An evaluation of pre-processing techniques for virtual loudspeaker binaural ambisonic rendering

Thomas Mckenzie, Damian Murphy, Gavin Kearney

To cite this version:

Thomas Mckenzie, Damian Murphy, Gavin Kearney. An evaluation of pre-processing techniques for virtual loudspeaker binaural ambisonic rendering. EAA Spatial Audio Signal Processing Symposium, Sep 2019, Paris, France. pp.149-154, 10.25836/sasp.2019.09. hal-02275172

HAL Id: hal-02275172
https://hal.science/hal-02275172
Submitted on 30 Aug 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
AN EVALUATION OF PRE-PROCESSING TECHNIQUES FOR VIRTUAL LOUDSPEAKER BINAURAL AMBISONIC RENDERING

Thomas McKenzie, Damian T. Murphy, Gavin Kearney
AudioLab, Communication Technologies Research Group, Department of Electronic Engineering, University of York, York, YO10 5DD, UK
thomas.mckenzie@york.ac.uk

ABSTRACT

Binaural Ambisonic rendering can be achieved using virtual loudspeakers through head-related impulse response (HRIR) convolution of the Ambisonic loudspeaker feeds. It is widely used in immersive applications such as virtual reality due to its sound field rotation capabilities and low channel count. Binaural Ambisonic reproduction is inaccurate at high frequencies, causing reduced localisation and timbral accuracy, but can be improved through offline pre-processing of the virtual loudspeaker HRIRs. This paper details a numerical and perceptual evaluation of several state of the art pre-processing technique combinations.

1. INTRODUCTION

Binaural Ambisonic rendering is well suited to virtual reality applications due to its sound field rotation capabilities. Ambisonic reproduction can theoretically replicate the original sound field exactly in the region of the head for frequencies up to what is commonly referred to as the ‘spatial aliasing frequency’, $f_{\text{alias}}$. Above $f_{\text{alias}}$, reproduction can become inaccurate due to the limited spatial accuracy of reproducing a physical sound field with a finite number of transducers, which in practice causes localisation blur [8], reduced localisation [1] and comb filtering spectral artefacts [2].

The standard approach to improving Ambisonic reproduction is to increase the order of Ambisonics, which allows for exact sound field reproduction up to a higher $f_{\text{alias}}$ [3, 4], though at the expense of more channels for storage, more microphone capsules for recording, and more convolutions in binaural reproduction. It is therefore highly desirable to explore alternative methods of improving low-order Ambisonic rendering. One common practice is to employ a dual-band decoder with basic Ambisonic decoding at low frequencies and maximum $r_E$ channel weighting above $f_{\text{alias}}$ [1, 5], which improves spectral, localisation and lateralisation reproduction at high frequencies.

This paper presents the method and results of numerical and perceptual evaluations of state-of-the-art pre-processing technique combinations for virtual loudspeaker binaural Ambisonic rendering. These pre-processing techniques are applied to the head-related impulse responses (HRIRs) used in the virtual loudspeaker binaural rendering stage in offline processes, such that the resulting binaural decoders are of the same size and require the same number of real-time convolutions. This paper investigates 1st, 2nd and 3rd order Ambisonics, with loudspeaker configurations comprising 6, 14 and 26 loudspeakers respectively, arranged in Lebedev grids [9].

All HRIRs used in this study were generic measurements from a diffuse-field equalised version of the Benschutz Neumann KU 100 database [6]. All computations was carried offline in MATLAB version 9.3.0 R2017b and Ambisonic encoding and decoding was achieved using the Politis Ambisonic library [7], which uses three-dimensional full normalisation (N3D) and Ambisonic channel number (ACN) ordering. All audio used was of 24-bit depth and 48 kHz sample rate. Ambisonics was rendered using mode-matching pseudo-inverse decoding and, unless otherwise stated, dual-band decoding was utilised with basic channel weightings at frequencies below $f_{\text{alias}}$ and maximum $r_E$ weightings above [1, 5]. In this study, $f_{\text{alias}}$ was approximated according to [8] with a speed of sound of 343 m/s and radius of the listening area as 9 cm (the approximate radius of the Neumann KU 100 dummy head) as 670 Hz, 1270 Hz and 1870 Hz, for 1st, 2nd and 3rd order Ambisonics, respectively.

