PARAMETRIC FIRST-ORDER AMBISONIC DECODING FOR HEADPHONES UTILISING THE CROSS-PATTERN COHERENCE ALGORITHM

Leo McCormack and Symeon Delikaris-Manias
Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland
leo.mccormack@aalto.fi

ABSTRACT

Binaural ambisonics decoding is a means of reproducing a captured or synthesised sound-field, as described by a spherical harmonic representation, over headphones. The majority of ambisonic decoders proposed to date are based on a signal-independent approach; operating via a linear mapping between the input spherical harmonic signals and the output binaural signals. While this approach is computationally efficient, an impractically high input order is often required to deliver a sufficiently accurate rendition of the original spatial cues to the listener. This is especially problematic, as the vast majority of commercially available Ambisonics microphones are first-order, which ultimately results in numerous perceptual deficiencies during reproduction. Therefore, in this paper, a signal-dependent and parametric binaural ambisonic decoder is proposed, which is specifically intended to reproduce first-order input with high perceptual accuracy. The proposed method assumes a sound-field model of one source and one non-isotropic ambient component per narrow-band. It then employs the Cross-Pattern Coherence (CroPaC) post-filter, in order to segregate these components with improved spatial selectivity. Listening test results indicate that the proposed method, when using first-order input, performs similarly to third-order Ambisonics reproduction.

1. INTRODUCTION

Reproduction of synthesised or captured sound-fields is an important component in many immersive audio applications, where flexibility, in terms of both content generation and playback setup, is highly favoured. Methods formulated in the spherical harmonic domain (SHD) [1] are often well-suited to this task, as the recording and reproduction operations may be decoupled; with spherical harmonic signals serving as an intermediary. This SHD-based ecosystem for sound-field capture and reproduction is popularly known as Ambisonics [2], where the generation of spherical harmonic signals and the subsequent reproduction of the sound scene that they describe, is referred to as ambisonic encoding and decoding, respectively. Regarding the latter, currently proposed decoders may be loosely categorised as either non-parametric (signal-independent) or parametric (signal-dependent). Non-parametric binaural reproduction relies on a complex, frequency-dependent, and linear mapping of the input signals to the binaural channels. Whereas parametric methods operate by imposing a set of assumptions regarding the composition of the sound-field and are signal-dependent. Methods that fall within this latter category often rely on the extraction of perceptually meaningful parameters in the time-frequency domain, with the aim of mapping input signals to the binaural channels in a more informed manner [3]. This paper is primarily concerned with parametric reproduction of first-order spherical harmonic input over headphones. A binaural decoder is proposed for this task, which utilises the Cross-Pattern Coherence (CroPaC) [4] spatial post-filter; in order to segregate the sound-field into one source and one non-isotropic ambient component per time-frequency tile during the analysis stage. The method then employs the optimal mixing approach described in [5], to synthesise the output binaural signals.

1.1 Non-parametric binaural ambisonics decoding

Binaural ambisonics decoding is conducted via the application of a matrix of filters, which appropriately maps the input spherical harmonic signals to the binaural channels in a linear manner. Therefore, no time-varying distortions are introduced into the output signals. The decoding filters may be derived by approximating the directivity patterns of the listener's head-related transfer functions (HRTF), using the spherical harmonic basis functions, in a least-squares (LS) sense. However, in order to sufficiently approximate and reproduce these complicated directional patterns, a dense grid of HRTF measurements and a high input order is required; often in the range of 15-20th order. For practical reasons, the input order is typically truncated to a much lower order than that of the spatial order of the HRTF measurement grid. This, in turn, results in direction-dependent timbral colourations in the binaural signals. In addition, Ambisonics reproduction is inherently limited by the spatial resolution of the input format. For lower-orders, this has been found to exhibit numerous perceptual deficiencies, including: localisation ambiguity, comb-filtering effects, poor externalisation, and a loss of envelopment [6–10].

