Enhancement Of Ambisonic Binaural Reproduction Using Directional Audio Coding With Optimal Adaptive Mixing

Conference Paper · October 2017

DOI: 10.1109/WASPAA.2017.8170059

Archontis Politis
Tampere University

Leo McCormack
Aalto University

Ville Pulkki
Aalto University

ENHANCEMENT OF AMBISONIC BINAURAL REPRODUCTION USING DIRECTIONAL AUDIO CODING WITH OPTIMAL ADAPTIVE MIXING

Archontis Politis, *Leo McCormack,* *Ville Pulkki*

archontis.politis, leo.mccormack, ville.pulkki@aalto.fi
Department of Signal Processing and Acoustics
School of Electrical Engineering
Aalto University, FI-02150 Espoo, Finland

ABSTRACT

Headphone reproduction of recorded spatial sound scenes is becoming increasingly relevant to immersive audiovisual applications. Popular non-parametric reproduction methods, such as first-order ambisonics (FOA), can now be surpassed through the use of perceptually-motivated parametric reproduction methods. One such established method is Directional Audio Coding (DirAC). The earlier version of DirAC for headphones was itself limited to FOA input and achieved binaural rendering through a virtual loudspeaker approach; resulting in a high computational overhead. Therefore, this paper proposes an improved DirAC method that directly synthesises the binaural cues based on the estimated spatial parameters. This method can accommodate higher-order Ambisonics (HOA) signals and has reduced computational requirements; thus, making it suitable for lightweight processing with fast update rates and head-tracking support. According to listening tests, when utilising only FOA signals, the method results in equivalent or higher spatial accuracy than third-order Ambisonics and significantly outperforms FOA reproduction.

Index Terms— Parametric spatial audio, Ambisonics, binaural, headphone reproduction

1. INTRODUCTION

The ability to reproduce recorded spatial sound scenes over head-tracked headphones, whilst preserving the full spatial information of the scene, is an increasingly crucial component of the emerging immersive audiovisual technologies; including: personal cinema systems, telepresence, virtual reality and augmented reality. Existing methods strike a balance between spatial accuracy and fidelity of the binaural signals and can be roughly categorised as either parametric or non-parametric methods. This work focuses on parametric methods that operate on B-format [or spherical harmonic (SH)] signals, as they encode spatial information of the scene for all directions equally and they may be obtained from both spatial mixing and real recordings; thus, making them suitable for the aforementioned applications.

With regard to headphone reproduction, B-format signals are commonly decoded using Ambisonics [1], a non-parametric method which aims to reconstruct the appropriate binaural cues by applying an ambisonic decoding matrix that integrates a set of head-related-transfer functions (HRTFs) [2, 3, 4, 5, 6, 7]. This ambisonic decoding results in a static matrix of filters that transform the B-format signals to binaural signals in a linear manner. Ambisonic decoding does not introduce any distortion to the resulting signals; however, like signal-independent beamforming, its performance is completely determined by the spatial resolution of the input format. In the case of FOA and lower-order Ambisonics, this limitation can affect the perceived spatial quality of the result; producing effects such as: directional blurring of point sources, localisation ambiguity, loss of externalisation, reduced sense of envelopment in reverberant conditions and strong colouration effects [8, 9, 10, 11, 12, 7, 13, 14].

Directional Audio Coding [15] is a parametric method for spatial sound reproduction and upmixing, which operates in a time-frequency transform domain and extracts a direction-of-arrival (DoA) parameter and a diffuseness parameter at each time-frequency point. The parameters are based on the energetic quantities of the active intensity and diffuseness, which DirAC interprets as narrowband perceived DoA and inter-aural coherence cues. While originally defined for FOA signals, DirAC has recently been extended to accommodate higher-order signals (HO-DirAC) [16, 17]. The parametric model of DirAC has been found effective in a number of listening tests [18, 19, 20, 17], improving perceptual deficiencies of lower-order Ambisonics, at an increased computational cost and implementation complexity. Laitinen et al. [19] applied DirAC to headphone reproduction, based on a virtual loudspeaker approach (VL-DirAC). While this method has been effectively demonstrated in real-time with head-tracking support utilising a variety of content, it is computationally demanding and prone to certain artefacts; especially in challenging scenarios for the parametric analysis. Apart from DirAC, the HARPEX method [21] is another parametric approach based on FOA and binaural rendering, with a dual plane-wave model and no diffuse component.