2. PRE-PROCESSING TECHNIQUES

The three pre-processing techniques tested in this paper are time alignment (TA), Ambisonic interaural level difference optimisation (AIO) and diffuse-field equalisation (DFE). TA is the complete removal of interaural time differences (ITDs) of the HRIRs at high frequencies [10, 11], which reduces the comb filtering caused by the off-centre position of the ears in the virtual loudspeaker array. TA has previously only been implemented for dense sets of HRIRs and is here applied to sparse virtual loudspeaker sets. When using TA, basic channel weighting is usually used for the whole frequency spectrum. The crossover frequency above which TA is implemented in this study is 2.5 kHz; chosen
according to the listening test results in [12]. AIO is an iterative pre-processing technique that brings the Ambisonic rendering of interaural level differences (ILDs) [13] closer to that of HRIRs. It is achieved by augmenting the high frequency magnitude of virtual loudspeaker HRIRs such that when rendering Ambisons with the augmented HRIRs, the ILD is reproduced with greater accuracy. DFE is the removal of direction-independent spectral artefacts in the Ambisonic rendered diffuse-field [14], which improves the overall spectral reproduction of binaural Ambisonic rendering, as well as the timbral consistency between different virtual loudspeaker configurations [15]. It works by creating Ambisonic renders at a large amount of directions on the sphere and obtaining an average of all the frequency responses. This is then inverted and the equalisation filter is applied to the virtual loudspeaker HRIRs.

In this paper, different pre-processing techniques are combined. Theoretically, by running one after the other, the resulting binaural Ambisonic decoder will produce greater results than just one of the pre-processing techniques. The order of pre-processing techniques is as follows: TA is implemented first as it affects the rendering of ILD and the diffuse-field response. AIO also affects the diffuse-field response, so follows TA. DFE is implemented last, as it corrects any changes in average frequency response and the other pre-processing techniques can affect the diffuse-field response. The five binaural Ambisonic decoders under test in this paper (along with their abbreviations), with order of implementation, are as follows:

- NPP: Standard Ambisonic (dual band)
- PP 1: AIO & DFE (dual band)
- PP 2: TA & DFE (basic)
- PP 3: TA & AIO & DFE (basic)
- PP 4: TA & AIO & DFE (dual band)

AIO produces the greatest benefits for dual-band decoding, but TA is recommended for basic weighted decoding. Therefore, in PP 3 and PP 4 with the combination of all three pre-processing techniques, both basic weighted and dual-band instances are included to ascertain the differences and determine which (if any) is superior.

3. NUMERICAL EVALUATION

To assess the effect of different pre-processing combinations on the spectral accuracy of binaural Ambisonic rendering, a perceptually motivated spectral difference (PSD) fast Fourier transfer (FFT) based model was used [16]. It weights input signals using ISO 226 equal loudness contours and a sone scale to account for the loudness-varying sensitivity of human hearing as well as equivalent rectangular bandwidth weightings to address the linear frequency sample spacing of FFTs.

PSD between binaural Ambisonic renders and HRIRs for 16,020 locations over the sphere (distributed using a 2° Gaussian grid) with different combinations of pre-processing techniques for 1st, 2nd and 3rd order Ambisons was calculated. Figure 1 shows the solid angle weighted average PSD value for each pre-processing combination for all tested orders of Ambisons, with whiskers to denote the maximum and minimum PSD values. As expected, higher orders of Ambisons produce improved spectral reproduction. In all tested orders of Ambisons, every pre-processing combination improves the overall spectral accuracy over standard dual-band decoding, but PP 4 (the dual-band combination of TA, AIO and DFE) produces the greatest improvements with the lowest solid-angle weighted PSD and the lowest value of maximum PSD for all 3 tested orders of Ambisons. To illustrate how PSD changes over the sphere, the PSD for each combination of 1st order pre-processing techniques for all locations on the sphere is presented in Figure 2 (to conserve space, 2nd and 3rd order plots are omitted).

To assess the effect of different pre-processing combinations on the accuracy of ILD reproduction in binaural Ambisonic rendering, the ILD between binaural Ambisonic renders and HRIRs for 16,020 directions over the sphere with different pre-processing combinations was estimated using the method in [17]. The solid angle weighted change in ILD (referred to here on in as ∆ILD) between HRIRs and binaural Ambisonic renders with different pre-processing combinations for all directions on the sphere and all tested orders of Ambisons is presented in Figure 3, with whiskers to denote the maximum ∆ILD value. As expected, higher orders of Ambisons produce improved ILD rendering, and in all tested orders of Ambisons, every pre-processing combination improves the solid-angle weighted ∆ILD over standard dual-band decoding. Interestingly however, different orders produce different results for which pre-processing combination offers the best ILD reproduction. For 1st and 3rd order, PP 4 (the dual-band combination of TA, AIO and DFE) produces the greatest improvements with the lowest solid-angle weighted ∆ILD and the lowest value of maximum ∆ILD, but for 2nd order, this is found at PP 3 (the basic channel weighted combination of TA, AIO and DFE). To illustrate how ∆ILD changes depending on the location on the sphere, Figure 4 shows ∆ILD for all locations on the sphere and each pre-processing combination for 1st order Ambisons (to reduce the overall amount of figures, 2nd and 3rd order plots are omitted).

