PARAMETRIC SPATIAL AUDIO EFFECTS BASED ON THE MULTI-DIRECTIONAL DECOMPOSITION OF AMBISONIC SOUND SCENES

Leo McCormack[1] *, Archontis Politis*[2] *, Ville Pulkki*[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
[2]Faculty of Information Technology and Communication Sciences, Tampere University, Finland
leo.mccormack@aalto.fi

ABSTRACT

Decomposing a sound-field into its individual components and respective parameters can represent a convenient first-step towards offering the user an intuitive means of controlling spatial audio effects and sound-field modification tools. The majority of such tools available today, however, are instead limited to linear combinations of signals or employ a basic single-source parametric model. Therefore, the purpose of this paper is to present a parametric framework, which seeks to overcome these limitations by first dividing the sound-field into its multi-source and ambient components based on estimated spatial parameters. It is then demonstrated that by manipulating the spatial parameters prior to reproducing the scene, a number of sound-field modification and spatial audio effects may be realised; including: directional warping, listener translation, sound source tracking, spatial editing workflows and spatial side-chaining. Many of the effects described have also been implemented as real-time audio plug-ins, in order to demonstrate how a user may interact with such tools in practice.

1. INTRODUCTION

The ability to manipulate spatial sound scenes, prior to reproducing them over the target playback setup, has a number of important applications. These include: head-tracked informed sound scene rotations during virtual and augmented reality rendering; providing spatial editing tools to audio engineers engaged in the production of immersive content; and offering users creative outlets, by way of spatial audio effects. A popular framework for synthesising, capturing, modifying, and reproducing spatial sound scenes is Ambisonics [1], which is based solely on linear mappings of the channel signals. The framework operates by decoupling the recording and playback setups through the use of spherical harmonic (SH) signals, which serve as an intermediary. The process of converting a monophonic signal or microphone array signals into SH signals is commonly referred to as Ambisonic *encoding*. Whereas, mapping these SH signals to the target loudspeaker setup or through binaural filters, is often called Ambisonic *decoding*. Spatial manipulations may be realised by applying linear transformations on the intermediate SH signals. Some of these transformations are robust and well-defined, such as sound-field mirroring and rotations [2], whereas other transformations, for example, directional warping and zooming [3, 4, 5], are generally less defined or can be limited by the SH order of expansion.

Alternatives to this purely linear framework include the signal-dependent approaches described in [6, 7, 8, 9], which operate by describing the sound scene based on estimated spatial parameters. These parametric methods can offer improved spatial resolution over their linear counterparts during rendering, and can present new and unique avenues for spatial audio effects processing that would otherwise not be realisable in a linear manner. Many of these parametric alternatives also operate on SH signals, which means that they retain much of the convenience of the Ambisonics framework and may be used interchangeably within traditional linear workflows. In [10, 11, 12], a number of spatial effects were described based on the manipulation of the analysed parameters provided by the first-order Directional Audio Coding (DirAC) parametric model [6]. First-order DirAC operates by estimating a single direction-of-arrival (DoA) and a diffuseness parameter per time-frequency tile, often based on the energetic properties of the active-intensity vector. A low diffuseness value means that most of the corresponding time-frequency tile energy is considered to be that of a plane-wave in the estimated DoA. Whereas, a high diffuseness value indicates that the tile corresponds to diffuse noise and/or reverberation. For reproduction, directional components are routed directly to the target playback setup via amplitude-panning or convolving them with the respective binaural filters. The diffuse components are then distributed to all output channels and decorrelated. In [10, 11, 12], it was described how these spatial parameters could be manipulated for the purpose of realising spatial audio effects; including: zooming, translation around the receiver, warping, and direct-to-diffuse (DDR) ratio control. Sound-field zooming was also explored in [13, 14], based on the manipulation of the diffuseness parameter within the first-order DirAC model.

Many existing parametric spatial audio effects have therefore been constrained to first-order SH input, and have operated based on the limited model of: a single dominant directional cue per time-frequency, accompanied by their respective spatial coherence cues (based on direct-to-diffuse ratio or diffuseness). In this work, the Coding and Multi-Parameterisation of Ambisonics Sound Scenes (COMPASS) model [8] is explored for the task of revisiting previously proposed spatial audio effects. The model supports arbitrary input order and can estimate the DoAs of multiple simultaneous sources; subsequently employing spatial-filtering to segregate the sound-field into its source and anisotropic ambient components. Owing to its greater flexibility, new spatial manipulations and effects are also described using the model, including: sound source tracking, spatial editing workflows, and flexible directional warping. To demonstrate how the effects may be realised in practice, many have also been implemented as VST audio plug-ins and can be downloaded from the companion web-page[1].

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

2. AMBISONICS FRAMEWORK

In this work it is assumed that the sound-field comprises the spatial distribution of plane-waves, $a(t, f, \gamma)$, at time t and frequency f, where γ is a unit vector at azimuth ϕ and elevation θ, respectively. The Nth order spherical harmonic transform (SHT) of this spatial distribution provides the sound-field coefficients $\mathbf{a}(t, f)$ as

$$\mathbf{a}(t, f) = \mathcal{SHT}\left\{a(t, f, \gamma)\right\} = \int_{\gamma} a(t, f, \gamma)\mathbf{y}(\gamma)\mathbf{d}\gamma, \quad (1)$$

where the integration is conducted over the surface of the unit sphere, and $\mathbf{y}(\gamma)$ denotes a vector of spherical harmonic (SH) values $Y_{nm}(\gamma)$ of order n and degree $m \in [-n, n]$. For a band-limited representation of order N, there are $Q = (N + 1)^2$ transformed signals and SHs in the vectors above. Furthermore, from henceforth, the established Ambisonics convention of real orthonormal SHs is employed.

