient, j is the averaged band index, f_j is its upper frequency index, $\Delta f_j = f_j - f_{j-1}$,

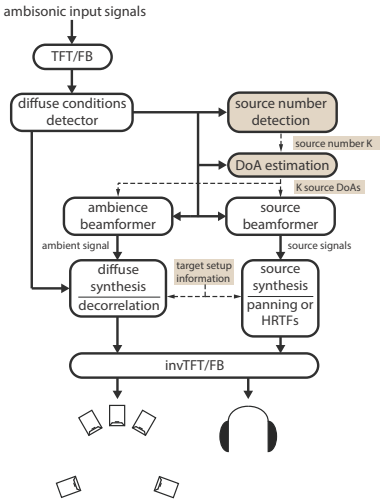

Fig. 1: Block diagram of the proposed COMPASS method. The TFT/FB blocks denote a time-frequency transform or filterbank.

and $f_0 = 0$. The coefficient $\alpha \in [0, 1]$ is linked to the decay time constant τ of the smoothing by $\alpha = e^{-R/(\tau f_s)}$, with R the hop size of the windowed transform or the decimation factor of the filter bank, and f_s the sample rate. The frequency averaging is performed over auditory-resolution inspired bands, such as equivalent rectangular bandwidth (ERB) ones.

The number of sources K is estimated from the distribution of the eigenvalues $\mathbf{U}$. More specifically, we use the SORTE method, which has been shown to be robust [18] and avoids manually adjusted thresholds. It is based on the differences of the sorted eigenvalues $\nabla\lambda_i = \lambda_i - \lambda_{i+1}$, for $i = 1, ..., M-1$. The number of sources according to SORTE is given by $K = \text{argmin}_k\, o(k)$, for $k = 1, ..., M-2$ with

$$o(k) = \begin{cases} \frac{\sigma^2_{k+1}}{\sigma^2_k}, & \sigma^2_k > 0 \\ +\infty, & \sigma^2_k = 0 \end{cases}, \quad \text{for } k = 1, ..., M-2, \tag{12}$$

and with the eigenvalue difference variances σ^2_k defined as

$$\sigma^2_k = \frac{1}{M-k} \sum_{i=k}^{M-1} \left(\nabla\lambda_i - \frac{1}{M-k} \sum_{i=k}^{M-1} \nabla\lambda_i \right)^2. \tag{13}$$

Additionally, a normalized measure of diffuseness $\psi \in [0, 1]$ is used in order to assess dominant diffuse conditions, in which case, estimation of source parameters is bypassed, and only ambience synthesis is applied. Diffuseness is based on the variance of the eigenvalues of the PSD matrix, as proposed in [19] and diffuse conditions are assumed if $\psi > 0.9$.

For the DoA estimation of the multiple sources, we use the MUSIC method [20]. We define a dense uniform grid of G directions $\mathbf{\Gamma}_\text{G} = [\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_G]$ and the associated SH matrix $\mathbf{Y}_\text{g} = [\mathbf{y}(\boldsymbol{\gamma}_1), ..., \mathbf{y}(\boldsymbol{\gamma}_G)]$. For K source components, we construct the noise subspace $\mathbf{V}_\text{n}$ from the eigenvectors corresponding to the lowest $M - K$ eigenvalues. The MUSIC spectrum is then given by

$$\mathbf{p}_\text{MUSIC} = \text{diag}\left[\mathbf{Y}_\text{g}^\text{T} \mathbf{V}_\text{n} \mathbf{V}_\text{n}^\text{H} \mathbf{Y}_\text{g} \right]. \tag{14}$$

The source DoAs $\mathbf{\Gamma}_\text{s} \in \mathbf{\Gamma}_\text{G}$ are found at the grid directions for which the K smallest local minima of (14) occur.