4. PERCEPTUAL EVALUATION

To assess the perceptual effect of different pre-processing combinations, listening tests were conducted using both simple and complex acoustic scenes. The tests followed the multiple stimulus with hidden reference and anchors (MUSHRA) paradigm, ITU-R BS.1534-3 [18]. Tests were conducted in a quiet listening room using an Apple MacBook Pro with a Fireface 400 audio interface, which has software controlled input and output levels. A single set of Sennheiser HD 650 circum-aural headphones were used, which were equalised using Kirkeby and Nelson’s least-
mean-square regularization method [19] from the RMS average of 11 impulse response measurements collected using Farina’s swept sine technique [20] and a Neumann KU 100. The range of inversion was 5 Hz–4 kHz, and in / out-band regularization of 25 dB and −2 dB respectively was employed to avoid sharp peaks in the inverse filters. 20 experienced listeners took part, aged between 22 and 41, with no reported hearing impairments.

4.1 Test Methodology

Tests were conducted using static binaural rendering with no head-tracking implemented. Listeners compared binaural Ambisonic renders created using the pre-processing combinations as throughout this paper. All scenarios were repeated once. Three types of stimuli were used in the listening test. The first was a pseudo-moving pink noise sound, generated using 45 bursts of pink noise played consecutively and lasting 0.05 seconds long each, panned between \((\theta, \phi) = (44^\circ, 0^\circ)\) and \((\theta, \phi) = (132^\circ, 0^\circ)\) in 2° increments. Each burst was windowed using a 50 sample hanning window, resulting in a full pseudo-moving sound lasting 2.25 seconds. The reference was made from direct HRIR convolutions, and a monophonic version of the HRIR reference low-passed at 3.5 kHz was used as the low anchor. The second was a synthesised complex scene which comprised of 8 monophonic percussion tracks panned to 8 of the centre vertices of the faces of a dodecahedron. The reference was created by summing direct HRIR convolutions of the 8 original monophonic tracks, and again a monophonic version of the reference low-passed at 3.5 kHz was used as the low anchor. The third stimuli type was a 5 second excerpt of a fourth-order Ambisonic recordings of a beach soundscape made using an mb acoustics em32 Eigenmike [21]. The recording was converted from Schmidt semi-normalised (SN3D) to N3D normalisation using the method in [13]. As the Eigenmike recording test could not use an HRIR render as a reference, listeners were in this case asked to rate the stimuli in terms of plausibility, which was defined as ‘a simulation in agreement with the listener’s expectation towards a corresponding real event’ [22]. An anchor was included as a monophonic version of the Ambisonic render with no
4.2 Results

Listening test data was checked for normality using the one-sample Kolmogorov-Smirnov test, which showed all data as non-normal. Therefore, results were analysed using non-parametric statistics. Figures 5, 6 and 7 present the median scores for orders 1, 2 and 3 with non-parametric 95% confidence intervals [23] for the moving noise, percussion and beach stimuli types, respectively.

In all scenarios, NPP was rated as the worst Ambisonic condition. To assess the statistical significance of the differences between pre-processing combinations, Friedman’s ANOVA tests were conducted on all test stimuli and orders. For the moving noise stimuli type, statistical significance was only found at 3rd order ($\chi^2(4) = 9.3, p = 0.05$; $\chi^2(4) = 7.3, p = 0.12$; $\chi^2(4) = 16.2, p < 0.01$ for 1st, 2nd and 3rd orders, respectively). For the percussion stimuli type, pre-processing combinations produced statistically significantly different results again only for 3rd order ($\chi^2(4) = 34.2, p < 0.01$ for 1st, 2nd and 3rd orders, respectively). For the beach stimuli type, pre-processing combinations produced statistically significantly different results for 3rd order ($\chi^2(4) = 34.2, p < 0.01$ for 1st, 2nd and 3rd orders, respectively).

5. DISCUSSION AND CONCLUSIONS

This paper has presented an evaluation of pre-processing technique combinations for virtual loudspeaker binaural Ambisonic rendering. It is clear that pre-processing produces an improvement for all tested orders, something that has been shown both numerically and perceptually. However, results are not as simple as to offer a definitive optimal pre-processing combination and therefore warrant further discussion and testing.

A discrepancy between the numerical and perceptual results for 1st order is notable where PP 4 clearly outperformed the other pre-processing combinations in spectral and ILD reproduction, but was not rated the highest for...
any test stimuli type in the perceptual tests.

Perceptual results differed with test stimuli types. PP 2 performed better for the percussion stimuli type, whereas pre-processing combinations with AIO (PP 1, PP 3 and PP 4) performed better for the other two stimuli types. One possible explanation for this is that there was greater lateralisation present in the moving noise and soundscape stimuli.

The improvement gained from implementing TA has been shown to increase with order. This has been reflected in both the numerical and perceptual evaluation results, for all the pre-processing combinations that include TA (PP 2, PP 3 and PP 4), and is a likely factor for the statistical significance in variation of results for 3rd order with all stimuli types.

Future work will look at comparing the pre-processing technique combinations presented in this paper to other state of the art Ambisonic decoding solutions such as Magnitude Least Squares [12] and directional equalisation strategies [24].

6. REFERENCES

Figure 7: Median scores of the beach stimuli tests with non-parametric 95% confidence intervals. Scores indicate perceived plausibility. LA and NPP denote low anchor and no pre-processing, respectively.