The end-to-end Ambisonics processing framework can be demonstrated compactly, based on the following series of signal-independent linear matrix operations

$$\mathbf{z}(t, f) = \mathbf{DT}\left[\mathbf{Y}_{\mathrm{s}}\mathbf{s}(t, f) + \mathbf{E}(f)\mathbf{x}(t, f)\right] = \mathbf{DTa}(t, f), \quad (2)$$

where $\mathbf{s}(t, f) = [s_1(t, f), ..., s_K(t, f)]^{\mathrm{T}}$ denotes K monophonic source signals, and $\mathbf{x}(t, f) = [x_1(t, f), ..., x_M(t, f)]^{\mathrm{T}}$ denotes M microphone array signals. These signals are encoded into the SH domain via matrices $\mathbf{Y}_{\mathrm{s}} = [\mathbf{y}(\gamma_1), ..., \mathbf{y}(\gamma_K)]$ and $\mathbf{E}(f) \in \mathbb{C}^{(N+1)^2 \times M}$, respectively. Since their base representations are the same, multiple SH recordings and/or encoded source signals may be combined simply via summation. For more information regarding computing $\mathbf{E}$, the reader is referred to [15, 16, 17]. $\mathbf{T}$ is then an optional spatial transformation matrix, which modifies the spatial properties of the sound scene directly in the SH domain. Examples of SH transformations include: rotations [2], directional warping of the sound distribution [3, 4], and directional loudness modifications [4, 5]. Finally, $\mathbf{D}$ is a decoding matrix, which defines a linear mapping of the SH signals to the L output channels, $\mathbf{z}(t, f) = [z_1(t, f), ..., z_L(t, f)]^{\mathrm{T}}$, of the reproduction system. For loudspeaker-based reproduction, the ambisonic decoding matrix $\mathbf{D}$ is of size $L \times Q$, and often derived based solely on the loudspeaker directions and transform order; available solutions include [18, 19]. Whereas, for binaural reproduction, the decoding matrix $\mathbf{D}(f)$ is instead frequency-dependent, since it is computed based on a grid of HRTF measurements, and available solutions include [20, 21].

3. PARAMETRIC FRAMEWORK

The parametric framework employed for this work is an extended formulation of the COMPASS method [8]; incorporating changes to provide greater freedom over the manipulation of its rendering behaviour. Considering the general case of a mixed sound-field, with a number of source signals of $K < Q$ and an additional diffuse component, the ambisonic signals may be expressed as

$$\mathbf{a}(t, f) = \mathbf{a}_{\mathrm{s}}(t, f) + \mathbf{a}_{\mathrm{d}}(t, f) = \mathbf{Y}_{\mathrm{s}}\mathbf{s}(t) + \mathbf{a}_{\mathrm{d}}(t, f). \quad (3)$$

Assuming that the source signals are uncorrelated with the diffuse signal and between themselves, their respective spatial covariance matrices are given as

$$\mathbf{C}_{\mathbf{a}}(t, f) = \mathbb{E}\left[\mathbf{a}(t, f)\mathbf{a}(t, f)^{\mathrm{T}}\right] = \mathbf{C}_{\mathbf{a},\mathrm{s}}(t, f) + \mathbf{C}_{\mathbf{a},\mathrm{d}}(t, f), \quad (4)$$

where $\mathbb{E}[.]$ denotes the expectation operator.

Note that the time-frequency indices are henceforth omitted from this section for the brevity of notation. The source covariance matrix can be expressed as

$$\mathbf{C}_{\mathbf{a},\mathrm{s}} = \mathbb{E}\left[\mathbf{a}_{\mathrm{s}}\mathbf{a}_{\mathrm{s}}^{\mathrm{T}}\right] = \mathbf{Y}_{\mathrm{s}}\mathbf{C}_{\mathbf{s}}\mathbf{Y}_{\mathrm{s}}^{\mathrm{T}} = \sum_{k=1}^{K} P_k\mathbf{y}(\gamma_k)\mathbf{y}^{\mathrm{T}}(\gamma_k), \quad (5)$$

where $\mathbf{C}_{\mathbf{s}} = \mathrm{diag}[\mathbf{p}_{\mathrm{s}}]$ is a diagonal matrix comprising the source powers $\mathbf{p}_{\mathrm{s}} = [P_1, ..., P_K]^{\mathrm{T}}$, with the total source power $P_{\mathrm{s}} = \sum_k P_k$. Given that $||\mathbf{y}(\gamma_1)||^2 = Q$, the power of the source and diffuse components are given as

$$P_{\mathbf{a},\mathrm{s}} = \mathbb{E}\left[||\mathbf{a}_{\mathrm{s}}||^2\right] = \mathbf{tr}\left[\mathbf{C}_{\mathbf{a},\mathrm{s}}\right] = QP_{\mathrm{s}}, \quad (6)$$

$$P_{\mathbf{a},\mathrm{d}} = \mathbb{E}\left[||\mathbf{a}_{\mathrm{d}}||^2\right] = \mathbf{tr}\left[\mathbf{C}_{\mathbf{a},\mathrm{d}}\right] = QP_{\mathrm{d}}, \quad (7)$$

where P_{d} is the power of the diffuse signals, and $\mathbf{tr}[.]$ denotes the trace operator.

3.1. Analysis

The parameter analysis conducted by the COMPASS method involves first detecting the number of sources, followed by determining their respective DoAs. Detection of the number of sources is commonly carried out based on the subspace principle of sensor array processing, whereby the eigenvalue decomposition (EVD) of the spatial correlation matrix is first computed as

$$\mathbf{C}_{\mathbf{a}} = \mathbf{VUV}^{\mathrm{H}} = \sum_{q=1}^{Q} \lambda_q\mathbf{v}_q\mathbf{v}_q^{\mathrm{H}} = \sum_{q=1}^{K} \lambda_q\mathbf{v}_q\mathbf{v}_q^{\mathrm{H}} + \sum_{q=K+1}^{Q} \lambda_q\mathbf{v}_q\mathbf{v}_q^{\mathrm{H}}, \quad (8)$$

where $\lambda_1 > ... > \lambda_q > ... > \lambda_Q \geq 0$ are the eigenvalues of the EVD, and $\mathbf{v}_q$ are their respective eigenvectors. It is then assumed that the lowest eigenvalues of $K < q \leq Q$ will all be equal or similar to the diffuse power P_{diff}, whereas the eigenvalues $1 \leq q \leq K$ should correspond to the powers of the sources, with $\lambda_q > P_{\mathrm{diff}}$. For a detailed overview of different source detection algorithms, the reader is referred to [22].

