A Super-Resolution Ambisonics-to-Binaural Rendering Plug-In

Peter Maximilian Giller[1], Christian Schörkhuber[1]

[1] *Kunstuniversität Graz, Inst. f. Elektronische Musik u. Akustik, 8010 Graz, Österreich,*

Email: peter-maximilian.giller@student.kug.ac.at

Introduction

Binaural reproduction of Ambisonic sound scenes has been studied extensively in the past few years. Yet, the achieved quality for first order Ambisonics (FOA) is often unconvincing. The existing approaches can be divided into static methods, operating independently from the input signal content, and dynamic methods, in which case rendering is based on parameters derived from the input signal. Static rendering can be achieved via Ambisonic decoding for an array of virtual loudspeakers and subsequent convolution with the respective head-related transfer function (HRTF), or, in case of the more recent least-squares (LS) methods, direct rendering via convolution with a decoder matrix obtained by minimizing the squared error between decoder and HRTF for all directions on the sphere. For low Ambisonic orders, the LS method suffers from poor externalization, and, most notably, a severe roll-off towards higher frequencies. As explained later on, both is related to inadequate reproduction of the HRTF. While various methods have been proposed to remedy timbral artifacts, the effectiveness of static rendering methods appears to be limited for lower-order input signals, particularly with regard to spatial resolution for FOA signals. Dynamic methods such as HARPEX [1], DirAC [2], COMPASS [3], or compressed-sensing based methods [4], either render binaural signals directly or may be employed as an intermediate 'upcoding' stage followed by binaural rendering with static methods. Dynamic rendering can be superior to static methods, but often at the expense of introducing artifacts and with the requirement for averaging or decorrelation.

In this contribution we present an open-source audio plug-in for signal-dependent binaural rendering of FOA. The plug-in is based on a recently proposed parametric extension of the constrained LS decoder [5], aiming at exact reproduction of both direct sound impinging from the most prominent source direction and diffuse sound. The quality of reproduction with regard to both spatial and timbral attributes was investigated in a listening experiment.

Least-squares decoder

The superposition of a number of plane waves can be written as the N-th order Ambisonics signal vector $\boldsymbol{z}(\omega,t) = \sum_i s_i(\omega,t)\,\boldsymbol{y}_N(\theta_i)$, where $s_i(\omega,t)$ is the i-th source signal with t and ω denoting time and frequency, and $\boldsymbol{y}_N(\theta_i) = \left[Y_0^0(\theta_i)\cdots Y_N^N(\theta_i)\right]^\mathsf{T}$ is a vector of length $L=(N+1)^2$ containing the real-valued spherical harmonics (SH) evaluated at the incidence directions θ_i. The estimated binaural signal for the left and right ear $\hat{\boldsymbol{x}}(\omega,t) = [\hat{x}_\mathsf{L}(\omega,t)\ \hat{x}_\mathsf{R}(\omega,t)]^\mathsf{T}$ is computed by multiplying a filter matrix $\boldsymbol{W}(\omega) = [\boldsymbol{w}_\mathsf{L}(\omega)\ \boldsymbol{w}_\mathsf{R}(\omega)]^\mathsf{H} \in \mathbb{C}^{2\times L}$ with the input signal vector:

$$\hat{\boldsymbol{x}}(\omega,t) = \boldsymbol{W}(\omega)\boldsymbol{z}(\omega,t) \tag{1}$$

The optimal coefficients $\boldsymbol{W}^*$ shall be determined such that the estimated signal, in a perceptual sense, becomes most similar to the desired signal

$$\boldsymbol{x}(\omega,t) = \sum_i s_i(\omega,t)\,\boldsymbol{h}(\omega,\theta_i), \tag{2}$$

i.e., the source signals s_i filtered with the HRTF $\boldsymbol{h}(\omega,\theta_i) = [H_\mathsf{L}(\omega,\theta_i)\ H_\mathsf{R}(\omega,\theta_i)]^\mathsf{T}$ for the respective direction. For compact notation, the time and frequency indices are omitted in the following. Employing the squared error for all directions on the sphere as a similarity measure, the optimal coefficients are given by

$$\boldsymbol{W}^* = \arg\min_{\boldsymbol{W}} \int_{\theta\in\mathcal{S}^2} \|\boldsymbol{W}\boldsymbol{y}_N(\theta) - \boldsymbol{h}(\theta)\|_2^2\,\mathrm{d}\theta. \tag{3}$$