Timbral colourations due to input order truncation especially affect high-frequencies, since the high-frequency energy is predominantly concentrated in the higher-order components. This loss of energy may be compensated for via diffuse-field equalisation [11]. However, this loss of high-frequency energy is largely due to the miss-match between the input order and the spatial order of the measurement grid, which is directly proportional to its density. Therefore, rather than applying post-equalisation filters, one may simply reduce the number of points in the HRTF measurement grid, such that its spatial order is more in-line with that of the input order; as suggested in [12]. This approach is often referred to as Spatial Re-sampling (SPR) or virtual loudspeaker decoding. In this case, rather than assigning high-frequency energy to higher-order components and subsequently discarding it, due to order truncation, the energy is instead aliased back into the lower-order components and preserved. However, while this approach improves upon the perceived timbral short-comings of lower-order binaural Ambisonic reproduction, it does not eliminate them, nor does it address the spatial deficiencies of the method.

The localisation ambiguities associated with lower-order binaural Ambisonic reproduction are due to a degradation of the reproduced binaural cues. There are two key causes for this. The first is due to the inherent low input spatial resolution, which leads to erroneously high signal coherence between the output channels; when generated in a linear manner. Whereas the second is due to the LS decoder itself, as it is unable to sufficiently fit the lower-order spherical harmonic patterns to the highly directive HRTF patterns. To address this latter limitation, an alternative method was proposed in [13], which conducts preliminary time-alignment of the Head-related impulse responses (HRIRs) and performs the LS fitting with an additional diffuse-field coherence constraint. The method essentially exploits prior knowledge of the bandwidth in which the inter-aural level differences (ILDs) are the dominant localisation cues; which is above approximately 1.5 kHz, as described by the Haas effect [14]. By discarding the phase information of the HRTFs at frequencies above 1.5 kHz, the LS fitting instead prioritises the delivery of the correct magnitude responses; rather than the phase. Thus it ultimately yields improved ILD cues and diminished inter-aural time difference (ITD) cues; but in a frequency range where ILD cues are more important for localisation. The same principle was also later employed in [15]. However, while these aforementioned approaches yield considerable improvements over traditional decoders, as shown with formal listening tests [13, 16], their performance with first-order input still deviates from that of higher-orders and directly binauralised scenes. This is especially problematic as the vast majority of commercially available Ambisonic microphones and available content are first-order.

1.2 Parametric binaural decoding

The inherent perceptual limitations associated with lower-order Ambisonics are primarily as a result of the erroneously high coherence between the output channels. In order to overcome these limitations, signal-dependent and parametric alternatives have been proposed [17–22]. These methods employ a sound-field model, which lays out a set of assumptions regarding the composition of the sound-field. The methods operate by extracting perceptually meaningful parameters in the time-frequency domain, and often employ dedicated rendering techniques for different components. The two main challenges when designing a parametric method are therefore: 1) identifying a perceptually robust sound-field model, and 2) employing the appropriate signal processing techniques in order to realise the model, with minimal artefacts incurred.

The most well-known and established parametric reproduction method is Directional Audio Coding (DirAC) [17], which employs a sound-field model consisting of one plane-wave and one diffuseness estimate per time-frequency tile. These parameters are derived from the active-intensity vector, in the case of first-order input. The plane-wave components are panned directly to the loudspeakers using Vector-base Amplitude Panning (VBAP) [23], and the diffuse components are sent to all loudspeakers and decorrelated. More recent formulations of DirAC also allow for multiple plane-wave and diffuseness estimates via spatially-localised active-intensity vectors, using higher-order input [18, 19]. In [24], a post-filter was proposed, which adaptively mixes between linearly decoded output and DirAC rendered outputs, in order to improve the output signal fidelity.