Once the number of sources has been detected, their DoAs can be estimated based on a number of different methods. Many of these involve scanning a grid of directions, followed by ascertaining the maxima or minima within the resulting activity-maps. However, if the employed detection algorithm operates in the subspace domain, then subspace DoA estimation methods, such as MUSIC [23] or ESPRIT [24, 25], are often convenient options in practice. For MUSIC, a dense grid of G directions $\mathbf{\Gamma}_{\mathrm{g}} = [\gamma_1, ..., \gamma_G]$ and the associated SH matrix $\mathbf{Y}_{\mathrm{g}} = [\mathbf{y}(\gamma_1), ..., \mathbf{y}(\gamma_G)]$ are employed. Assuming K directional components in the scene, the noise subspace $\mathbf{V}_{\mathrm{n}}$ may be constructed from the eigenvectors corresponding to the lowest $Q - K$ eigenvalues. The MUSIC spectrum is then given by

$$\mathbf{p}_{\mathrm{MUSIC}} = \mathrm{diag}[\mathbf{Y}_{\mathrm{g}}^{\mathrm{T}}\mathbf{V}_{\mathrm{n}}\mathbf{V}_{\mathrm{n}}^{\mathrm{H}}\mathbf{Y}_{\mathrm{g}}]. \quad (9)$$

The source DoAs $\tilde{\mathbf{\Gamma}}_{\mathrm{s}} \in \mathbf{\Gamma}_{\mathrm{g}}$ are then found at the grid directions for which the K minima of (9) occur.

3.2. Synthesis

Once the number of sources and their respective DoAs have been estimated, the source beamforming matrix $\mathbf{W}_{\mathrm{s}} \in \mathbb{R}^{K \times Q}$ may be

computed based on the following regularised inversion

$$\mathbf{W}_\mathrm{s} = (\mathbf{Y}_\mathrm{s}^\mathrm{T}\mathbf{Y}_\mathrm{s} + \beta^2\mathbf{I}_K)^{-1}\mathbf{Y}_\mathrm{s}^\mathrm{T}, \tag{10}$$

where β is a regularisation parameter, and $\mathbf{I}_K \in \mathbb{R}^{K\times K}$ is an identity matrix. Note that this inversion has the effect of producing beamformers of unity gain in the estimated directions, while placing nulls towards the other DoAs. However, in practice, β can be DoA separation dependent, in order to bypass the null steering during cases where DoAs fall within the same angle as the main-lobe of the beamformer; i.e. reverting to $\mathbf{W}_\mathrm{s} = (1/K)\mathbf{Y}_\mathrm{s}^\mathrm{T}$ in such cases, in order to improve beamformer stability.

The estimated source signals and source powers may then be obtained as

$$\mathbf{s} = \mathbf{W}_\mathrm{s}\mathbf{a}, \tag{11}$$

$$\mathbf{p}_\mathrm{s} = \mathrm{diag}[\mathbf{W}_\mathrm{s}\mathbf{C}_\mathbf{a}\mathbf{W}_\mathrm{s}^\mathrm{T}]. \tag{12}$$

To reproduce the source signals over a L-channel target playback setup, the appropriate spatialisation gains are required $\mathbf{g}(\gamma) = [g_1(\gamma), ..., g_L(\gamma)]^\mathrm{H}$, which can be, for example, amplitude-panning gains for loudspeaker playback, HRTFs for binaural playback, or SH weights of optionally higher order than the input order (i.e. for upscaling). The spatialisation is therefore applied as

$$\mathbf{z}_\mathrm{s} = \mathbf{G}_\mathrm{r}\mathbf{s} = \mathbf{G}_\mathrm{r}\mathbf{W}_\mathrm{s}\mathbf{a}, \tag{13}$$

where $\mathbf{G}_\mathrm{r} = [\mathbf{g}(\gamma_1), ..., \mathbf{g}(\gamma_K)]$ are the spatialisation gains for all of the source signals in the reproduction directions $\mathbf{\Gamma}_\mathrm{r}$, which do not necessarily need to be the same as the estimated source directions $\tilde{\mathbf{\Gamma}}_\mathrm{s}$ used for the beamforming.

For the ambient rendering, the source signals are first re-encoded back into the SH domain and subtracted from the input scene as

$$\mathbf{a}_\mathrm{d} = \mathbf{a} - \mathbf{Y}_\mathrm{s}\mathbf{s} = \mathbf{a} - \mathbf{Y}_\mathrm{s}\mathbf{W}_s\mathbf{a} = \mathbf{W}_\mathrm{d}\mathbf{a}, \tag{14}$$

$$\mathbf{W}_\mathrm{d} = \mathbf{I}_Q - \mathbf{Y}_\mathrm{s}\mathbf{W}_\mathrm{s}. \tag{15}$$

This SH domain residual is then used to obtain the ambient signals as

$$\mathbf{z}_\mathrm{d} = (1/V)\mathbf{G}_\mathrm{v}\mathcal{D}[\mathbf{Y}_\mathrm{v}^\mathrm{T}\mathbf{a}_\mathrm{d}] = (1/V)\mathbf{G}_\mathrm{v}\mathcal{D}[\mathbf{Y}_\mathrm{v}^\mathrm{T}\mathbf{W}_\mathrm{d}\mathbf{a}], \tag{16}$$

where $\mathbf{Y}_\mathrm{v} \in \mathbb{R}^{V\times(N+1)^2}$ are SH weights for a uniform spherical arrangement of V *virtual* directions, $\mathbf{G}_\mathrm{v} \in \mathbb{C}^{L\times V}$ are spatialisation gains to map the virtual directions signals to the target playback setup, and $\mathcal{D}[.]$ denotes a decorrelation operation on the enclosed signals to enforce diffuse characteristics, if desired.

The final output signals are then obtained by simply summing the two streams

$$\mathbf{z} = \mathbf{z}_\mathrm{s} + \mathbf{z}_\mathrm{d}. \tag{17}$$

4. PARAMETRIC SPATIAL AUDIO EFFECTS

Due to the decoupling of the source and ambient rendering, it is possible to apply different effects to each of the two streams. For example, conventional linear operations such as mirroring, rotations [2], warping [3, 4], and directional loudness modifications [4, 5], may be applied to only the ambient part of the sound-field, $\mathbf{a}_\mathrm{d}$, as described by (14). However, this work instead focuses on how to realise sound-field modifications and effects based on parameter manipulations.