Building on the HO-DirAC formulation in [17], this work proposes a new architecture for headphone reproduction that overcomes the drawbacks of VL-DirAC. Additionally, the proposed formulation incorporates support for HOA signals, if available, and the optimal adaptive mixing solution established in [22], in order to synthesise the binaural cues directly from the spatial parameters.

2. AMBISONIC BINAURAL DECODING

It is assumed that the spatial sound scene is recorded or mixed to the $(N+1)^2$-length vector of B-format signals $\mathbf{x}_N$ of order N[1]. It is further assumed that the signals have been transformed into the time-frequency domain [via a short-time Fourier transform (STFT)

[1] The spherical harmonic conventions commonly used in HOA are followed here, with orthonormalised real SHs of mode-number (n, m) and channel indexing $q = 1, ..., (N+1)^2$ with $q = n^2 + n + m + 1$.

or perfect reconstruction filterbank], with time and frequency indices denoted as (t, f). Ambisonic decoding to headphones can be performed with a $2 \times (N+1)^2$ matrix of filters $\mathbf{D}_{\text{bin}}$ based on a set of individualised or generic HRTFs

$$\mathbf{y}_{\text{bin}}(t,f) = \mathbf{D}_{\text{bin}}(f)\mathbf{x}_N(t,f). \tag{1}$$

The ambisonic decoding matrix depends on the available order N and can be designed based on a virtual loudspeaker approach [2, 5], or by directly expressing the HRTFs in the SH domain [6]; for a description of the two approaches see e.g. [23]. Rotation of the sound scene, necessary for head-tracking, can be easily performed with Ambisonics through appropriate $(N+1)^2 \times (N+1)^2$ SH rotation matrices $\mathbf{M}_{\text{rot}}$

$$\mathbf{y}_{\text{bin}}^{(\text{rot})}(t,f) = \mathbf{D}_{\text{bin}}(f)\mathbf{M}_{\text{rot}}(\alpha,\beta,\gamma)\mathbf{x}_N(t,f), \tag{2}$$

where α, β, γ are rotation angles sent by the head-tracker. For implementation details on the rotation matrices see e.g. [23].

3. LEGACY DIRECTIONAL AUDIO CODING

Legacy DirAC utilises only the FOA signals $\mathbf{x}_1$, which relate directly to the acoustic pressure and particle velocity at the recording point. It operates by estimating the quantities of the active intensity vector $\mathbf{i}_a$, which expresses the net energy flow, energy density E, and the diffuseness ψ; where the latter represents the diffuse-to-total energy ratio. The opposite direction of $\mathbf{i}_a$ provides the DoA azimuth-elevation angles θ, ϕ. The spatial parameter vector of DirAC comprises of these parameters: $\mathbf{p}_1 = [\theta, \phi, E, \psi]$. During synthesis, preliminary ambisonic decoding is first performed with a matrix $\mathbf{D}_{\text{ls}}$. Then the loudspeaker signals are further enhanced parametrically, where the parameters are utilised to separate the signals into directional and diffuse parts and to re-distribute them appropriately between the loudspeakers. This is achieved by means of vector-base amplitude panning (VBAP) gains [24] for the directional stream, and decorrelation filters for the diffuse stream. The VL-DirAC is similar to the legacy loudspeaker version of DirAC, employing a virtual set of loudspeaker signals, which are further convolved with their respective HRTFs.