3.2. Separation and power estimation

Knowing the source DoAs, we can estimate the source and ambient signals and their powers. Regarding the directional part, we consider an $K \times M$ beamforming matrix $\mathbf{W}_\text{s}$ producing nulls to all estimated DoAs apart from the source of interest, given by the solution to the constraints $\mathbf{W}_\text{s}\mathbf{Y}_\text{s} = \mathbf{I}_K$, which is the pseudo-inverse $\mathbf{Y}_\text{s}^+$ of $\mathbf{Y}_\text{s}$ for the estimated DOAs

$$\mathbf{W}_\text{s} = \mathbf{Y}_\text{s}^+ = (\mathbf{Y}_\text{s}^\text{T}\mathbf{Y}_\text{s})^{-1}\mathbf{Y}_\text{s}^\text{T}. \tag{15}$$

The estimated amplitudes $\mathbf{s}$ of the source signals and the source powers $\mathbf{p}_\text{s}$ are then

$$\mathbf{s} = \mathbf{W}_\text{s}\mathbf{a} \tag{16}$$

$$\mathbf{p}_\text{s} = \text{diag}\left[\mathbf{W}_\text{s}\mathbf{C}_\mathbf{a}\mathbf{W}_\text{s}^\text{H} \right]. \tag{17}$$

Regarding the ambient part, we aim to estimate directly its ambisonic image $\mathbf{a}_\text{d}$. It is computed simply as the residual after the encoded source signals have been extracted from the ambisonic signals, and is thus expected to contain mostly reverberant and ambient components. Using (16)

$$\mathbf{a}_\text{d} = \mathbf{a} - \mathbf{Y}_\text{s}\mathbf{s} = \mathbf{a} - \mathbf{Y}_\text{s}\mathbf{W}_\text{s}\mathbf{a} = \mathbf{W}_\text{d}\mathbf{a} \tag{18}$$

$$\mathbf{W}_\text{d} = \mathbf{I}_M - \mathbf{Y}_\text{s}\mathbf{W}_\text{s} = \mathbf{I}_M - \mathbf{Y}_\text{s}\mathbf{Y}_\text{s}^+, \tag{19}$$

where the $M \times M$ beamforming matrix $\mathbf{W}_\text{d}$ defines an orthogonal projection on the nullspace of $\mathbf{Y}_\text{s}^\text{T}$. Finally, the ambient power is

$$P_\text{d} = \frac{1}{M}\text{trace}\left[\mathbf{W}_\text{d}\mathbf{C}_\mathbf{a}\mathbf{W}_\text{d}^\text{H} \right]. \tag{20}$$

3.3. Synthesis

The source components should be distributed to the output channels with maximum directional concentration from their analyzed DoAs. Such distribution functions corresponds to amplitude panning gains or HRTFs, and they can be considered as synthesis steering vectors. We denote a vector of such real or complex spatialization gains as $\mathbf{g}(\boldsymbol{\gamma}) = [g_1(\boldsymbol{\gamma}), ..., g_L(\boldsymbol{\gamma})]^\mathbf{T}$. Having estimated the source signal amplitudes of (16), associated with their DoAs, the source signals can be spatialized as

$$\mathbf{z}_\text{s} = \mathbf{G}_\text{s}\mathbf{s} = \mathbf{G}_\text{s}\mathbf{W}_\text{s}\mathbf{a}. \tag{21}$$

where $\mathbf{G}_\text{s} = [\mathbf{g}(\boldsymbol{\gamma}_1), ..., \mathbf{g}(\boldsymbol{\gamma}_K)]$ is the $L \times K$ matrix of spatialization gains for the estimated directions $\mathbf{\Gamma}_\text{s}$. However, instead of applying directly the above synthesis matrix, we prefer to let the linear ambisonic decoding $\mathbf{D}$ of (4) achieve a preliminary spatialization, and apply an additional adaptive matrix $\mathbf{A}_\text{s}$ to achieve the spatialization operation of (21), which is the solution to $\text{argmin}\|\mathbf{A}_\text{s}\mathbf{D} - \mathbf{G}_\text{s}\mathbf{W}_\text{s}\|_\text{F}^2$