If the sphere is sampled at a suitable set of quadrature points $\mathcal{M}$, the solution to the discretized problem with the SH matrix $\boldsymbol{Y}_{\mathcal{M}N} = [\boldsymbol{y}_N(\theta_i)]_{\theta_i\in\mathcal{M}} \in \mathbb{R}^{L\times|\mathcal{M}|}$ and the HRTF set $\boldsymbol{H}_\mathcal{M} = [\boldsymbol{h}(\theta_i)]_{\theta_i\in\mathcal{M}} \in \mathbb{C}^{2\times|\mathcal{M}|}$ is given by

$$\boldsymbol{W}^* = \mathsf{SHT}_N(\boldsymbol{H}_\mathcal{M}) = \boldsymbol{H}_\mathcal{M}\boldsymbol{Y}_{\mathcal{M}N}^\dagger, \tag{4}$$

where $\mathsf{SHT}_N(\cdot)$ denotes the order-N spherical harmonics transform, and $\boldsymbol{Y}_{\mathcal{M}N}^\dagger$ denotes the Moore-Penrose pseudo-inverse of $\boldsymbol{Y}_{\mathcal{M}N}$. The product $\boldsymbol{W}\boldsymbol{y}_N$ implicitly contained in the decoder equation (1) can be regarded as a linear combination of a set of orthogonal basis functions (the SHs) weighted by the coefficients $\boldsymbol{W}$ to approximate the HRTF for all directions on the sphere. The accuracy of the approximation depends on the order N. Perfect approximation would require N to be equal or greater than the maximum modal order of the HRTF. However, as spatial complexity of HRTFs increases with frequency, modes up to order 35 substantially contribute to the total energy [6, 7]. A low input order N therefore leads to a spectral roll-off towards high frequencies, making the unconstrained least squares decoder basically inapplicable for FOA. Suggested improvements, which include spherical sub-sampling [6], global diffuse field [8] or inter-aural phase equalization [7], and Magnitude Least-Squares optimization [9], can reduce the aforementioned problems to a certain extent, but tend to be insufficient for FOA.

Implemented Method

Our implementation is based on the recently proposed Linearly and Quadratically Constrained Least-Squares (LQC-LS) approach [5]. In addition to minimizing the least-squares error on the sphere, two constraints are imposed upon the optimization problem in order to yield an accurate reproduction of both direct and diffuse sound. We model a sound field in the Ambisonics domain

$$\boldsymbol{z} = \boldsymbol{z}_{\mathsf{dir}} + \boldsymbol{z}_{\mathsf{diff}} \tag{5}$$

with limited order N as the superposition of a direct component $\boldsymbol{z}_{\mathsf{dir}} = s\,\boldsymbol{y}_N(\theta_0)$ corresponding to a single plane wave s coming from the direction of arrival θ_0 (DOA), and a diffuse component $\boldsymbol{z}_{\mathsf{diff}}$. Similarly, the binaural signal

$$\boldsymbol{x} = \boldsymbol{x}_{\mathsf{dir}} + \boldsymbol{x}_{\mathsf{diff}} \tag{6}$$

is modeled as the superposition of a direct component $\boldsymbol{x}_{\mathsf{dir}} = s\,\boldsymbol{h}(\theta_0)$ and a diffuse component $\boldsymbol{x}_{\mathsf{diff}}$. The estimated binaural signal

