Improving Binaural Rendering with Bilateral Ambisonics and MagLS

Isaac Engel, Dan Goodman, Lorenzo Picinali

Imperial College London, Exhibition Road, London, United Kingdom, Email: isaac.engel@imperial.ac.uk

Introduction

Binaural signals are often rendered from sound fields described in the spherical harmonics (SH) domain, i.e. Ambisonics. In practice, these sound fields typically have a limited spatial order due to hardware constraints or the need for efficient acoustic simulations. To perform a binaural rendering, the sound field must be combined with a head-related transfer function (HRTF), which inherently has a high spatial order. This mismatch causes the HRTF's order to be truncated, leading to audible artefacts in the binaural signal, such as colouration, poor localization, and low loudness stability [1, 2]. This problem can be mitigated by preprocessing the HRTF with various methods. Two state-of-the-art examples are:

1. **Time-alignment** [3]: The time-domain components of the HRTF, or head-related impulse responses (HRIRs), are aligned by means of removing their initial onsets. This requires bilateral Ambisonics to render binaural signals (see next section).
2. **Magnitude Least Squares (MagLS)** [4]: The HRTF's phase is disregarded above a given cutoff frequency (f_c), assuming it is perceptually irrelevant. This method does not require bilateral Ambisonics.

Both these methods aim to reduce the HRTF's effective spatial order, so that less information is lost in the truncation process. Consider, for instance, the HRTF used in this study (see Evaluation for details): a spatial order of $N = 16$ is required to preserve 90% of its energy at 10 kHz (see Fig. 1). However, time-aligning the HRTF reduces that threshold to $N = 5$, and removing its phase altogether (setting it to zero) reduces it further to $N = 3$. Note that the last plot means to illustrate the strong influence of the HRTF's phase on its effective spatial order, but MagLS does not actually set the phase to zero.

Bilateral Ambisonics

Bilateral Ambisonics has been recently proposed as an effective method to produce high-quality binaural renderings from order-limited Ambisonics sound fields [3]. It achieves so by generating two Ambisonics sound fields at the ears' positions, rather than a single one at the head centre (see Fig. 2), and employing time-aligned HRTFs.

It has been shown that this 'ear-centric' coordinate system allows to produce accurate binaural signals at lower spatial orders than the traditional 'head-centric' one (e.g. with MagLS), although at the cost of two downsides. First, the dual coordinate system makes head rotations non-trivial to implement at the rendering stage. While this is a crucial feature when reproducing static (e.g. pre-recorded) sound fields, it might be irrelevant for dynamic simulations where the auditory scene is expected to change and the rendering is recomputed in real time, as the head orientation can be updated alongside the sound field. The second downside of bilateral Ambisonics is the need for two sound fields rather than a single one. While this may inconvenient for actual recordings, as it requires a non-standard microphone setup, it is less relevant in the case of simulations, where two receivers can be easily implemented. Therefore, we do not expect that bilateral Ambisonics will become a prevalent method for the binaural rendering of recorded sound fields, but they could be useful for dynamic binaural simulations.

Proposed method: BiMagLS

A new HRTF preprocessing method, BiMagLS, is proposed to improve bilateral Ambisonics rendering. In essence, it first time-aligns the HRTF and then transforms it to the SH domain via MagLS, which is expected to reduce magnitude errors in the rendered binaural signals. Additionally, a smooth transition around the cutoff is proposed to improve the MagLS method.

Choosing the cutoff frequency

The authors of MagLS recommended to employ a cutoff of $f_c = 2$ kHz, as they observed that listeners could not perceive phase errors above that frequency for typical sound signals [4]. However, as noted by Zotter and Frank [6] (sec. 4.11), it is better to vary f_c so it matches the so-called aliasing frequency (f_a):

$$f_a \simeq \frac{Nc}{2\pi r}, \qquad \text{[Hz]} \qquad (1)$$

where N is the spatial order, c is the speed of sound and r is the distance between the receiver and the origin, i.e. the head radius (in this work, we assume the typical values $c = 343\,\mathrm{ms}^{-1}$ and $r = 0.0875$ m). This is because phase errors increase dramatically above f_a, so there is no point in trying to reconstruct phase accurately in that range and is preferable to optimise for magnitude instead.

However, BiMagLS uses ear-centric coordinates, so r approaches zero and, thus, f_a tends to infinity. Therefore, f_c can be chosen without worrying about order dependence. In this study, the cutoff for BiMagLS is chosen as the maximum value (the least restrictive) between 3 kHz (the most conservative recommendation in [4]) and the f_a calculated for the non-aligned HRTF, as shown in Fig. 3.