High Angular Resolution plane-wave Expansion (HARPEX) [20], is another example of a parametric method, which operates by extracting two plane-wave components per frequency using first-order input. The Sparse-Recovery method [21] extracts a number of plane-waves, which corresponds to up to half the number of input channels of arbitrary order. The COding and Multi-Parameterisation of Ambisonic Sound Scenes (COMPASS) method [25] also extracts multiple source components; up to half the number of input channels. However, it employs an additional residual stream that encapsulates the remaining diffuse and ambient components in the scene. An alternative parameterisation of the sound-field was also presented in [26, 27], which circumvents the modelling of the sound-field with conventional parameters, such as source direction or diffuseness. It considers only the perceived quality of the individual output channels and the perceived quality of the spatial attributes of the reproduction. In [22] an adaptive LS-binaural decoder was proposed, which constrains the covariance matrix of the decoded audio to reflect that of the covariance matrix of the listeners HRTFs. The method does not require explicit estimation of the direct/diffuse balance, nor does it need to employ de-correlation.

1.2.1 Parametric binaural decoding with optimal mixing

The main point of criticism regarding parametric reproduction methods is the incursion of time-varying artefacts. These occur either due to the input scene not conforming to the assumed sound-field model, or due to the rendering techniques not being sufficiently robust. Therefore, in [28], an *optimal mixing* technique was proposed, which attempts to synthesise the output using a linear combination of linearly decoded prototype signals, as much as possible; thus retaining much of the high single-channel fidelity in the output. The method then employs decorrelation only if the output inter-channel dependencies still deviate from the target, thus mitigating potential decorrelation artefacts; such as the temporal smearing of transients.

The approach is formulated in the covariance domain and relies on the construction of time-varying narrow-band target covariance matrices, which are dictated by the sound-field model of the parametric method. It then employs traditional linearly decoded audio as a prototype. The approach has been employed previously in [19] using the DirAC sound-field model, and also in the closed-form solution of [22] for binaural reproduction. However, in principle, any sound-field model may employ this optimal mixing approach for the synthesis stage.

1.3 Motivation for this work

Further development of first-order ambisonic decoders is of particular interest; given the prevalence of first-order commercially available Ambisonic microphone arrays and material accessible on the internet. Recent advancements have shown improvements in the perceived performance of linear binaural ambisonic decoding [13, 15]. However, these decoders still rely on input orders which may be considered impractical for wide-spread adoption today. Especially given the quadratic growth in the number of channels (and microphones) with increasing order, leading to more expensive arrays and higher bandwidths. Therefore, the need for robust signal-dependent parametric alternatives for lower-orders is well established.

In this paper, a first-order binaural decoder is proposed [1], which employs a parametric sound-field model that assumes one source and one directional ambient component per time-frequency tile. The model is inspired by the multi-source and non-isotropic ambient model employed by the COMPASS method [25]. However, given the low-resolution of first-order input, the proposed method employs an additional CroPaC spatial post-filter [4] to better isolate the source component, and an inverse CroPaC post-filter to obtain the ambient component. The proposed method also extracts instantaneous direction-of-arrival (DoA) estimates via steered-response power (SRP) activity-maps and peak-finding; thus forgoing the need for long temporal averaging of input covariance matrices, as required by sub-space alternatives. The method also employs the optimal mixing solution of [5], in order to minimise the amount of decorrelated signal energy in the output, and synthesise the target inter-channel dependencies in a linear manner, as much as possible.

1.4 Organisation of the paper

Section 2 describes linear binaural ambisonic decoding, which is employed as a basis for the proposed method. Section 3 details the proposed method. Subjective evaluation of the proposed algorithm, through listening tests, is described in Section 4, and Section 5 concludes the paper.

[1] A VST audio plug-in of the proposed decoder may be found here: http://research.spa.aalto.fi/publications/papers/sasp19-parametric/

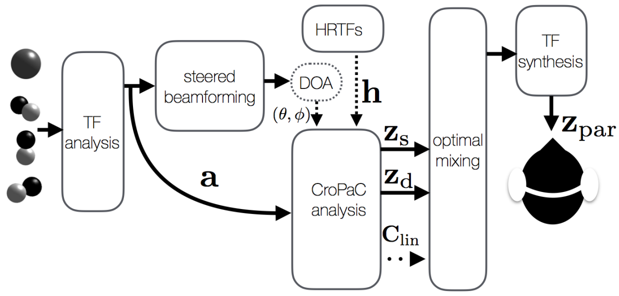

Figure 1: Overall block diagram of the proposed method.