$$\hat{\boldsymbol{x}} = \boldsymbol{W}(\theta_0)\,\boldsymbol{z} \tag{7}$$

is obtained by multiplication of the Ambisonic signal with the direction-dependent decoder matrix $\boldsymbol{W}$. A plane wave from direction θ_0 is encoded correctly with respect to the HRTF if $\hat{\boldsymbol{x}}_{\mathsf{dir}} = \boldsymbol{x}_{\mathsf{dir}}$ holds. With $\boldsymbol{x}_{\mathsf{dir}} = s\,\boldsymbol{h}(\theta_0)$ for the desired and $\hat{\boldsymbol{x}}_{\mathsf{dir}} = s\boldsymbol{W}\boldsymbol{y}_N(\theta_0)$ for the estimated binaural signal, a linear constraint on $\boldsymbol{W}$ is attained:

$$\boldsymbol{W}\boldsymbol{y}_N(\theta_0) = \boldsymbol{h}(\theta_0) \tag{8}$$

For correct encoding of the diffuse component, the diffuse field energy and the inter-aural coherence of the estimated signal $\hat{\boldsymbol{x}}$ should match with the desired signal $\boldsymbol{x}$. These parameters are captured by the auto-covariance matrices of the signals. The authors of [5] show that a matrix $\boldsymbol{W}$ can be found which transforms the covariance matrix $\boldsymbol{C}_{\boldsymbol{z}}$ of the input signal to the desired output covariance matrix:

$$\boldsymbol{W}\boldsymbol{C}_{\boldsymbol{z}}\boldsymbol{W}^{\mathsf{H}} = \boldsymbol{C}_{\boldsymbol{x}} \tag{9}$$

Assuming that the direct component is not correlated with the diffuse component, this can be achieved with a second, quadratic constraint

$$\boldsymbol{W}\boldsymbol{W}^{\mathsf{H}} = \boldsymbol{C}_{\boldsymbol{h}}, \tag{10}$$

where $\boldsymbol{C}_h$ is the diffuse field covariance matrix of the HRTF set. Since the solution to these constraints is not necessarily unique, the remaining degrees of freedom are used to minimize the mean squared error for all directions on the sphere analogously to (3) subject to the linear and quadratic constraints (8) and (10).

Precomputation of the filter weights

The decoder requires knowledge of the DOA within each frequency band. Accordingly, the optimization problem would need to be solved for each time-frequency bin. To reduce the amount of computations at runtime, the filter weights are precomputed offline for a dense set of directions, transformed to the SH domain, and stored in an HDF5 file. The SH domain representation of the weights is given by

$$\boldsymbol{W}_{nm} = \int_{\theta\in\mathcal{S}^2} \boldsymbol{W}(\theta) Y_n^m(\theta)\mathrm{d}\theta \tag{11}$$

and the estimated binaural signal can be computed at the expansion order Q with

$$\hat{\boldsymbol{x}} = \sum_{n=0}^{Q}\sum_{m=-n}^{n} \boldsymbol{W}_{nm} Y_n^m(\theta_0)\boldsymbol{z}, \tag{12}$$

where the DOA θ_0 is estimated based on the intensity vector. While the closed-form solution in [5] might also be feasible in real-time, decoupling the optimization from the rendering step brings the advantage of being able to use arbitrary SH-domain weights for rendering in principle. The convenience of the SH representation in comparison to a look-up table is that it does not require explicit interpolation or quantization of the DOA. Furthermore, order truncation may allow for a trade-off between reproduction quality and computational cost if appropriate. Evaluation of the SHs is implemented in an efficient way using pre-factored expressions [10].

Processing of sparse HRTFs

Precomputation of the decoding weights requires evaluating the HRTFs for all directions over the sphere. HRTF interpolation can be performed conveniently using a high-order SHT of the measured HRTF set. However, the number or distribution of sampling points used in the measurement procedure does not always allow for robust computation of the expansion coefficients up to high orders. This is often the case for human subjects where the measurement has to be performed within reasonable time and certain angles are difficult to cover (usually the area beneath the listener). A typical example is shown in Fig. 1. The sound sources were located on an arc rotated around the inter-aural axis. The measurement grid is less dense at lateral positions, and the area around the south pole is not covered at all. Fig. 1 (*left*) shows the magnitude of the right ear at 1 kHz. The SHT of an HRTF $\boldsymbol{H} = [\boldsymbol{h}(\theta_i)]_{\theta_i\in\mathcal{M}}$ can be computed using the LS method via multiplication with the pseudo-inverse $\boldsymbol{Y}_N^{\dagger}$ of the SH matrix $\boldsymbol{Y}_N = [\boldsymbol{y}_N(\theta_i)]_{\theta_i\in\mathcal{M}}$. In case of inappropriate sampling, $\boldsymbol{Y}_N^{\dagger}$ may be ill-conditioned for high orders N, leading to numerical instability. The inverse