Smooth transition (improvement for MagLS)

We use the popular MagLS implementation proposed by Zotter and Frank, which achieves good results through a simple and fast iterative procedure [6] (sec. 4.11). However, one issue of this implementation is that it produces a sharp transition in the HRTF's spectrum at the cutoff frequency, which introduces audible artefacts (e.g. hiss-

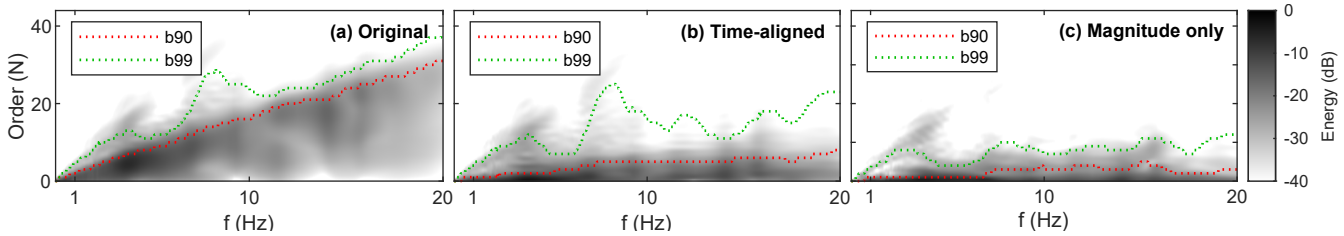

Figure 1: SH spectra [5] of the HRTF employed in this study: (a) before preprocessing, (b) after time-alignment, and (c) after setting phase to zero. The lowest spatial order that contains x% of the HRTF's energy for each frequency is indicated as 'bx'.

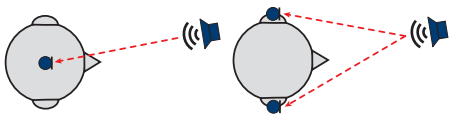

Figure 2: Standard (left) and bilateral (right) Ambisonics.

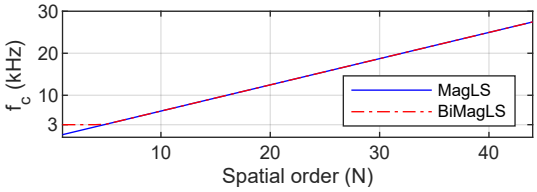

Figure 3: Cutoff frequencies employed in this study for MagLS ($f_c = f_a$) and BiMagLS ($f_c = \max\{3\,\text{kHz}, f_a\}$).

ing) in the binaural signal, as discussed by Ben-Hur et al. in [3].

We propose to define a transition interval of k octaves centred at f_c where the calculation of the SH coefficients changes smoothly from one side of the cutoff frequency to the other. Below f_c, the HRTF's order-N truncated SH coefficients for a given frequency $[\mathbf{H}_N^{\text{low}}(f)]$ are given by the spatial Fourier transform (SFT):

$$\mathbf{H}_N^{\text{low}}(f) = \mathcal{SFT}_N[\mathbf{H}(f, D)] = \mathbf{H}(f, D)\mathbf{Y}_N^{\dagger}(D), \quad (2)$$

where $\mathbf{H}(f, D)$ is the HRTF for a set of directions (D), $\mathcal{SFT}_N$ is the order-N truncated SFT, $\mathbf{Y}_N(D)$ are the real-valued spherical harmonics up to order N for directions D, and $\dagger$ denotes the pseudoinverse. Note that Eq. 2 provides the least-squares solution of the SFT [5].

Above the cutoff, the SH coefficients are obtained from a modified version of $\mathbf{H}(f, D)$ whose phase has been simplified in order to lower its effective SH order (see Fig. 1). For each frequency bin, the target phase is set to the phase that was estimated for the previous bin, which ensures that phase does not undergo large variation as the frequency increases. This process is performed by iterating through the frequency bins, as follows:

$$\phi(f, D) = \angle[\mathbf{H}_N(f)\mathbf{Y}_N(D)], \quad (3)$$

$$\mathbf{H}_N^{\text{high}}(f + \Delta f) = |\mathbf{H}(f + \Delta f, D)|e^{i\phi(f,D)}\mathbf{Y}_N^{\dagger}(D), \quad (4)$$

where Δf is the size of the frequency bins in the HRTF data, $\angle()$ is the argument operator, and $|\cdot|$ is the modulus operator. Note that, so far, the procedure is identical to the one described in [6]. Within the transition interval, we propose to obtain the SH coefficients as a weighted sum of $\mathbf{H}_N^{\text{low}}(f)$ and $\mathbf{H}_N^{\text{high}}(f)$, using a scalar weight (α) between 0 and 1 which is proportional to the frequency and increases logarithmically. Thus, for a transition interval of k octaves, the SH coefficients are obtained as:

$$\alpha(f) = 0.5 + \frac{k\log(f/f_c)}{\log(2)}, \quad (5)$$

$$\mathbf{H}_N^{\text{tr}}(f) = \alpha(f)\mathbf{H}_N^{\text{high}}(f) + [1 - \alpha(f)]\mathbf{H}_N^{\text{low}}(f), \quad (6)$$

where log() is the natural logarithm. Note that α is equal to zero at the lowest point of the transition interval and equal to one at the highest point, which ensures continuity. The final SH coefficients are obtained by iterating through all available frequencies from lowest to highest:

$$\mathbf{H}_N(f) = \begin{cases} \mathbf{H}_N^{\text{low}}(f) & f < 2^{-k/2} f_c \\ \mathbf{H}_N^{\text{tr}}(f) & 2^{-k/2} f_c < f < 2^{k/2} f_c \\ \mathbf{H}_N^{\text{high}}(f) & f > 2^{k/2} f_c. \end{cases} \quad (7)$$

Setting k too small would lead to a less effective mitigation of the artefacts reported in [3], whereas a too large k would increase the phase errors below the cutoff and the magnitude errors above it. Informal listening tests by the authors suggested that $k = 1$ was a good compromise, as it eliminated the artefacts without increasing the magnitude and phase errors importantly (see Fig. 4).

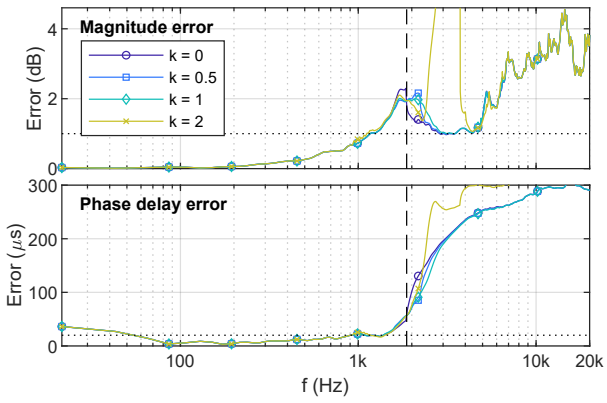

Figure 4: HRTF mean interpolation errors (top: left-ear magnitude; bottom: interaural phase delay) for MagLS, $N = 3$ and varying k. Dotted lines denote the approximate just noticeable differences (JNDs). The dashed line indicates f_a.

Evaluation

We compared the following preprocessing methods in terms of their effectiveness when interpolating an HRTF in the SH domain with a limited spatial order:

1. MagLS: as in [6] (sec. 4.11.2) with $f_c = f_a$ and smooth transition of one octave ($k = 1$).
2. Basic bilateral Ambisonics (BBA): time-aligning the HRTF via ear-alignment, as in [3].
3. BiMagLS: time-aligning the HRTF and then applying MagLS with $f_c = \max\{3\,\text{kHz}, f_a\}$ and $k = 1$.

The three methods were used to interpolate an HRTF (subject 1 of HUTUBS database, acoustically measured for 11950 directions [7]) in the SH domain to a set of target directions (110 points in an approximate Lebedev grid) for orders $N = 1$ to $N = 44$. For the MagLS and BiMagLS methods, the whole HRTF was initially time-shifted so that its mean phase was equal to zero, which helps the MagLS phase estimation—the time shift was later undone on the SH coefficients. The interpolated HRTFs were finally evaluated as follows:

1. An initial analysis of interpolation error and interaural time and level differences (ITD/ILD) was done for $N = 3$. For this, the HRTF was interpolated to 180 equispaced directions in the horizontal plane.
2. Magnitude errors for the 110 target directions were quantified with the perceptual spectral difference (PSD) [8], which translates magnitude differences to sones via equal loudness contours (we assumed 65 dB_{SPL}) and averages them across frequencies via equivalent rectangular bandwidth (ERB) weights.
3. Localisation of anechoic sources for the 110 directions was simulated with the model in [9], implemented in the Auditory Modeling Toolbox [10]. This model employs a Bayesian framework that simulates listener uncertainty, meaning that the estimates for a 'wide' source should display a large variance.

All the code employed in this study is available online at the BinauralSH repository (`github.com/isaacengel/binauralsh`, last viewed on 15/08/2021).