2. LINEAR BINAURAL AMBISONICS DECODING

Due to the signal- and frequency-dependent nature of the proposed algorithm (described in Section 3) it is assumed that the input ambisonic signals have been transformed into the time-frequency (t, f) domain; where t and f denote the time and frequency indices, respectively. The ambisonic signals, $\mathbf{a}$, may be synthesised by mapping monophonic signals onto spherical harmonic basis functions or captured utilising a microphone array with subsequent suitable encoding

$$\mathbf{a} = [a_{00}, a_{1(-1)}, a_{10}, \dots, a_{N(N-1)}, a_{NN}]^{\mathrm{T}} \in \mathbb{C}^{(N+1)^2 \times 1}, \tag{1}$$

where a_{nm} are the individual ambisonic signals for each order, n, and degree, m, up to the maximum order, N. It is assumed, henceforth, that the input ambisonic signals conform to the ortho-normalised (N3D) and Ambisonic Channel Numbering (ACN) conventions.

Ambisonic signals may be decoded for headphone playback as

$$\mathbf{z}_{\mathrm{lin}}(t, f) = \mathbf{D}_{\mathrm{bin}}(f)\mathbf{a}(t, f), \tag{2}$$

where $\mathbf{z}_{\mathrm{lin}}(t, f) \in \mathbb{C}^{2\times 1}$ are the output binaural signals, and $\mathbf{D}_{\mathrm{bin}}(f) \in \mathbb{C}^{2\times (N+1)^2}$ is a binaural ambisonic decoding matrix, derived using one of a number of approaches [12, 13, 15].

3. PROPOSED PARAMETRIC BINAURAL AMBISONICS DECODER

The proposed first-order decoder employs a sound-field model comprising one source component and one non-isotropic ambient component per time-frequency tile. The method first estimates the source DoA via steered-response power (SRP) beamforming and subsequent peak-finding. The source stream is then segregated by steering a beamformer toward the estimated DoA, and employing an additional CroPaC post-filtering operation to improve its spatial selectivity. The ambient stream is then simply the residual, once the source component has been subtracted from the input sound-field. The two streams are then binauralised and fed into an optimal mixing unit, along with the ambisonic prototype covariance matrix, in order to generate the binaural output. A block diagram of the proposed method is depicted in Fig. 1.

3.1 Analysis

3.1.1 DoA estimation

A DoA estimator based on SRP beamforming and peak-finding [29] is suggested for the proposed rendering method. Note that the chosen scanning grid should preferably take the angular resolution of human hearing into account [14]. This approach yields instantaneous estimates of the source direction, $(\theta_\mathrm{s}, \phi_\mathrm{s})$, which is in keeping with the instantaneous CroPaC post-filter values. Therefore, the need for temporal averaging of input covariance matrices is avoided in the analysis stage, as would be required by subspace-based alternatives.

3.1.2 CroPaC post-filter

The cross-correlation between the omni and dipole may be utilised as a spatial post-filter [4]

$$G(t, f) = \frac{2}{\sqrt{3}} \Re[a_{00}{}^*(t, f) a_{11}(t, f)], \tag{3}$$

where $\Re$ denotes the real operator and * denotes the complex conjugate. This value is then normalised with the input sound-field energy and half-wave rectified

$$\hat{G}(t, f) = \max \left[\frac{G(t, f)}{|a_{00}(t, f)|^2 + \sum_{-1}^{1} |a_{1m}(t, f)/\sqrt{3}|^2}, \lambda \right], \tag{4}$$

where $\lambda \in [0, ..., 1]$ is a parameter which influences the severity of the post-filter, similarly to the spectral floor of a traditional post-filter. Note that the spatial selectivity of the CroPaC post-filter is sharper than a conventional first-order beam pattern, due to the multiplication (rather than summation) of the two signals in (3).