The method can be expressed in a simplified form as (we omit the (t, f) indices below for compactness)

$$\mathbf{p}_1 = \mathcal{A}\left[\mathbf{x}_1\right], \tag{3}$$

$$\mathbf{z}_{\text{ls}} = \frac{\sqrt{1-\psi}}{L}\mathbf{G}(\theta,\phi)\mathbf{D}_{\text{ls}}\mathbf{x}_1 + \mathcal{D}\left[\sqrt{\psi}\mathbf{D}_{\text{ls}}\mathbf{x}_1\right], \tag{4}$$

$$\mathbf{y}_{\text{bin}} = \mathbf{H}_{\text{ls}}\mathbf{z}_{\text{ls}}, \tag{5}$$

where L is the number of virtual loudspeakers, $\mathbf{G}(\theta, \phi)$ is an $L \times L$ diagonal matrix of VBAP gains for the analysed DoA, $\mathbf{D}_{\text{ls}}$ is an $L \times 4$ ambisonic decoding matrix and $\mathcal{D}[\cdot]$ denotes a decorrelation operation of the enclosed signals. The first equation (3) denotes the energetic analysis of DirAC from the FOA signals (see e.g. [17] for details); while (4) describes the synthesis, and (5) the binaural rendering of the virtual loudspeaker signals with the set of HRTFs contained in a $2 \times L$ filter matrix $\mathbf{H}_{\text{ls}}$. Head-tracking support is integrated through the appropriate rotation of the analysed DoAs and the ambisonic signals prior to decoding [19].

4. PROPOSED DIRAC WITH OPTIMAL MIXING

One issue with VL-DirAC, is that it does not take into account what is already achieved by the ambisonic decoding, in terms of directional spatialisation and diffuse reproduction. Instead, time-variant panning and diffuse gains are applied to the linearly-decoded signals $\mathbf{D}_{\text{ls}}\mathbf{x}_1$ without considering jointly the combined linear and parametric rendering, apart from a mean correction for energy preservation [18]. Higher perceptual quality and robustness should be attained if the adaptive gains and decorrelation are applied only to the extent that is necessary after the linear decoding. Such a reformulation of DirAC synthesis is presented by Vilkamo et al. in [22], termed here as *optimal mixing*, based on the analysed parameters. Another issue of VL-DirAC is its inability to make use of HOA signals at both analysis and synthesis, to obtain a more accurate model of the sound scene and improve performance in scenarios that are challenging for the basic intensity-diffuseness model. The HO-DirAC of [17] takes advantage of the higher spatial resolution of HOA signals in order to extract multiple intensity-diffuseness estimates at spatially separated sectors. HO-DirAC fully integrates the optimal mixing formulation of [22] and preserves the single-channel quality of the linear decoding, while enhancing spatialisation cues that Ambisonics may fail to deliver; such as sharp point source sounds and spatially incoherent diffuse sounds. In [17], HO-DirAC was formulated for arbitrary loudspeaker set-ups, while in this paper it is reformulated for efficient headphone reproduction.

If B-format signals $\mathbf{x}_N$, where N is the order, are available, then the Nth-order HO-DirAC analysis will first determine spatially selective signals defining spatial sectors that cover the sphere uniformly. The standard DirAC parameters in each sector may then be extracted separately

$$\begin{aligned}\mathbf{p}_N = \mathcal{A}\left[\mathbf{W}_N\mathbf{x}_N\right] = [&E_1, \psi_1, \theta_1, \phi_1, ...\\ &E_S, \psi_S, \theta_S, \phi_S],\end{aligned} \tag{6}$$

where $\mathbf{W}_N$ is a $4S \times (N+1)^2$ beamforming matrix that generates the appropriate sector analysis signals, and S is the order-dependent number of sectors; for more details on the structure of $\mathbf{W}_N$ please refer to [17]. Note that if only FOA signals are available, the analysis reduces to the basic DirAC analysis of (3). During synthesis, instead of virtual loudspeaker decoding, the signals are directly decoded linearly to binaural signals using an appropriate ambisonic decoding matrix $\mathbf{y}_{\text{lin}} = \mathbf{D}_{\text{bin}}\mathbf{x}_N$, similar to (1). The binaural signals $\mathbf{y}_{\text{lin}}$ serve as the input to the optimal mixing block. Therefore, instead of the DirAC enhancement being applied to $L >> 2$ virtual loudspeakers, it operates only on two signals.