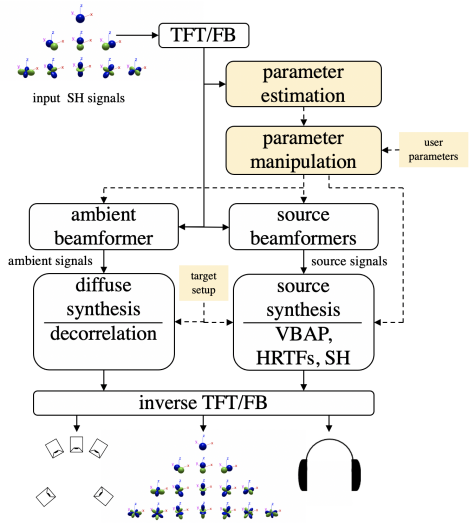

Figure 1: A block diagram for the parametric framework. Note that TFT refers to a time-frequency transform, such as a short-time Fourier transform (STFT) or filterbank.

For notation convenience, the employed parametric model and processing may be abstracted by the following: the analysis $\mathcal{A}$ of the input scene, to obtain the estimated source DoAs, is given as

$$\tilde{\mathbf{\Gamma}}_\mathrm{s}(t,f) = \mathcal{A}[\mathbf{a}(t,f)], \tag{18}$$

whereas the synthesis of the output source $\mathcal{S}_\mathrm{s}$ and ambient $\mathcal{S}_\mathrm{d}$ streams are denoted as

$$\mathbf{z}_\mathrm{s}(t,f) = \mathcal{S}_\mathrm{s}[\mathbf{a}(t,f), \tilde{\mathbf{\Gamma}}_\mathrm{s}(t,f), \mathbf{\Gamma}_\mathrm{r}(t,f), \mathbf{G}_\mathrm{r}(t,f)], \tag{19}$$

$$\mathbf{z}_\mathrm{d}(t,f) = \mathcal{S}_\mathrm{d}[\mathbf{a}(t,f), \tilde{\mathbf{\Gamma}}_\mathrm{s}(t,f)]. \tag{20}$$

Note that, by default, the reproduction directions are identical to the estimated DoAs $\mathbf{\Gamma}_\mathrm{r} = \tilde{\mathbf{\Gamma}}_\mathrm{s}$, and the reproduction gains $\mathbf{G}_\mathrm{r}$ also correspond to the $\tilde{\mathbf{\Gamma}}_\mathrm{s}$ directions. Therefore, the framework reverts back to the standard COMPASS rendering if no parameter manipulation is conducted.

4.1. Direct-to-diffuse balance manipulation

Since the direct and ambient streams are decoupled in the presented framework, a trivial parameter control method is to incorporate a biasing term during synthesis. This can have the effect of offering the user a means of emphasising the "natural" reverberation present in the scene (as described by the sound-field model), or de-emphasising it (akin to de-reverberation). The frequency-dependent biasing term, λ, may be applied simply as

$$\mathbf{z}_\mathrm{s}(t,f) = \lambda(f)\,\mathcal{S}_\mathrm{s}[\mathbf{a}(t,f), \tilde{\mathbf{\Gamma}}_\mathrm{s}(t,f), \mathbf{\Gamma}_\mathrm{r}(t,f), \mathbf{G}_\mathrm{r}(t,f)], \tag{21}$$

$$\mathbf{z}_\mathrm{d}(t,f) = \big(1-\lambda(f)\big)\,\mathcal{S}_\mathrm{d}[\mathbf{a}(t,f), \tilde{\mathbf{\Gamma}}_\mathrm{s}(t,f)], \tag{22}$$

which may be used either clinically, to correct for any model mismatches, or creatively; for example, the biasing term can be manipulated based on a time-varying modulator or any other external device.

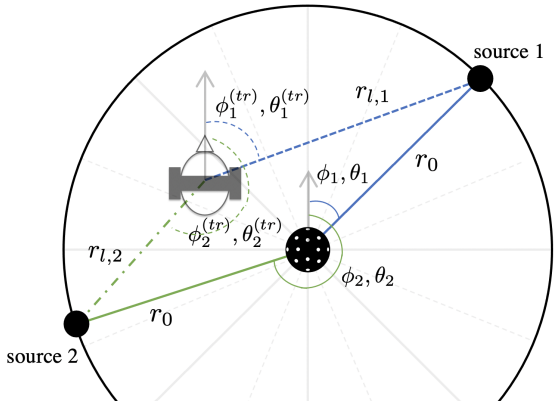

Figure 2: An example of how the reproduction directions may be manipulated to account for a translated listener position, based on first projecting the two sources onto a sphere of known or assumed radius [11].

4.2. Listener translation and acoustical zooming

Due to the recent resurgence of virtual and augmented reality devices, listener translation methods are becoming increasingly relevant. The framework can accommodate this effect by conducting the beamforming in the estimated $\tilde{\mathbf{\Gamma}}_\mathrm{s}$ directions as normal, but manipulating the reproduction directions based on a translated listener position $\mathbf{\Gamma}_\mathrm{r}^{(tr)}$. For a single SH receiver, the distance between the sound sources and the receiver position r_0 must either be known or assumed. After which, all source DoA estimates are then projected onto an arbitrarily shaped surface that defines the assumed or known source distances for all directions from the perspective of the receiver position. Based on the knowledge of the source DoAs, their distances, and the translated listener position with respect to the receiver position, the new reproduction directions can be computed using trigonometry; see e.g. Fig. 2. Furthermore, denoting the distance between the sound source position and the translated listener position as r_l, the inverse distance attenuation law may be used to compensate for the levels of the source beamformers as r_0/r_l. The ambient stream can therefore remain unchanged, with the parameter manipulation only effecting the source stream as

$$\mathbf{z}_\mathrm{s}(t,f) = \mathcal{S}_\mathrm{s}[\mathbf{a}(t,f), \tilde{\mathbf{\Gamma}}_\mathrm{s}(t,f), \mathbf{\Gamma}_\mathrm{r}^{(tr)}(t,f), \mathbf{G}_\mathrm{r}^{(tr)}(t,f)], \quad (23)$$

where $\mathbf{G}_\mathrm{r}^{(tr)}$ are the distance compensated reproduction gains, which, assuming that the sources are projected onto a sphere, is given as

$$\mathbf{G}_\mathrm{r}^{(tr)} = [\frac{r_0}{r_{l,1}}\mathbf{g}^{(tr)}(\gamma_1), ..., \frac{r_0}{r_{l,K}}\mathbf{g}^{(tr)}(\gamma_K)]. \quad (24)$$

Note that for arbitrary projections, r_0 instead becomes source direction-dependent. Furthermore, since the framework operates in the time-frequency domain, the distance-based gain compensation terms can be frequency-dependent. This therefore allows the framework to also accommodate near-field/proximity effects [26, 27], or other more creative distance-dependent filters.