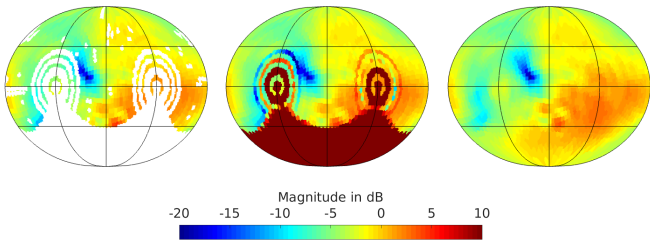

Figure 1: *Left to right*: Interpolated magnitude at 1 kHz (right ear) of: the original HRTF, the iSHT of the original data, and the iSHT of the complemented data.

SHT (iSHT) of the original data at order $N = 20$ shown in Fig. 1 (*center*) exhibits excessive levels in the sparsely covered areas. One possible approach is to regularize the least squares fit as shown in [11], but this method is likely to reduce the accuracy at the existing data points. The authors of [12] propose to first use a non-regularized low-order SH approximation to generate additional data points for missing directions. After complementing the HRTF set with the new samples, a higher-order SHT can be performed on the combined set. While, naturally, this method can not reconstruct missing data, it ensures numerical stability, has little effect on the measured samples, does not lead to severe artifacts in the unknown regions, and is computationally cheap. In our present implementation, the complementary positions are determined by comparison with a spherical t-57 design. New data points are generated at every coordinate further away than 3° from any of the available points. As Fig. 1(c) (*right*) shows, the magnitude of the complemented data is plausible in the unknown regions, and the original values are not noticeably altered.

Listening Experiment

We evaluated the present implementation of the LQC-LS method in a listening experiment. We created a virtual sound scene using Ambisonic room impulses (ARIRs) of a simulated room with volume $V = 479\,\mathrm{m}^3$, reverberation time $T_{30} = 1.4\,\mathrm{s}$, and, accordingly, a critical distance of $r_c \approx 1\,\mathrm{m}$. The sound scene consisted of two speakers simultaneously speaking. Two different anechoic speech signals were convolved with the ARIRs of omni-directional sources located at a distance of $r = 3r_c$ and at the angles $\varphi = \{90°, -30°\}$. We manipulated the direct-to-reverberant energy ratio (DRR) in order to create the following three conditions: (*i*) anechoic condition, created by removing the reverberant part of the ARIRs ($DRR \to \infty$), (*ii*) moderate reverberation with $DRR = 0\,\mathrm{dB}$ as expected at distance r_c, created by damping the reverberant part by 6 dB, and (*iii*) strong reverberation with $DRR = -6\,\mathrm{dB}$ corresponding to $3r_c$ using the unmodified ARIRs. Conditions with multiple sources and varying reverberation were chosen because methods relying on DOA estimation in the time-frequency domain tend to be sensible regarding DOA fluctuations. A reference signal rendered with order $N = 35$ using the LS method was compared with different binaural signals in a MUSHRA-like test with hidden reference and two anchors. We compared the LQC-LS decoder with SH-domain weights using SH orders $Q = 20$ and $Q = 10$ (considered sufficient), and $Q = 3$ (considered insufficient). Furthermore, the Magnitude Least-Squares (MagLS) decoder [9] was compared as a state-of-the-art static rendering method.[1] The unconstrained LS decoder as well as the omni-directional component of the Ambisonics signal (Mono) were used as anchors. All stimuli except for the reference were generated with first-order input signals and Neumann KU-100 HRTFs with 2702 measurement positions [13] were used in the experiment. The experiment consisted of two parts: the subjects were asked to rate the similarity to the reference (*i*) regarding the reproduction of *space* ('externalization, distance, and direction of the sources'), and (*ii*) regarding the reproduction of *sound* ('sound color and quality of reproduction') on a scale from 0 to 100, where 100 corresponds to *identical*, 50 to *different*, and 0 to *very different*. 17 experienced listeners less than 40 years old participated in the experiment.