The optimal mixing solves the problem of finding the mixing matrix that needs to be applied to the inputs in such a manner as to ensure that the outputs have the desired binaural cues; while preserving the high single-channel quality of the linearly-decoded signals as much as possible. As the binaural cues relate to inter-channel level and phase differences and inter-channel coherence, they are all captured in the complex narrowband covariance matrix of the binaural signals. The problem is set as

$$\mathbf{y}_{\text{bin}} = \mathbf{A}\mathbf{y}_{\text{lin}} + \mathbf{B}\mathcal{D}\left[\mathbf{y}_{\text{lin}}\right], \tag{7}$$

with the mixing matrices $\mathbf{A}, \mathbf{B}$ being the solution to

$$\mathbf{A}\mathbf{C}_{\text{lin}}\mathbf{A}^H + \mathbf{B}\tilde{\mathbf{C}}_{\text{lin}}\mathbf{B}^H = \mathbf{C}_{\text{model}}, \tag{8}$$

where $\mathbf{C}_{\text{lin}} = \mathcal{E}\left[\mathbf{y}_{\text{lin}}\mathbf{y}_{\text{lin}}^H\right]$ is the covariance matrix of the input signals, and $\tilde{\mathbf{C}}_{\text{lin}}$ is a diagonal covariance matrix of their decorrelated

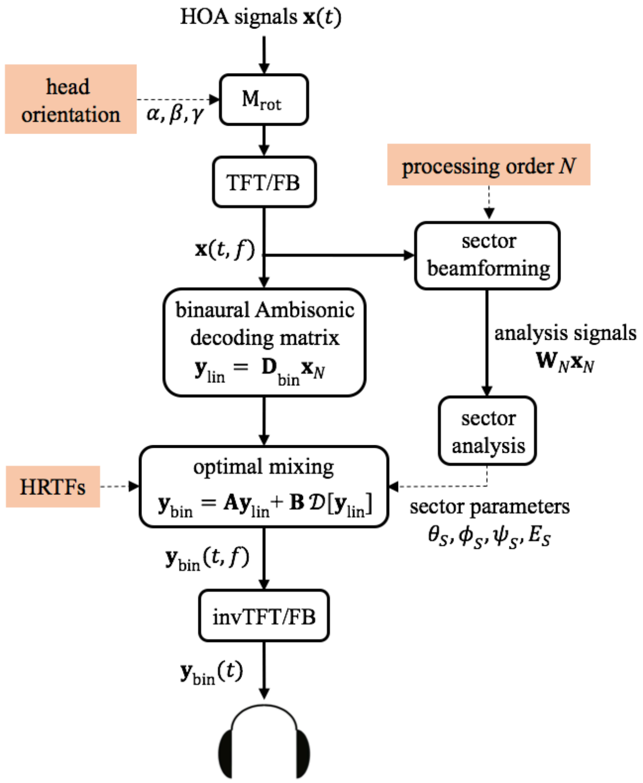

HOA signals $\mathbf{x}(t)$
head orientation
α, β, γ
$\mathrm{M_{rot}}$
TFT/FB
processing order N
$\mathbf{x}(t, f)$
sector beamforming
binaural Ambisonic decoding matrix
$\mathbf{y}_{\mathrm{lin}} = \mathbf{D}_{\mathrm{bin}}\mathbf{x}_N$
analysis signals
$\mathbf{W}_N\mathbf{x}_N$
sector analysis
HRTFs
optimal mixing
$\mathbf{y}_{\mathrm{bin}} = \mathbf{A}\mathbf{y}_{\mathrm{lin}} + \mathbf{B}\,\mathcal{D}[\mathbf{y}_{\mathrm{lin}}]$
sector parameters
$\theta_S, \phi_S, \psi_S, E_S$
$\mathbf{y}_{\mathrm{bin}}(t, f)$
invTFT/FB
$\mathbf{y}_{\mathrm{bin}}(t)$