Note that this effect, in this parametric context, was evaluated perceptually using the COMPASS framework in [28] with positive results. It has also previously been explored in [11, 12] using the first-order DirAC model, which based the synthesis on only the omni-directional component. Therefore, two key differences are that the framework presented here includes a source beamforming stage, which can improve source signal isolation, and the method can also accommodate multiple simultaneous sources. The acoustic zooming techniques described in [13, 14] are then examples of a less explicitly defined translation, which are based instead on the manipulation of the DirAC diffuseness parameter. They operate based on the knowledge that a reduction in the reverberation level, also has the effect of perceptually bringing sound sources closer to the listener; i.e. reducing the perceived externalisation of the sources. Therefore, similarly, the presented framework may also be used for acoustical zooming in this way, by instead manipulating the source and ambient stream balance; as described in Section 4.1. Note that in [13, 14], the technique was intended for use in teleconferencing applications, enabling the zooming-in function on the video to be accompanied by the respective acoustical zooming. Due to the recent rise in popularity of over-the-web streaming of Ambisonic sound scenes [29, 30], user controllable acoustical zooming methods are becoming more widespread.

4.3. Spatial editing

At each time frame, the parametric framework provides up to K number of DoA estimates. However, in a multi-directional model such as COMPASS, the number of DoA estimates can vary across time frames, and the order in which the estimates are presented to the beamforming and spatialisation stages can also change. For reproduction purposes, this limitation of not being able to associate DoA estimates with their respective sources across time frames, matters little. However, for spatial editing applications, decomposing the sound-field into its broad-band sound objects (or "stems"), along with their corresponding unique identification numbers, can be particularly useful. This object-based decomposition can either be based on beamformers that are steered towards manually defined user markers, or automatically through temporal data association methods. Examples of source tracking algorithms on Ambisonic signals for multiple simultaneous sources include [31, 32].

The resulting broad-band source objects can then be re-balanced, re-ordered, and/or re-directed, prior to reconstructing the sound scene. This also allows traditional single-channel processing methods, such as delays, equalisers, phasers, and dynamic range compressors, to be applied to specific source objects within the overall sound scene independently. Additionally, the broad-band beamformer signals may be subjected to frequency-dependent spatial post-filters, for example [33, 9], in order to improve their spatial selectivity and deactivate them during periods of source inactivity. Similar approaches, using multi-channel Wiener filters on the Ambisonic signals as an alternative, have also been explored in [34].

4.4. Directional transformations

As mentioned in Section 4.2, when rendering the source stream, the source beamformer directions may remain informed by the estimated DoAs, but the reproduction directions can instead be based on the estimated DoAs after they are subjected to a directional transformation. Additionally, a gain factor may also be applied to the spatialisation gains, which extends the flexibility of the framework. Based on these options, there are many directional transformations that are easily imaginable. These include the traditional

sound scene mirroring and rotations, which are already well defined as linear Ambisonic operations. Other existing Ambisonic operations, such as warping and directional loudness modifications, are also possible. However, importantly, in this parametric context, these effects are only applied to the directional source stream, with the ambient components remaining unchanged. Furthermore, parametric methods are less bound by the input SH order, and may yield improved spatial resolution over their linear counterparts. Linear Ambisonic transformations, such as directional filtering and warping [4], can also result in an output order that is significantly higher than the input order; thus, requiring many more channels than a parametric equivalent. Some directional transformations are also unique to parametric methods, including: direction randomisation, and source deletion if it is within a user defined field of view.

As an example, a simple, yet effective, directional transformation is now described. Here, user defined marker directions, given as Cartesian coordinates $\mathbf{u} \in \mathbb{R}^3$, first serve as control points; from which a direction-dependent biasing term is derived, which pulls nearby DoA estimates towards these control points. The effect can be considered as directional warping or "focusing". A maximum operating range is defined as θ_0, beyond which no biasing is applied, and the angle θ_{rot} at which a DoA estimate, $\mathbf{v}_\mathrm{s} \in \mathbb{R}^3$, needs to be rotated towards the user defined control point is computed as

$$\theta_{\mathbf{vu}} = \cos^{-1}(\mathbf{v}_\mathrm{s}^\mathrm{T}\mathbf{u}), \tag{25}$$

$$\theta_{rot} = \theta_{\mathbf{vu}}(1 - \min[\theta_{\mathbf{vu}}/\theta_0, 1]^\alpha), \tag{26}$$

where $\alpha > 1$ determines how drastically the estimates are pulled towards the control point. The Rodrigues equation may then be used to apply the rotation along the surface of the unit sphere as

$$\mathbf{v}_\mathrm{r} = \mathbf{v}_\mathrm{s}\cos(\theta_{rot}) + (\mathbf{k}\times\mathbf{v}_\mathrm{s})\sin(\theta_{rot}) + \mathbf{k}(\mathbf{k}\mathbf{v}_\mathrm{s}^\mathrm{T})[1-\cos(\theta_{rot})], \tag{27}$$

where $\times$ denotes the cross-product operation and $\mathbf{k} = \mathbf{v}_\mathrm{s} \times \mathbf{u}$. Note that this is performed individually per DoA and recursively per control point. An example of this particular directional transformation is depicted in Fig. 3.

Finally, by stacking the directionally transformed $\mathbf{v}_\mathrm{r}$ angles into $\mathbf{\Gamma}_\mathrm{r}^{(dt)}$, the modified source stream can be obtained as

$$\mathbf{z}_\mathrm{s}(t,f) = \mathcal{S}_\mathrm{s}[\mathbf{a}(t,f), \tilde{\mathbf{\Gamma}}_\mathrm{s}(t,f), \mathbf{\Gamma}_\mathrm{r}^{(dt)}(t,f), \mathbf{G}_\mathrm{r}(t,f)]. \tag{28}$$

Note that when user defined markers are used, this directional biasing acts as a spatial focusing effect. However, if the control points are informed by a source tracker, as described in Section 4.3, this approach can instead act as a means of stabilising the source rendering stream during parametric reproduction; which may be used as an alternative to long temporal averaging functions.