The CroPaC spatial filter may be steered in the source direction, by first rotating the spherical harmonic signals using an appropriate rotation matrix, $\mathbf{M}_\mathrm{rot}(\theta_\mathrm{s}, \phi_\mathrm{s}) \in \mathbb{R}^{(N+1)^2 \times (N+1)^2}$, as

$$\mathbf{a}_\mathrm{rot}(t, f) = \mathbf{M}_\mathrm{rot}(\theta_\mathrm{s}, \phi_\mathrm{s}) \mathbf{a}(t, f), \tag{5}$$

where $\mathbf{a}_\mathrm{rot}(t, f) \in \mathbb{C}^{(N+1)^2 \times 1}$ are the resulting rotated signals, which are then employed for the post-filter estimation (3). For details regarding the calculation of this rotation matrix, the reader is directed to [30].

3.2 Synthesis

The source stream, $\mathbf{z}_\mathrm{s}(t, f)$, is obtained by directly applying the HRTF gains to the extracted source signal as

$$\mathbf{z}_\mathrm{s}(t, f) = \mathbf{h} \frac{\hat{G}(t, f)}{(N+1)^2} \mathbf{w}^\mathrm{T} \mathbf{a}(t, f), \tag{6}$$

where $\mathbf{h} \in \mathbb{C}^{2 \times 1}$ and $\mathbf{w} \in \mathbb{R}^{(N+1)^2 \times 1}$ are HRTF gains and static beamforming weights [1], corresponding to the analysed DoA at each time-frequency tile, respectively. A visual depiction of the improved spatial selectivity of the source beamformer with the post-filter is given in Fig. 2.

The ambient stream, $\mathbf{z}_\mathrm{d}(t, f)$, is then obtained by subtracting the source signal from the input scene, and decoding it to headphones using a binaural ambisonic decoder

$$\mathbf{z}_\mathrm{d}(t, f) = \mathbf{D}_\mathrm{bin} [\mathbf{a}(t, f) - \frac{\hat{G}(t, f)}{(N+1)^2} \mathbf{y} \mathbf{w}^\mathrm{T} \mathbf{a}(t, f)], \tag{7}$$

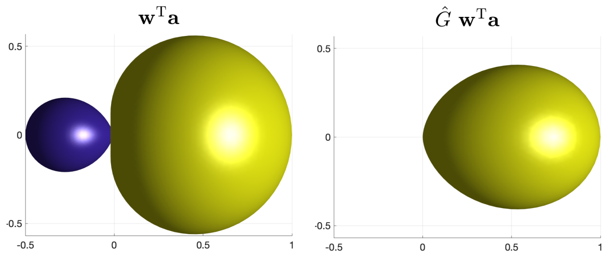

Figure 2: A first-order hyper cardioid beamformer without (left) and with (right) the CroPaC post-filter ($\lambda = 0$).

where $\mathbf{y} \in \mathbb{R}^{(N+1)^2 \times 1}$ are the spherical harmonic weights for the analysed DoA.

Optionally, the ambient stream may be decorrelated, in order to further minimise the inter-channel coherence between the binaural channels

$$\hat{\mathbf{z}}_\mathrm{d}(t, f) = \mathcal{D} [\mathbf{z}_\mathrm{d}(t, f)] \tag{8}$$

where $\mathcal{D}[.]$ denotes a decorrelation operation on the enclosed signals.

3.2.1 Optimal mixing

Rather than summing the source (6) and ambient (8) streams together, to acquire the binaural output, an alternative synthesis approach is suggested. This approach is based on the covariance domain framework, termed here as *optimal mixing*, and is described in [5, 28]. The method employs linearly decoded signals, $\mathbf{z}_\mathrm{lin}(t, f)$, as a prototype, which has the following base-line covariance matrix (note that the time-frequency indices have been omitted for the brevity of notation)

$$\mathbf{C}_\mathrm{lin} = \mathrm{E}[\mathbf{z}_\mathrm{lin} \mathbf{z}_\mathrm{lin}^\mathrm{H}] = \mathbf{D}_\mathrm{bin} \mathbf{C}_\mathrm{a} \mathbf{D}_\mathrm{bin}^\mathrm{H}, \tag{9}$$

where $\mathbf{C}_\mathrm{a} = \mathrm{E}[\mathbf{a}\mathbf{a}^\mathrm{H}] \in \mathbb{C}^{(N+1)^2 \times (N+1)^2}$ is the covariance matrix of the input signals.