Results

The initial analysis for $N = 3$ revealed that, as expected, both magnitude and phase errors generally increased with frequency, particularly in the case of MagLS phase errors, which increased dramatically above the aliasing frequency (Fig. 5). BiMagLS displayed smaller magnitude errors than BBA, particularly above the 3 kHz cutoff, where the improvements were of approximately 1 to 2 dB. In contrast, BiMagLS displayed larger phase delay errors than BBA (expected due to the MagLS preprocessing), but they were prevalent mostly above 4 kHz and therefore should not be perceived by human listeners [4].

Both bilateral methods displayed accurate ITDs compared to MagLS for $N = 3$ (Fig. 6), which is attributed to the latter producing large phase errors at relatively low frequencies. In contrast, both MagLS and BiMagLS produced more accurate ILDs than BBA, likely due to their magnitude optimisation at high frequencies (Fig. 7).

As shown in Fig. 8, BiMagLS displayed a smaller PSD (proxy for magnitude error) than BBA for all spatial orders, which agreed with the initial analysis. For lateral localisation precision (circular standard deviation), MagLS showed poor results for orders $N \leq 3$, likely due to large phase errors and subsequent degraded ITDs, while the two bilateral methods performed well, with BiMagLS leading by a small margin (more accurate ILDs). Finally, BiMagLS showed the best polar precision, outperforming BBA by 10° for $N < 5$. This is attributed to its lower magnitude errors at critical frequencies for elevation perception, such as between 6 and 9 kHz [12].

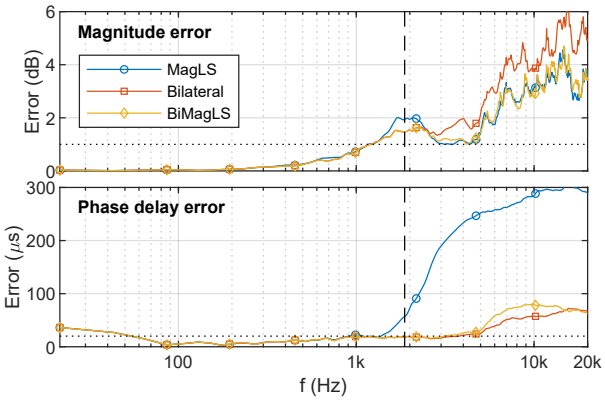

Figure 5: HRTF mean interpolation errors (top: left-ear magnitude; bottom: interaural phase delay) for $N = 3$. Dotted and dashed lines denote the approximate JNDs and f_a.

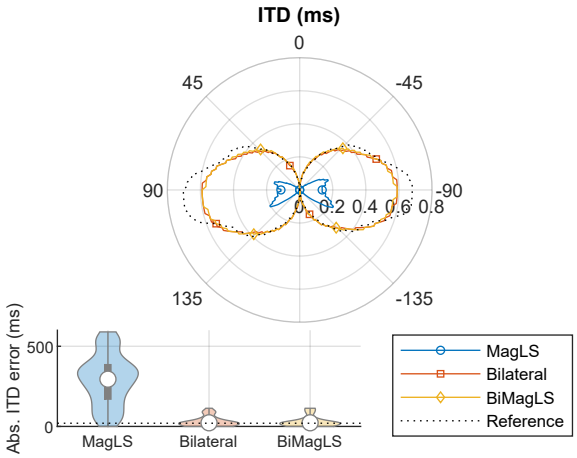

Figure 6: Top: ITDs of the interpolated HRTFs ($N = 3$) and the original HRTF (reference) in the horizontal plane. Bottom: absolute ITD errors, where dotted lines are approximate JNDs in anechoic conditions [11].

Summary and future work

An HRTF preprocessing method for the binaural rendering of order-limited Ambisonics sound fields, BiMagLS, has been proposed. By applying MagLS preprocessing to the time-aligned HRTF, it reduces magnitude errors at high frequencies and achieves a more accurate (model-predicted) localisation of anechoic sound sources.

Also, an improvement to the iterative MagLS method [6] has been proposed, which smooths out the sharp transition at the cutoff frequency by applying a simple scalar weight. Informal tests suggest that this mitigates considerably the audible artefacts (e.g. hissing) reported in [3].