4.5. Spatial morphing and modulations

In traditional music production, it is common to process signals based on the analysed parameters of other signals. For example, a dynamic range compressor can be used to attenuate a bass guitar signal based on the gain factors derived from analysing a kick drum signal. This either allows the signals to combine more cohesively, or the processing can be exaggerated to serve as an audio effect. The same principles may be used in this spatial audio context, where instead it is the spatial parameters analysed from one sound scene that may be used to synthesise a second sound scene.

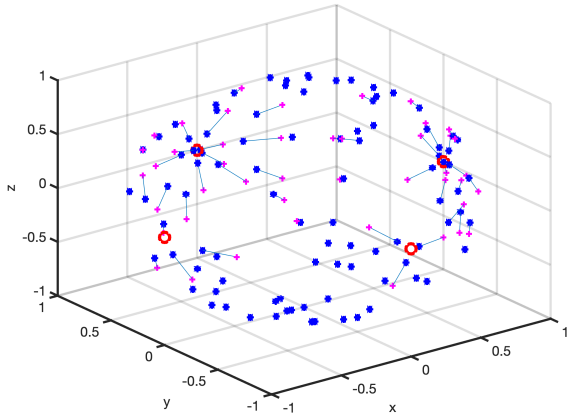

Figure 3: An example of the described direction warping function. The original DoA estimates are depicted as magenta coloured crosses, and as blue asterisks after they have been biased towards the red circle control points ($\alpha = 1.35$).

This has been referred to as spatial morphing (carrier/modulator analogy) before in [10]. The analysis is first conducted based on scene A as

$$\tilde{\mathbf{\Gamma}}_\mathrm{s}^{(A)}(t,f) = \mathcal{A}[\mathbf{a}^{(A)}(t,f)], \tag{29}$$

with the synthesis applied to a different scene B, but using the estimated parameters from scene A, as

$$\mathbf{z}_\mathrm{s}(t,f) = \mathcal{S}_\mathrm{s}[\mathbf{a}^{(B)}(t,f), \tilde{\mathbf{\Gamma}}_\mathrm{s}^{(A)}(t,f), \mathbf{\Gamma}_\mathrm{r}^{(A)}(t,f), \mathbf{G}_\mathrm{r}^{(A)}(t,f)], \tag{30}$$

$$\mathbf{z}_\mathrm{d}(t,f) = \mathcal{S}_\mathrm{d}[\mathbf{a}^{(B)}(t,f), \tilde{\mathbf{\Gamma}}_\mathrm{s}^{(A)}(t,f)]. \tag{31}$$

If scene A and B are the same, then the processing reverts to the standard COMPASS rendering. Other spatial morphing and modulations are also easily imaginable based on mixing and matching the reproduction directions and/or reproduction gains for the two scenes. It is also possible to modulate only the source stream or the ambient stream.

5. PRACTICAL IMPLEMENTATIONS

This section describes a number of real-time VST audio plug-ins, which were developed in order to demonstrate the spatial audio effects and sound-field modifications that are discussed in Section 4. All of the plug-ins employ the modified COMPASS framework detailed in Section 3, and were developed using JUCE[2] and the Spatial_Audio_Framework[3].

5.1. BinauralVR

As with the existing COMPASS Binaural decoder plug-in described in [35], the BinauralVR decoder plug-in offers the ability to alter the balance between the source and ambient streams, as covered in Section 4.1, using frequency-dependent sliders. Additionally, the BinauralVR decoder supports listener translation

[2]https://github.com/juce-framework/JUCE

[3]https://github.com/leomccormack/Spatial_Audio_Framework

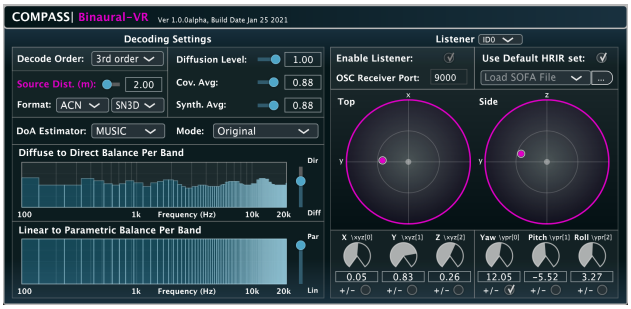

Figure 4: The user interface for the binauralVR plug-in, which supports listener translation around a single receiver and multiple simultaneous listeners.

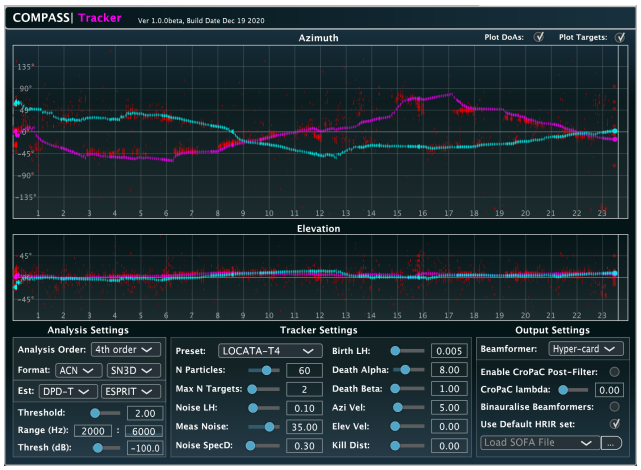

Figure 5: The user interface for the acoustic tracking and beamforming plug-in. The DoA estimates fed to the tracker are depicted in red, and the two target trajectories in magenta and cyan.

around a single receiver, as described in Section 4.2. The user must first select the assumed distance of the sources. For simplicity, it is assumed that all sources are projected onto the surface of a sphere. The plug-in may then be informed of the listener position and orientation, either via its user interface, or by sending the Cartesian coordinates and rotation angles outputted by an external tracking device; such as a virtual or augmented reality headset.

A single instance of the plug-in also allows multiple listeners to experience the same scene, but based on their own position and head orientation. Since the analysis and synthesis stages are decoupled, the analysis need only be conducted once. Therefore, the framework represents an efficient means of delivering dynamic renderings for multiple simultaneous listeners, which has not been explored before in this parametric reproduction context. Extending the plug-in for multi-receiver operation, where no assumption of the source distance is required, is a topic of future work.