A time-varying and narrow-band target covariance matrix is then required, which may be defined in this case as

$$\mathbf{C}_\mathrm{target} = \mathrm{E}[(\mathbf{z}_\mathrm{s} + \mathbf{z}_\mathrm{d})(\mathbf{z}_\mathrm{s} + \mathbf{z}_\mathrm{d})^\mathrm{H}]. \tag{10}$$

The optimal mixing solution then provides the values for matrices $\mathbf{A} \in \mathbb{C}^{2 \times 2}$ and $\mathbf{B} \in \mathbb{C}^{2 \times 2}$, in the following equation

$$\mathbf{A} \mathbf{C}_\mathrm{lin} \mathbf{A}^\mathrm{H} + \mathbf{B} \tilde{\mathbf{C}}_\mathrm{lin} \mathbf{B}^\mathrm{H} = \mathbf{C}_\mathrm{target} \tag{11}$$

where $\tilde{\mathbf{C}}_\mathrm{lin} = \mathrm{diag} \left[\mathcal{D}[\mathbf{z}_\mathrm{lin}] \mathcal{D}[\mathbf{z}_\mathrm{lin}]^\mathrm{H} \right]$ is a diagonal matrix consisting of the diagonal entries of the covariance matrix of a decorrelated version of the linearly decoded prototype. More information regarding the derivation and calculation of these mixing matrices is given in [5].

The output audio $\mathbf{z}_\mathrm{par}$ may then be obtained as

$$\mathbf{z}_\mathrm{par} = \mathbf{A} \mathbf{z}_\mathrm{lin} + \mathbf{B} \mathcal{D}[\mathbf{z}_\mathrm{lin}], \tag{12}$$

which ideally exhibits all of the target inter-channel dependencies, as dictated by the target covariance matrix $\mathrm{E}[\mathbf{z}_\mathrm{par} \mathbf{z}_\mathrm{par}^\mathrm{H}] \simeq \mathbf{C}_\mathrm{target}$. However, note that in practice, due to the need for regularisation of the input covariance matrices, the solution is never exact.

4. EVALUATION

A multiple-stimulus test was conducted, in order to compare the perceived performance of the proposed approach (*CroPaC1*), with both first-order (*Ambi1*) and third-order (*Ambi3*) spatial re-sampling ambisonic decoders [12]. Note that the optimal mixing approach (12) was employed, using the same ambisonic decoder for the prototype $\mathbf{z}_{\text{lin}}$.

Synthetic test scenes were created as follows: multiple speakers (*speakers*), a modern funk band (*groove*), and a mix comprising of a speaker, clapping, water fountain and piano (*mix*). The individual sources were binauralised directly for the reference test cases, and encoded into first- and third- order signals and passed through their respective decoders for the *_dry* test cases. A shoebox image-source room simulator [2] was employed to obtain reverberant (*_rev*) counterparts. The room simulator was configured to resemble a small hall. The image-sources arriving at the receiver position were binauralised directly for the reference cases, and encoded into spherical harmonic signals for the reverberant test cases.

The evaluation consisted of three parts, which addressed the spatial, timbral, and overall differences between the methods. For the evaluation of **spatial** differences, the mean spectra of the reference was imposed onto the test cases. Therefore, timbral differences between the methods were greatly mitigated, but still retained their original spatial characteristics. The test subjects were explicitly instructed to ignore any remaining timbral differences, which may not have been addressed by the equalisation. The anchor was obtained as an omni-directional spherical harmonic component, replicated to each binaural channel and also equalised. For evaluating the **timbral** differences, the mean spectra of each test case, was imposed onto an omni-directional signal and sent to both left and right channels. Therefore, the spatial differences between the methods were eliminated, thus retaining only the timbral differences. The anchor was obtained as the mean spectra of the reference, replicated to each binaural channel and low-passed filtered at 4 kHz. Finally, for assessing the **overall** differences, only the broad-band RMS of the reference was used to normalise the test cases. Therefore, all timbral and spatial differences between test cases remained, and the test participants graded the samples based on their subjective weighting of the reproduced attributes.