Figure 1: Block diagram of the proposed DirAC binaural rendering with optimal mixing. The TFT/FB blocks denote a time-frequency transform or perfect reconstruction filterbank.

version $\mathcal{D}\left[\mathbf{y}_{\mathrm{lin}}\right]$. The desired target dependencies, contained in the model covariance matrix $\mathbf{C}_{\mathrm{model}}$ are prescribed according to the analysed spatial parameters. Note that the solution first tries to meet the energies and coherences imposed by $\mathbf{C}_{\mathrm{model}}$, via linear mixing of the input signals through $\mathbf{A}$, without employing decorrelation; therefore, avoiding potential decorrelation artefacts such as temporal smearing of transients. However, if the target dependencies cannot be satisfied in this first step, decorrelation is employed on the input signals and these are further mixed to the outputs through $\mathbf{B}$. For example, such a case could arise when the sound-field is analysed as being completely diffuse and $\mathbf{y}_{\mathrm{lin}}$ fails to meet the appropriate binaural correlations, due to high inter-channel coherence of the Ambisonic decoding. The solution to (8) is detailed in [25]. Apart from satisfying the constraint in (8), the solution aims to minimise phase differences between the enhanced $\mathbf{y}_{\mathrm{bin}}$ of (7) and $\mathbf{y}_{\mathrm{lin}}$ as much as possible; hence preserving the high single-channel quality of the linear decoding and increasing robustness. The overall structure of the proposed method is presented in Fig. 1.

The final step to complete the method is the definition of the target covariance matrix $\mathbf{C}_{\mathrm{model}}$. Assuming S sectors are used in the analysis, there are S sets of parameters $\mathbf{p}_s = [E_s, \psi_s, \theta_s, \phi_s]$. The following assumptions are made:

a) the energy of the diffuse part for the sth sector is $\psi_s E_s$,

b) the energy of the directional part for the sth sector is $(1-\psi_s)E_s$,

c) the diffuse components are uncorrelated between sectors, and

d) the directional components are uncorrelated with the diffuse components.

Based on these assumptions, the covariance matrix of the directional and diffuse components of a single sector is

$$\mathbf{C}_{\mathrm{dir}}^{(s)} = (1-\psi_s)E_s\mathbf{h}(\theta_s, \phi_s)\mathbf{h}^{\mathrm{H}}(\theta_s, \phi_s), \tag{9}$$

$$\mathbf{C}_{\mathrm{diff}}^{(s)} = \psi_s E_s \mathbf{U}, \tag{10}$$

where $\mathbf{h} = [h_{\mathrm{L}}, h_{\mathrm{R}}]^{\mathrm{T}}$ is the vector of HRTFs for direction θ, ϕ. The matrix $\mathbf{U}$ is a diffuse energy distribution matrix dependent on the reproduction system. In the case of binaural signals, it should conform to the binaural coherence curve $c_{\mathrm{bin}}(f)$ under diffuse-field excitation. Such a coherence curve can be computed from the set of HRTFs, e.g. as proposed in [26], or it can be approximated by a parametric model such as in [27]. A suitable distribution matrix is

$$\mathbf{U} = \begin{bmatrix} \alpha & c_{\mathrm{bin}} \\ c_{\mathrm{bin}} & \beta \end{bmatrix}, \tag{11}$$

where α, β are factors that distribute the diffuse energy between the left and right ears with $\alpha + \beta = 1$, and determined by $[\alpha, \beta] = \mathrm{diag}\left[\tilde{\mathbf{C}}_{\mathrm{lin}}\right] / \mathrm{trace}\left[\tilde{\mathbf{C}}_{\mathrm{lin}}\right]$. The total target $\mathbf{C}_{\mathrm{model}}$ combines all sector contributions

$$\mathbf{C}_{\mathrm{model}} = \sum_{s=1}^{S} \mathbf{C}_{\mathrm{dir}}^{(s)} + \mathbf{C}_{\mathrm{diff}}^{(s)}. \tag{12}$$

Matrix $\mathbf{C}_{\mathrm{model}}$ essentially contains the binaural inter-aural level differences, phase differences and inter-aural coherence determined by the directional analysis; its definition, along with (8), concludes the proposed solution.