5.2. Tracker

The Tracker plug-in serves as a demonstration of how multi-source tracking may be used within the parametric spatial audio context. The tracking algorithm employed internally by the plug-in, which can adapt to sound sources that vary in number and position over

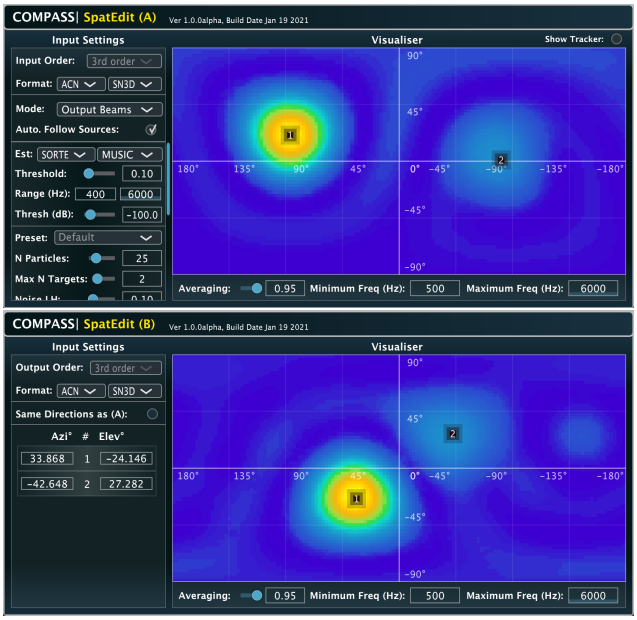

Figure 6: The user interface for the first SpatEdit plug-in instance (top), and the second instance (bottom). Between the two instances, the user has access to either the source beamformer or residual signals, which they may manipulate as they choose prior to the reconstruction of the SH scene.

time, is described in detail in [32]. Since the target velocities are also considered by the tracker model, it can still distinguish between sources whose trajectories cross-over one another. It visualises the azimuth and elevation angles of the DoA estimates and the tracked target trajectories over a 24 second history on its user interface; as shown in Fig. 5.

The plug-in can also steer a beamformer towards each target direction and output their signals (one target signal per output channel), which is akin to decomposing the scene into its individual broad-band "stems". The corresponding target directions are also accessible via the plug-in's automation data, which allows the stems to also be optionally spatialised in their respective original (or transformed) directions by other plug-ins. The spatial post-filter described in [33] is also included, in order to improve the beamformer's performance in noisy/reveberant environments.

5.3. SpatEdit

The SpatEdit plug-in is intended to be used with two instances. The first instance of the plug-in allows the user to place markers on an equirectangular representation of the sphere. Alternatively, the markers can automatically follow the directions of sound sources, through use of the tracker, which is fed by the analysed DoA estimates. The DoA estimates are then replaced by these marker directions, and the source and ambient beamforming is conducted as normal. The source beamformer signals are then outputted by the first instance of the plug-in, where the user can then apply any conventional single-channel audio effect, re-balance their levels, or re-order the beamformer signals. These manipulated beamformer signals are then passed to the second instance of the plug-in, which also receives the residual signals from the first plug-in instance in-

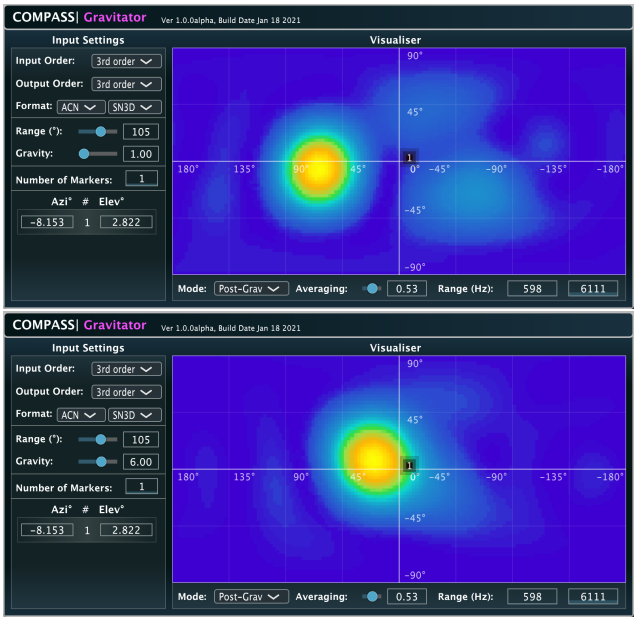

Figure 7: The user interface for the Gravitator plug-in, when $\alpha = 1$ (top; i.e. not being applied) and when $\alpha = 6$ (bottom).

ternally, and the COMPASS synthesis is conducted to obtain the output SH signals. Alternatively, the residual stream signals may be outputted by the first plug-in instance instead, which therefore allows conventional linear Ambisonics transformations to be applied to only the ambient parts of the scene. Note that to aid marker placement, the sound-field is also visualised, based on the steered response power (SRP) approach [36], and projected behind the markers on the same equirectangular spherical window.

5.4. Gravitator

The Gravitator plug-in implements the directional transformation example described in Section 4.4, which results in the directional focusing of source components within the scene towards user marker directions. The user defines these marker directions using an equirectangular representation of the sphere, along with the maximum operating range (θ_0) and "gravity" (α). The SRP method is employed to visualise the sound-field using the same equirectangular representation, for either pre- or post-effect processing. The plug-in output is also SH signals, but of an optionally higher order than that of the input order; therefore, if the gravity parameter is set to 1, then the processing reverts back to the standard COMPASS Upmixer plug-in [35].

5.5. SideChain

The SideChain plug-in serves as an example of direct-to-diffuse balance manipulations, as described in Section 4.1, and spatial modulation, as described in Section 4.5. The plug-in accepts one Ambisonic sound scene using input channels 1-16, and another Ambisonic sound scene using input channels 17-32. The synthesis of one scene is then conducted based on the analysed parameters of the other scene. The output of the plug-in is then SH signals of optionally higher order than that of the input signals.

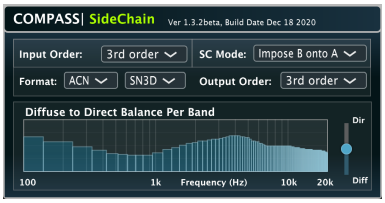

Figure 8: The user interface for the SideChain plug-in, when using the spatial parameters from scene B to process scene A.