[1] Implementation available: `https://plugins.iem.at/docs/plugindescriptions/#binauraldecoder`

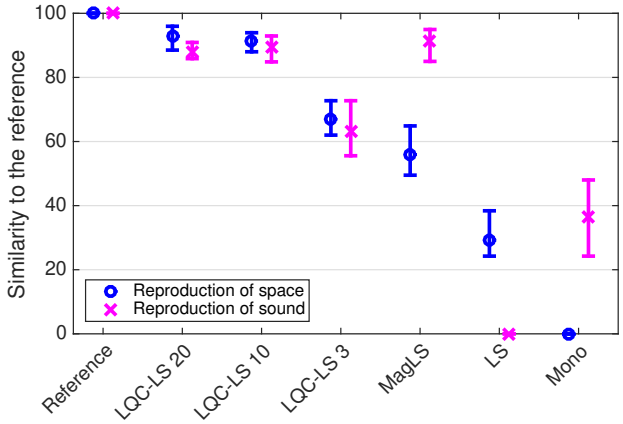

Figure 2: Median and 95 % confidence intervals of the similarity ratings for combined room conditions.

Discussion of the results

Fig. 2 shows the median and 95 % confidence intervals of the normalized ratings with all DRR conditions combined.

Reproduction of space. A Wilcoxon signed rank test showed that the differences between all ratings are significant, except for LQC-LS 20 and 10. Both LQC-LS 20 and 10 are rated very similar to the reference. As expected, the order reduction to $Q = 3$ leads to a lower rating, however, it is still rated higher than the MagLS decoder.

Reproduction of sound. The test showed that all ratings are significantly different except for LQC-LS 20/10, and MagLS. The MagLS decoder is rated much higher compared to the first part, while the rating of LQC-LS 20/10 is still high, but with a downward tendency.

An Analysis of Deviations (ANODE) revealed a significant interaction between the factors DRR and rendering method. Fig. 3 shows the ratings of the two parts for each of the three conditions. It can be observed that the similarity rating of MagLS improves if the DRR is decreased. This is likely caused by different effects: The resolution of direct sound becomes less important in a diffuse sound field while, at the same time, externalization is supported by reverberation. Furthermore, the sound color of the static MagLS decoder optimized for all directions on the sphere can be expected to become more similar to the reference in increasingly diffuse conditions.

While LQC-LS 20/10 is also rated very similar to the reference, there appears to be a slight downward tendency for diffuse conditions. This might be due to the fact, that the increase of diffuse sound leads to an increase of both

errors and fluctuations of the estimated DOA, probably affecting the spatial reproduction and leading to a less transparent sound. However, no severe artifacts like, e.g., crackling or musical noise have been reported. On the contrary, the upward tendency of ratings for the LQC-LS 3 with decreasing *DDR* might be due to a reduced effect of order reduction on the reproduction of diffuse sound or smaller differences of order-reduced beams between neighboring directions.