$$\mathbf{A}_\text{s} = \mathbf{G}_\text{s}\mathbf{W}_\text{s}\mathbf{D}^\text{H}(\mathbf{D}\mathbf{D}^\text{H})^{-1} \tag{22}$$

$$\mathbf{z}_\text{s} = \mathbf{A}_\text{s}\mathbf{D}\mathbf{a}. \tag{23}$$

Synthesis of the ambient component may use only the basic linear decoding of (4)

$$\mathbf{z}_\text{d} = \mathbf{D}\mathbf{a}_\text{d} = \mathbf{D}\mathbf{W}_\text{d}\mathbf{a}, \tag{24}$$

or if enhanced ambience rendering is desired, a decorrelation stage can be injected inside the decoding of (4) to achieve an incoherent spatial distribution

$$\mathbf{z}_\text{d} = (1/V)\mathbf{G}_\text{vls}\mathcal{D}\left[\mathbf{Y}_\text{vls}^\text{T}\mathbf{a}_\text{d} \right] = (1/V)\mathbf{G}_\text{vls}\mathcal{D}\left[\mathbf{Y}_\text{vls}^\text{T}\mathbf{W}_\text{d}\mathbf{a} \right], \tag{25}$$

where $\mathcal{D}[\cdot]$ denotes decorrelation of the signal set enclosed in the brackets. Decorrelation may deemed necessary for low orders, in which the linear decoding may fail to deliver sufficiently incoherent signals under diffuse conditions. We note that contrary to most parametric playback methods that employ a diffuse component [8], COMPASS does not force isotropic conditions and preserves the directionality of the ambience and reverberation which may have significant perceptual qualities [21].

3.4. Rendering and control parameters

Three parameters β, γ, δ controlling different aspects of the rendering are introduced, such that

$$\mathbf{z}(t,f) = \mathbf{M}(t,f,\beta,\gamma,\delta)\mathbf{a}(t,f), \tag{26}$$

where $\mathbf{M}$ is the final synthesis matrix derived from the analysis and control parameters. The first parameter $\beta \in [0,1]$ imposes temporal smoothing on the synthesis matrix similar to (11), an operation common in parametric spatial sound processing [22, 8]

$$\mathbf{M}(t,f,\beta,\gamma,\delta) = \beta\mathbf{M}(t-1,f,\beta,\gamma,\delta) + (1-\beta)\mathbf{R}(t,f,\gamma,\delta), \tag{27}$$

where $\mathbf{R}$ is the instantaneous synthesis matrix computed for the current frame. The second control parameter $\gamma \in [0,1]$ cross-fades between purely linear ambisonic decoding and the adaptive one, so that the user can control the amount of parametric enhancement according to the sound scene. The third parameter $\delta(f) \in [0,1]$ controls the source-to-ambience ratio and can be frequency-dependent. Based on (23) and (24), the final instantaneous rendering matrix is

$$\mathbf{R}(\gamma,\delta) = \delta\left[\gamma\mathbf{A}_\mathrm{s} + (1-\gamma)\mathbf{I}_L\right]\mathbf{D} + (1-\delta)\mathbf{D}\left[\gamma\mathbf{W}_\mathrm{d} + (1-\gamma)\mathbf{I}_M\right]. \tag{28}$$

4. EVALUATION

In order to assess the performance of COMPASS, we conducted a multiple-stimulus with hidden reference and anchor (MUSHRA) listening test [23], for headphone playback. Five sound scenes were simulated for anechoic and reverberant conditions with a varying numbers of sources. Anechoic scenes included only propagation delays and directional encoding for the source distances and DoAs, while their reverberant counterparts introduced full reverberation simulated using the image source method [24]. The absorption profiles were tuned to match frequency-dependent target reverberation times, ranging from 0.6–1.2 seconds, specified in octave bands, and including attenuation due to air absorption. The propagation filter for each image source was then convolved with the appropriate HRTF, so that a final binaural spatial room impulse response (SRIR) was generated for each source in the sound scene. By convolving each SRIR with the source signals, a reference binaural version of the sound scene was generated. The same image source propagation filters were additionally encoded to third-order Ambisonics (TOA), resulting in ambisonic SRIRs. Convolution with the source signals resulted in a TOA encoding of the overall sound scene. A first-order ambisonic (FOA) version was obtained by keeping only the first 4 channels of the TOA signals.