6. CONCLUSIONS

This paper has presented a general framework for creating parametric spatial audio effects. The framework is based on the Coding and Multi-Parameterisation of Ambisonic Sound Scenes (COMPASS) method, which employs analysed spatial parameters to divide the input Ambisonic sound scene into its multiple source and ambient components. It is demonstrated, through formulations and VST audio plug-in implementations, that the framework can represent a convenient and intuitive means of developing spatial audio effects based on simple manipulations of the spatial parameters.

7. ACKNOWLEDGMENTS

This research has received funding from the Aalto University Doctoral School of Electrical Engineering.

8. 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] Joseph Ivanic and Klaus Ruedenberg, "Rotation matrices for real spherical harmonics. Direct determination by recursion," *The Journal of Physical Chemistry A*, vol. 102, no. 45, pp. 9099–9100, 1998.

[3] Hannes Pomberger and Franz Zotter, "Warping of 3D ambisonic recordings," in *Proceedings of the 3rd International Symposium on Ambisonics & Spherical Acoustics*, 2011.

[4] Matthias Kronlachner and Franz Zotter, "Spatial transformations for the enhancement of ambisonic recordings," in *Proceedings of the 2nd International Conference on Spatial Audio, Erlangen*, 2014.

[5] Pierre Lecomte, Philippe-Aubert Gauthier, Alain Berry, Alexandre Garcia, and Christophe Langrenne, "Directional filtering of ambisonic sound scenes," in *2018 AES International Conference on Spatial Reproduction-Aesthetics and Science*. Audio Engineering Society, 2018.

[6] 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*, pp. 89–138. John Wiley & Sons, 2017.

[7] Svein Berge and Natasha Barrett, "High angular resolution planewave expansion," in *Proceedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, 2010, pp. 6–7.

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

[9] Leo McCormack and Symeon Delikaris-Manias, "Parametric first-order ambisonic decoding for headphones utilising the cross-pattern coherence algorithm," in *EAA Spatial Audio Signal Processing Symposium*, 2019, pp. 173–178.

[10] Archontis Politis, Tapani Pihlajamäki, and Ville Pulkki, "Parametric spatial audio effects," in *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12)*, 2012.

[11] Tapani Pihlajamäki and Ville Pulkki, "Projecting simulated or recorded spatial sound onto 3D-surfaces," in *Audio Engineering Society Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio*, 2012.

[12] Ville Pulkki, Archontis Politis, Tapani Pihlajamäki, and Mikko-Ville Laitinen, "Spatial sound scene synthesis and manipulation for virtual reality and audio effects," in *Parametric Time-Frequency Domain Spatial Audio*, pp. 347–361. John Wiley & Sons, 2017.

[13] Richard Schultz-Amling, Fabian Kuech, Oliver Thiergart, and Markus Kallinger, "Acoustical zooming based on a parametric sound field representation," in *Audio Engineering Society Convention 128*. Audio Engineering Society, 2010.

[14] Markus Kallinger, Giovanni Del Galdo, Fabian Kuech, and Oliver Thiergart, "Dereverberation in the spatial audio coding domain," in *Audio Engineering Society Convention 130*. Audio Engineering Society, 2011.

[15] Earl G Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*, Academic press, 1999.

[16] Heinz Teutsch, *Modal array signal processing: principles and applications of acoustic wavefield decomposition*, vol. 348, Springer, 2007.

[17] Franz Zotter and Matthias Frank, *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*, Springer Nature, 2019.

[18] 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.

[19] Franz Zotter, Hannes Pomberger, and Markus Noisternig, "Energy-preserving ambisonic decoding," *Acta Acustica united with Acustica*, vol. 98, no. 1, pp. 37–47, 2012.

[20] Markus Zaunschirm, Christian Schörkhuber, and Robert Höldrich, "Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3616–3627, 2018.

[21] Christian Schörkhuber, Markus Zaunschirm, and Robert Höldrich, "Binaural rendering of ambisonic signals via magnitude least squares," in *Proceedings of the DAGA*, 2018, vol. 44, pp. 339–342.

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

[23] Ralph Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[24] Byeongho Jo and Jung-Woo Choi, "Parametric direction-of-arrival estimation with three recurrence relations of spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 145, no. 1, pp. 480–488, 2019.

[25] Byeongho Jo, Franz Zotter, and Jung-Woo Choi, "Extended Vector-Based EB-ESPRIT Method," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

[26] Richard O Duda and William L Martens, "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, 1998.

[27] Jérôme Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format," in *23rd AES International Conference: Signal Processing in Audio Recording and Reproduction*. Audio Engineering Society, 2003.

[28] Maximilian Kentgens, Andreas Behler, and Peter Jax, "Translation of a higher order ambisonics sound scene based on parametric decomposition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 151–155.

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

[30] Thomas Deppisch, Nils Meyer-Kahlen, Benjamin Hofer, Tomasz Latka, and Tomasz Zernicki, "HOAST: A higher-order ambisonics streaming platform," in *Audio Engineering Society Convention 148*. Audio Engineering Society, 2020.

[31] Srđan Kitić and Alexandre Guérin, "TRAMP: Tracking by a realtime ambisonic-based particle filter," in *Proceedings of the LOCATA Challenge Workshop - Satellite Event IWAENC*, 2018.

[32] Leo McCormack, Archontis Politis, Simo Särkkä, and Ville Pulkki, "Real-time tracking of multiple acoustical sources utilising Rao-Blackwellised particle filtering," in *Accepted at the 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021.

[33] Symeon Delikaris-Manias, Juha Vilkamo, and Ville Pulkki, "Signal-dependent spatial filtering based on weighted-orthogonal beamformers in the spherical harmonic domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1511–1523, 2016.

[34] Alexis Favrot and Christof Faller, "Wiener-based spatial B-Format equalization," *Journal of the Audio Engineering Society*, vol. 68, no. 7/8, pp. 488–494, 2020.

[35] Leo McCormack and Archontis Politis, "SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods," in *Audio Engineering Society Conference: International Conference on Immersive and Interactive Audio*, 2019.

[36] Maximo Cobos, Amparo Marti, and Jose J Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Processing Letters*, vol. 18, no. 1, pp. 71–74, 2010.