5. IMPLEMENTATION

A real-time implementation of the system was developed for evaluation as an audio plug-in. The implementation supports B-format signals up to 7th-order and head-tracking. Contrary to the rotation of the analysed DoAs, which was performed in VL-DirAC [19], it was found to be more efficient to rotate the B-format signals directly, due to the faster update rates, as shown in Fig. 1. The rotation angles are updated for every analysis frame.

The time-frequency transform selected for the implementation is based on the STFT, with analysis and synthesis windows optimised to suppress temporal aliasing; the source code of the implementation is provided in [28]. The temporal resolution of the transform is determined by a hop size of 2.7 msec (128 samples at 48kHz sample rate). The transform has a uniform resolution of 128 bands, but employs additional filters to increase low-frequency resolution similar to hybrid filterbanks for spatial audio coding [29], resulting in a total of 133 bands. The spatial parameters $\mathbf{p}_N$ of (6) are obtained instantaneously to capture rapid variations of the sound scene, while the definition of input and target covariance matrices $\mathbf{C}_{\mathrm{lin}}, \mathbf{C}_{\mathrm{model}}$ are computed across multiple windows to capture and provide meaningful signal statistics for a robust synthesis.

The beamforming coefficients $\mathbf{W}_N$ in (6), the Ambisonic decoding matrix $\mathbf{D}_{\text{bin}}$ and the binaural coherence matrix $\mathbf{U}$ of (11) are precomputed and stored. The decoding matrix $\mathbf{D}_{\text{bin}}$ is computed using a densely measured set of HRTFs from one of the authors. The decoding filters and HRTFs are pre-processed with the same hybrid STFT as applied during runtime and converted to spectral coefficients. During construction of the target covariance matrix, the HRTF vector $\mathbf{h}$ in (9) is interpolated from the measured set to the respective analysed direction (θ_s, ϕ_s) using triangular interpolation on the measurement grid [30].

6. EVALUATION

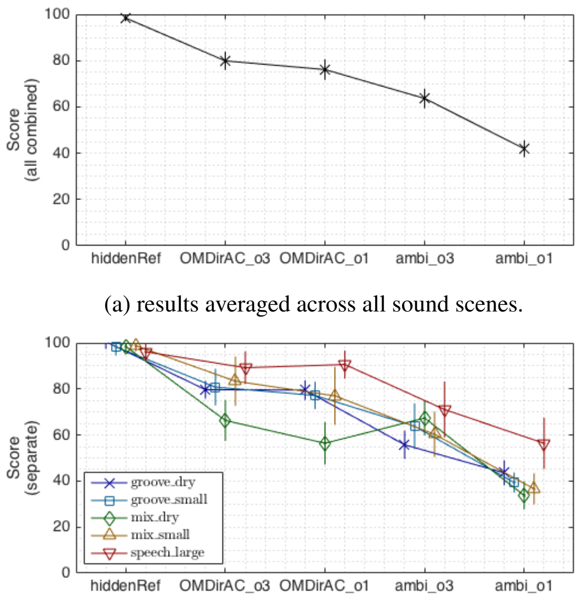
A multiple-stimulus listening test was conducted in order to assess the performance of OM-DirAC. All the stimuli used in the listening tests are available online[2].

Five synthetic sound scenes were simulated with a varying number of sound sources in anechoic and reverberant environments. Room reverberation was simulated with the image source method. All direct paths and image sources were quantised directly to 28 plane wave signals covering the sphere, without employing panning. These 28-channel signals were played back through real loudspeakers in an anechoic chamber from their corresponding directions, to assess the naturalness of the synthetic scenes. The 28-channel scenes served as a reference to assess the different methods and were specifically designed to be critical of basic parametric analysis. There are two free-field sound scenes, labelled here as *groove_dry* and *mix_dry*. The former consists of individual dry recordings of a band distributed on the front hemisphere horizontally, while the latter incorporates clapping, a fountain, piano and female speech, with three of the sound sources placed horizontally and one above the listener. *groove_small* and *mix_small*, are the reverberant versions of their free-field counterparts. The final sample *speech_large*, is comprised of female speech in front of the listener simulated in a large hall.