A total of 14 expert listeners participated in the listening tests in purpose-built headphone booths. The present authors did not take part in the tests. The means and 95% confidence intervals of the results are shown in Fig. 3. It can be observed that the spatial and timbral characteristics of the reference were more closely reproduced using the proposed method, when compared to conventional ambisonics with the same first-order input. Furthermore, the proposed method yielded scores more inline with that of third-order ambisonics. It should be highlighted that third-order ambisonics employs four times the number of input channels than that of the proposed method. This, therefore, represents a significant reduction in bandwidth, without compromising the perceived spatial accuracy or fidelity.

[2] The shoe-box room simulator employed for the reverberant test cases may be found here: https://github.com/polarch/shoebox-roomsim

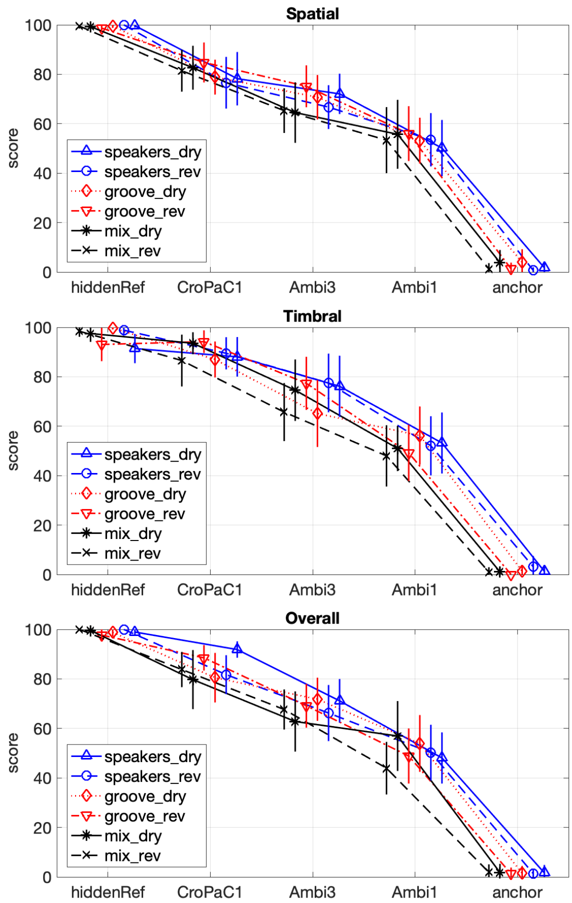

Figure 3: The means and 95% confidence intervals for each individual sound scene. The evaluation criteria were: spatial only, timbral only, and overall (top-bottom).

5. CONCLUSION

This paper has proposed a first-order parametric binaural ambisonic decoder, which employs a sound-field model comprising one source and one directional ambient component per time-frequency tile. The proposed approach first isolates the source components using a spherical harmonic domain beamformer, modulated by the Cross-Pattern Coherence (CroPaC) spatial post-filter. The ambient stream is then simply the residual, once the source components have been subtracted from the input sound-field. The proposed approach is inspired by the COMPASS method [25]. However, along with the CroPaC post-filter, it also employs instantaneous source direction estimation and synthesises the output in a linear manner as much as possible; in order to improve the fidelity of the output signals. Formal listening tests indicate that the proposed first-order decoder performs similarly to (or exceeds) third-order spatial re-sampling ambisonics decoding, in terms of the perceived spatial and timbral attributes of the reproduction.

6. ACKNOWLEDGEMENTS

This research has received funding from the Aalto ELEC Doctoral School. Many thanks are also extended to Dr. Archontis Politis for the insightful discussions regarding the method.

7. REFERENCES

[1] B. Rafaely, *Fundamentals of spherical array processing*, vol. 8. Springer, 2015.

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

[3] V. Pulkki, S. Delikaris-Manias, and A. Politis, "Parametric time-frequency domain spatial audio," 2018.