There are three free-field sound scenes, labelled here as *bandDry*, *speakersDry*, and *orchestraDry*, comprising of dry recordings of a band of 4 instruments, 3 speakers, and a 24-instrument orchestra, distributed on the front hemisphere. Two reverberant scenes consist of the band and a soundscape with handclaps, a fountain, piano and female speech, labelled as *bandRev* and *soundscapeRev* respectively. The ambisonic scenes were decoded binaurally using (4) for both FOA (*Ambi_1*), and TOA (*Ambi_3*), and the same signals were also processed by the proposed method resulting in *COMPASS_1* and *COMPASS_3*. Three MUSHRA tests were constructed. In the first, the listeners were instructed to rate perceived distance from the reference, with 100 being indistinguishable, considering overall quality, including timbral and spatial differences, and artifacts. The two additional tests were aimed to separate timbral and spatial effects. In the first one, the RMS magnitude response between left and right signals of the binaural reference was used to match the RMS response of each method to the reference, hence minimizing timbral differences while preserving spatial ones and artifacts. In the second one, the reference signal RMS response was matched to the response of the test cases, creating coloured versions of it matched timbrally to the outputs of the methods but eliminating spatial differences. The three tests are labelled *overall*, *spatial*, and *timbral*. All stimuli found in the tests are available online[1].

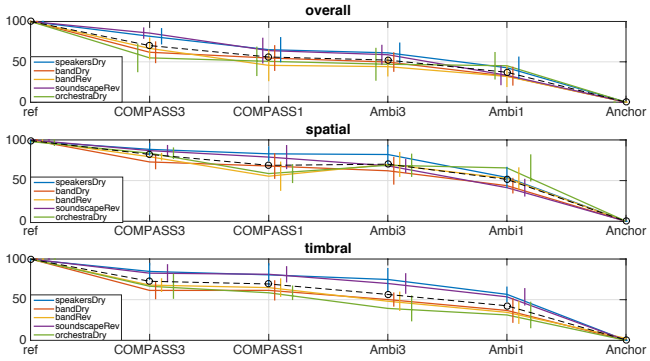

Fig. 2: MUSHRA mean scores and 95% confidence intervals across all subjects. Circles indicate the mean scores across all sound scenes.

Twelve expert listeners participated in the test. According to the mean scores in Fig. 2, *COMPASS_1* with FOA input has an overall quality similar to linear ambisonic decoding with TOA input (*Ambi_3*). Furthermore, COMPASS is closer in timbre to the reference compared to ambisonic decoding, for both FOA and TOA input. Spatially, *COMPASS_1* achieves a performance close to TOA decoding (*Ambi_3*), while *COMPASS_3* with TOA input outperforms TOA decoding. The results resemble the ones presented by the authors using DirAC in [9], under similar sound scenarios. Comparisons of COMPASS with DirAC and other parametric approaches is scheduled for future work.

5. CONCLUSION

A new method of ambisonic signal analysis and synthesis is proposed, based on a more general model of the sound scene compared to previous approaches, with applications to flexible high-resolution reproduction, enhancement of low-order signals, and spatial modifications of the sound scene. The method extracts multiple source signals and an ambient residual that is not necessarily isotropic, and renders them to loudspeakers or headphones with control between linear and fully parametric rendering, or between foreground and ambient background components. Listening test results indicate that the method improves spatial, timbral, and overall perceived quality compared to Ambisonics with the same order of input.

[1] http://research.spa.aalto.fi/publications/papers/icassp18-compass/

6. REFERENCES

[1] Michael A Gerzon, "Periphony: With-height sound reproduction," *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, 1973.