The reference test cases *hiddenRef*, were obtained by convolving the 28-channel signals with their respective HRTFs and summing the resulting binaural signals. Ambisonic encoders were applied to each of the 28-channel sound scenes, in order to obtain both FOA and third-order ambisonic signals. The sum of the omnidirectional signals served as a low-quality anchor. The test cases for both first and third-order Ambisonics and OM-DirAC *ambi_o1*, *ambi_o3*, *OMDirAC_o1*, *OMDirAC_o3*, were obtained by passing the corresponding ambisonic signals through their respective offline decoders. Note that the standalone Ambisonic decoders are identical to those used within the OM-DirAC implementation.

The test subjects were instructed to rate the test case they perceived to be closest to the reference, in terms of *overall quality* and *spatial accuracy*, as 100; to rate the test case furthest from the reference as 0; and to rate the remaining four test cases relative to each-other, the reference and anchor. Instructions on overall quality were to include any perceived spatial or temporal artefacts. Since previous studies [18, 19, 20, 17], have found that lower-order Ambisonics may colour the output spectrum compared to the reference, all of the test cases were equalised to spectrally match their reference. This served to reduce the likelihood of large variances in the results due to the easily remedied spectral differences between methods.

There were 13 expert listeners that volunteered to participate. It can be observed that for the majority of sound scenes (Fig. 2b) the first and third order variants of OM-DirAC are perceived as being closer to the reference (in terms of overall quality and spatial accuracy) when compared to their respective first and third order variants of Ambisonics. However, it is evident in the *mix_dry* sound scene that spatial artefacts induced by the lack of individual sectors in *OMDirAC_o1* have negatively impacted scores; although, the scores are still significantly higher than the *ambi_o1* test case, while using the same FOA signals. It can also be seen that the increased number of sectors in the third order case *OMDirAC_o3*, has reduced spatial artefacts to a certain extent; however, the performance is not significantly different to the *ambi_o3* test case for the *mix_dry* sound scene. Regardless, it must be stressed that this particular sound scene does not represent a likely recording scenario and most recordings will contain some degree of reverberation, which can be seen to mask these spatial artefacts to some degree in the reverberant counterpart *mix_small*.



(a) results averaged across all sound scenes.

(b) results for each sound scene.

Figure 2: The means and 95 % confidence intervals of the listening test results.

7. CONCLUSIONS

This work presents a new DirAC approach for head-tracked headphone playback, which represents a clear improvement over existing implementations by reducing the computational requirements and artefacts arising from model mismatch; while also improving the robustness and overall perceived spatial accuracy. According to the listening test results, based on comparisons with a binaural reference, the method outperforms FOA decoding for all tested sound scenes and HOA (third-order) decoding in a number of cases using only FOA signals. When using HOA (third-order) signals, the method is improved further and performs better than ambisonic decoding for all cases bar one; attaining scores which more closely match the reference.

[2] The samples used for the listening test can be found on the companion website: http://research.spa.aalto.fi/publications/papers/waspaa17-omdirac/

8. REFERENCES

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

[2] B. Wiggins, I. Paterson-Stephens, and P. Schillebeeckx, “The analysis of multi-channel sound reproduction algorithms using HRTF data.” in *19th Int. Conf. of AES*, Schloss Elmau, Germany, 2001.

[3] M. Noisternig, T. Musil, A. Sontacchi, and R. Holdrich, “3D binaural sound reproduction using a virtual ambisonic approach,” in *IEEE Int. Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems (VECIMS’03)*, Lugano, Switzerland, 2003, pp. 174–178.