[4] S. Delikaris-Manias and V. Pulkki, "Cross pattern coherence algorithm for spatial filtering applications utilizing microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2356–2367, 2013.

[5] J. Vilkamo and T. Backstrom, "Time–frequency processing: Methods and tools," in *Parametric Time-Frequency Domain Spatial Audio*, pp. 3–23, John Wiley & Sons, 2018.

[6] O. Santala, H. Vertanen, J. Pekonen, J. Oksanen, and V. Pulkki, "Effect of listening room on audio quality in ambisonics reproduction," in *Audio Engineering Society Convention 126*, Audio Engineering Society, 2009.

[7] S. Braun and M. Frank, "Localization of 3d ambisonic recordings and ambisonic virtual sources," in *1st International Conference on Spatial Audio,(Detmold)*, 2011.

[8] 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," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2711–2721, 2013.

[9] S. Bertet, J. Daniel, E. Parizet, and O. 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.

[10] P. Stitt, S. Bertet, and M. van Walstijn, "Off-centre localisation performance of ambisonics and hoa for large and small loudspeaker array radii," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 937–944, 2014.

[11] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4087–4096, 2017.

[12] B. Bernschütz, A. V. Giner, C. Pörschmann, and J. Arend, "Binaural reproduction of plane waves with reduced modal order," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 972–983, 2014.

[13] M. Zaunschirm, C. Schörkhuber, and R. 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.

[14] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[15] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural rendering of ambisonic signals via magnitude least squares," in *Proceedings of the DAGA*, vol. 44, pp. 339–342, 2018.

[16] H. Lee, M. Frank, and F. Zotter, "Spatial and timbral fidelities of binaural ambisonics decoders for main microphone array recordings," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, Audio Engineering Society, 2019.

[17] V. Pulkki, "Directional audio coding in spatial sound reproduction and stereo upmixing," in *Audio Engineering Society Conference: 28th International Conference: The Future of Audio Technology–Surround and Beyond*, Audio Engineering Society, 2006.

[18] A. Politis, J. Vilkamo, and V. 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.

[19] A. Politis, L. McCormack, and V. Pulkki, "Enhancement of ambisonic binaural reproduction using directional audio coding with optimal adaptive mixing," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,*, 2017.

[20] S. Berge and N. Barrett, "High angular resolution planewave expansion," in *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics May*, pp. 6–7, 2010.

[21] A. Wabnitz, N. Epain, and C. T. Jin, "A frequency-domain algorithm to upscale ambisonic sound scenes," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 385–388, IEEE, 2012.

[22] C. Schörkhuber and R. Höldrich, "Linearly and quadratically constrained least-squares decoder for signal-dependent binaural rendering of ambisonic signals," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, Audio Engineering Society, 2019.

[23] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the audio engineering society*, vol. 45, no. 6, pp. 456–466, 1997.

[24] O. Thiergart, G. Milano, and E. A. Habets, "Combining linear spatial filtering and non-linear parametric processing for high-quality spatial sound capturing," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 571–575, IEEE, 2019.

[25] A. Politis, S. Tervo, and V. Pulkki, "Compass: Coding and multidirectional parameterization of ambisonic sound scenes," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6802–6806, IEEE, 2018.

[26] J. Vilkamo and S. Delikaris-Manias, "Perceptual reproduction of spatial sound using loudspeaker-signal-domain parametrization," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 10, pp. 1660–1669, 2015.

[27] S. Delikaris-Manias, J. Vilkamo, and V. Pulkki, "Parametric binaural rendering utilizing compact microphone arrays," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 629–633, IEEE, 2015.

[28] J. Vilkamo, T. Bäckström, and A. Kuntz, "Optimized covariance domain framework for time–frequency processing of spatial audio," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 403–411, 2013.

[29] D. P. Jarrett, E. A. Habets, and P. A. Naylor, "3d source localization in the spherical harmonic domain using a pseudointensity vector," in *2010 18th European Signal Processing Conference*, pp. 442–446, IEEE, 2010.

[30] J. Ivanic and K. 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.