[2] Mark A Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *Journal of the Audio Engineering Society*, vol. 53, no. 11, pp. 1004–1025, 2005.

[3] Franz Zotter and Matthias Frank, "All-round ambisonic panning and decoding," *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 807–820, 2012.

[4] Benjamin Bernschütz, Arnau Vázquez Giner, Christoph Pörschmann, and Johannes Arend, "Binaural reproduction of plane waves with reduced modal order," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 972–983, 2014.

[5] Amir Avni, Jens Ahrens, Matthias Geier, Sascha Spors, Hagen Wierstorf, and Boaz Rafaely, "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2711–2721, 2013.

[6] Stéphanie Bertet, Jérôme Daniel, Etienne Parizet, and Olivier Warusfel, "Investigation on localisation accuracy for first and higher order Ambisonics reproduced sound sources," *Acta Acustica united with Acustica*, vol. 99, no. 4, pp. 642–657, 2013.

[7] Audun Solvang, "Spectral impairment of two-dimensional higher order Ambisonics," *Journal of the Audio Engineering Society*, vol. 56, no. 4, pp. 267–279, 2008.

[8] Ville Pulkki, Archontis Politis, Mikko-Ville Laitinen, Juha Vilkamo, and Jukka Ahonen, "First-order Directional Audio Coding (DirAC)," in *Parametric Time-Frequency Domain Spatial Audio*, p. 89. John Wiley & Sons, 2017.

[9] Archontis Politis, Juha Vilkamo, and Ville Pulkki, "Sector-based parametric sound field reproduction in the spherical harmonic domain," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 852–866, 2015.

[10] Svein Berge and Natasha Barrett, "A new method for B-format to binaural transcoding," in *40th Int. Conf. of AES*, Tokyo, Japan, 2010.

[11] Archontis Politis, Tapani Pihlajamäki, and Ville Pulkki, "Parametric spatial audio effects," in *15th Int. Conf. Digital Audio Effects (DAFx-12)*, York, UK, 2012.

[12] Andrew Wabnitz, Nicolas Epain, and Craig T Jin, "A frequency-domain algorithm to upscale ambisonic sound scenes," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 385–388.

[13] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[14] Oliver Thiergart, Maja Taseska, and Emanuël AP Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 2182–2196, 2014.

[15] Archontis Politis and David Poirier-Quinot, "JSAmbisonics: A Web Audio library for interactive spatial sound processing on the web," in *Interactive Audio Systems Symposium*, York, UK, 2016.

[16] Ronald H Hardin and Neil JA Sloane, "McLaren's improved snub cube and other new spherical designs in three dimensions," *Discrete & Computational Geometry*, vol. 15, no. 4, pp. 429–441, 1996.

[17] Juha Vilkamo and Tom Backstrom, "Time–frequency processing: Methods and tools," in *Parametric Time-Frequency Domain Spatial Audio*, p. 3. John Wiley & Sons, 2017.

[18] Keyong Han and Arye Nehorai, "Improved source number detection and direction estimation with nested arrays and ULAs using jackknifing," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 6118–6128, 2013.

[19] Nicolas Epain and Craig Jin, "Spherical harmonic signal covariance and sound field diffuseness," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 10, pp. 1796–1807, 2016.

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

[21] Olli Santala and Ville Pulkki, "Directional perception of distributed sound sources," *The Journal of the Acoustical Society of America*, vol. 129, no. 3, pp. 1522–1530, 2011.

[22] Christof Faller, "Upmixing and beamforming in professional audio," in *Parametric Time-Frequency Domain Spatial Audio*, p. 329. John Wiley & Sons, 2017.

[23] ITU-R Recommendation BS.1534-1, "Method for the subjective assessment of intermediate quality levels of coding systems," Tech. Rep., International Telecommunication Union (ITU), Geneva, Switzerland, 2003.

[24] Patrick M Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *The Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1527–1529, 1986.