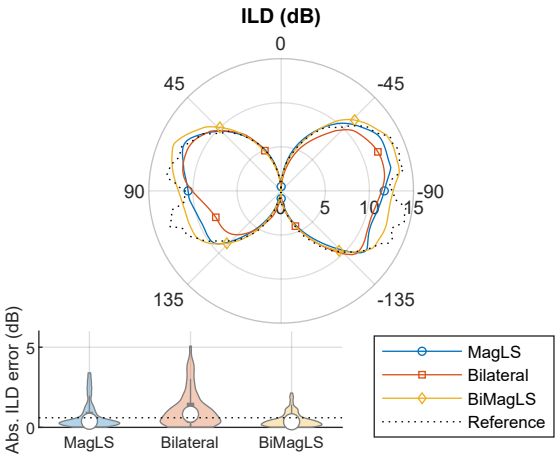

Figure 7: Top: ILDs of the interpolated HRTFs ($N = 3$) and the original HRTF (reference) in the horizontal plane. Bottom: absolute ILD errors, where dotted lines are approximate JNDs in anechoic conditions [11].

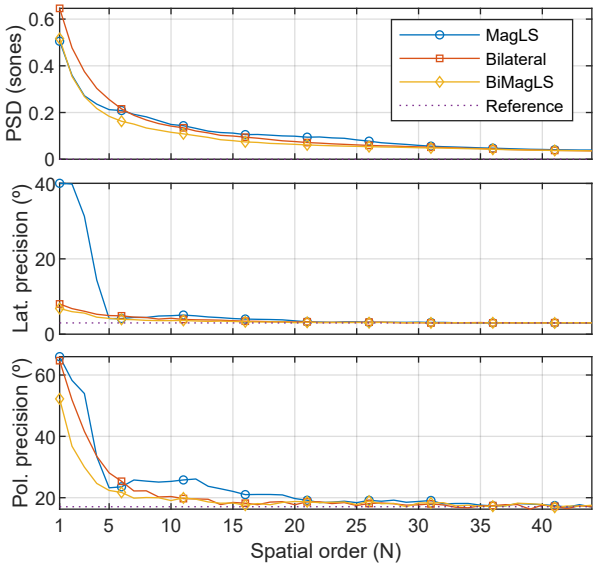

Figure 8: Evaluation results for the interpolated HRTFs, as a function of spatial order. Top: absolute left-ear PSD with respect to the reference HRTF, averaged across the 110 directions. Mid/bottom: predicted lateral/polar localisation precision, i.e. circular standard deviation of the estimates.

As future work, it is suggested to conduct formal listening tests to verify the perceptual models' predictions regarding BiMagLS and to validate the smooth transition method. Reverberant sound fields may be considered, for which the perceptual differences between spatial orders are expected to be smaller than in the anechoic case [13].

References

[1] 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.

[2] Z. Ben-Hur, D. L. Alon, B. Rafaely, and R. Mehra, "Loudness stability of binaural sound with spherical harmonic representation of sparse head-related transfer functions," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, p. 5, 2019.

[3] Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely, "Binaural Reproduction Based on Bilateral Ambisonics and Ear-Aligned HRTFs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 901–913, 2021.

[4] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Fortschritte Der Akustik–DAGA*, 2018, pp. 339–342.

[5] Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely, "Efficient Representation and Sparse Sampling of Head-Related Transfer Functions Using Phase-Correction Based on Ear Alignment," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2249–2262, 2019.

[6] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer Nature, 2019.

[7] B. Fabian, D. Manoj, P. Robert, W. Jan Joschka, S. Fabian, V. Daniel, G. Peter, and W. Stefan, "The HUTUBS head-related transfer function (HRTF) database," 2019.

[8] C. Armstrong, T. McKenzie, D. T. Murphy, and G. Kearney, "A Perceptual Spectral Difference Model for Binaural Signals," in *Audio Engineering Society Convention 145*, 2018.

[9] J. Reijniers, D. Vanderelst, C. Jin, S. Carlile, and H. Peremans, "An ideal-observer model of human sound localization," *Biological Cybernetics*, vol. 108, no. 2, pp. 169–181, 2014.

[10] P. L. Søndergaard and P. Majdak, "The Auditory Modeling Toolbox," in *The Technology of Binaural Listening*, ser. Modern Acoustics and Signal Processing, J. Blauert, Ed. Springer, 2013, pp. 33–56.

[11] S. Klockgether and S. van de Par, "Just noticeable differences of spatial cues in echoic and anechoic acoustical environments," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. EL352–EL357, 2016.

[12] B. Zonooz, E. Arani, K. P. Körding, P. A. T. R. Aalbers, T. Celikel, and A. J. Van Opstal, "Spectral Weighting Underlies Perceived Sound Elevation," *Scientific Reports*, vol. 9, no. 1, p. 1642, 2019.

[13] I. Engel, C. Henry, S. V. Amengual Garí, P. W. Robinson, and L. Picinali, "Perceptual implications of different Ambisonics-based methods for binaural reverberation," *The Journal of the Acoustical Society of America*, vol. 149, no. 2, pp. 895–910, 2021.