[4] L. S. Davis, R. Duraiswami, E. Grassi, N. A. Gumerov, Z. Li, and D. N. Zotkin, “High order spatial audio capture and its binaural head-tracked playback over headphones with HRTF cues,” in *119th Convention of AES*, New York, NY, USA, 2005.

[5] F. Melchior, O. Thiergart, G. Del Galdo, D. de Vries, and S. Brix, “Dual radius spherical cardioid microphone arrays for binaural auralization,” in *127th Convention of AES*, New York, NY, USA, 2009.

[6] N. R. Shabtai and B. Rafaely, “Binaural sound reproduction beamforming using spherical microphone arrays,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 101–105.

[7] 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.

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

[9] O. Santala, H. Vertanen, J. Pekonen, J. Oksanen, and V. Pulkki, “Effect of listening room on audio quality in Ambisonics reproduction,” in *126th Convention of the AES*, Munich, Germany, 2009.

[10] S. Braun and M. Frank, “Localization of 3D ambisonic recordings and ambisonic virtual sources,” in *1st Int. Conf. on Spatial Audio (ICSA)*, Detmold, Germany, 2011.

[11] G. Kearney, M. Gorzel, H. Rice, and F. Boland, “Distance perception in interactive virtual acoustic environments using first and higher order ambisonic sound fields,” *Acta Acustica united with Acustica*, vol. 98, no. 1, pp. 61–71, 2012.

[12] 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.

[13] 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.

[14] L. Yang and B. Xie, “Subjective evaluation on the timbre of horizontal ambisonics reproduction,” in *Int. Conf. on Audio, Language and Image Processing (ICALIP)*, Shanghai, China, 2014.

[15] V. Pulkki, “Spatial sound reproduction with Directional Audio Coding,” *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.

[16] V. Pulkki, A. Politis, G. Del Galdo, and A. Kuntz, “Parametric spatial audio reproduction with higher-order B-format microphone input,” in *134th Convention of AES*, Rome, Italy, 2013.

[17] 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.

[18] J. Vilkamo, T. Lokki, and V. Pulkki, “Directional Audio Coding: Virtual microphone-based synthesis and subjective evaluation,” *Journal of the Audio Engineering Society*, vol. 57, no. 9, pp. 709–724, 2009.

[19] M.-V. Laitinen and V. Pulkki, “Binaural reproduction for Directional Audio Coding,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2009.

[20] A. Politis, M.-V. Laitinen, J. Ahonen, and V. Pulkki, “Parametric spatial audio processing of spaced microphone array recordings for multichannel reproduction,” *Journal of the Audio Engineering Society*, vol. 63, no. 4, pp. 216–227, 2015.

[21] S. Berge and N. Barrett, “A new method for B-format to binaural transcoding,” in *40th Int. Conf. of AES*, Tokyo, Japan, 2010.

[22] J. Vilkamo and V. Pulkki, “Minimization of decorrelator artifacts in Directional Audio Coding by covariance domain rendering,” *Journal of the Audio Engineering Society*, vol. 61, no. 9, pp. 637–646, 2013.

[23] A. Politis and D. Poirier-Quinot, “JSAmbisonics: A Web Audio library for interactive spatial sound processing on the web,” in *Interactive Audio Systems Symposium*, York, UK, 2016.

[24] 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.

[25] 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.

[26] A. Politis, “Diffuse-field coherence of sensors with arbitrary directional responses,” *arXiv preprint arXiv:1608.07713*, 2016.

[27] C. Borß and R. Martin, “An improved parametric model for perception-based design of virtual acoustics,” in *35th Int. Conf. of AES*, London, UK, 2009.

[28] J. Vilkamo, “Alias-free short-time Fourier Transform,” https://github.com/jvilkamo/afSTFT.

[29] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegard, “Low complexity parametric stereo coding,” in *116th Convention of AES*, Berlin, Germany, 2004.

[30] H. Gamper, “Head-related transfer function interpolation in azimuth, elevation, and distance,” *The Journal of the Acoustical Society of America*, vol. 134, no. 6, pp. EL547–EL553, 2013.