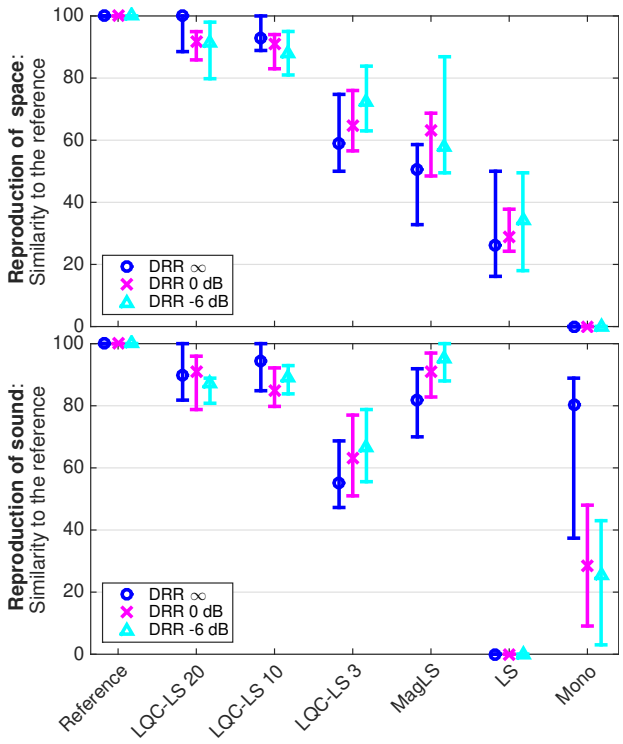

Figure 3: Median and 95 % confidence intervals of the similarity ratings regarding space (*top*) and sound (*bottom*) for each of the three room conditions.

Conclusion

In this contribution we presented the implementation of the recently proposed LQC-LS method. We evaluated the implemented method in a listening experiment for a speech scenario with varying reverberation. The results show a high similarity rating of the implemented method to the reference signal with regard to both spatial reproduction and sound quality. Although the formulation in [5] is for arbitrary orders, we decided to restrict the present implementation to FOA since the benefit of the signal-dependent rendering is most noticeable at low orders while the computational cost increases with increasing order. The implementation allows the use of custom HRTFs in the SOFA format [14]. However, we recommend using high-resolution artificial head HRTFs. Our implementation, the *AdaptiveBinauralDecoder*[2], is available open-source and free of charge as part of the *IEM Plug-in Suite*.

[2] `https://plugins.iem.at/docs/adaptivebinauraldecoder`

References

[1] N. Barrett and S. Berge, "A new method for B-Format to binaural transcoding," in *Audio Eng. Soc. Int. Conf.: Spatial Audio: Sense the Sound of Space*. Audio Eng. Soc., 2010.

[2] M.-V. Laitinen and V. Pulkki, "Binaural reproduction for directional audio coding," in *Appl. of Signal Process. to Audio and Acoust.* IEEE, 2009.

[3] 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)*. IEEE, 2018, pp. 6802–6806.

[4] A. Wabnitz, N. Epain, and C. T. Jin, "A frequency-domain algorithm to upscale ambisonic sound scenes," in *IEEE Int. Conf. on Acoust., Speech and Signal Process.* IEEE, 2012, pp. 385–388.

[5] C. Schörkhuber and R. Höldrich, "Linearly and quadratically constrained least-squares decoder for signal-dependent binaural rendering of ambisonic signals," in *Inter. Conf. on Immersive and Interactive Audio*. Audio Eng. Soc., 2019.

[6] 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.

[7] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural rendering of Ambisonic signals by HRIR time alignment and a diffuseness constraint," in *J. Acoust. Soc. Am.*, 2018.

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

[9] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural rendering of ambisonic signals via magnitude least squares," in *Conf.: Fortschritte der Akustik, Munich*. DAGA, 03 2018.

[10] P.-P. Sloan, "Efficient spherical harmonic evaluation," *J. of Comput. Graph. Techn. (JCGT)*, vol. 2, no. 2, pp. 84–83, September 2013.

[11] D. N. Zotkin, R. Duraiswami, and N. A. Gumerov, "Regularized HRTF fitting using spherical harmonics," in *Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*. IEEE, 2009, pp. 257–260.

[12] J. Ahrens, M. R. Thomas, and I. Tashev, "HRTF magnitude modeling using a non-regularized least-squares fit of spherical harmonics coefficients on incomplete data," in *Signal & Information Process. Assoc. Annu. Summit and Conf.* IEEE, 2012.

[13] B. Bernschütz, "A spherical far field HRIR/HRTF compilation of the Neumann KU 100," in *Proc. of the DAGA*, 2013, p. 29.

[14] P. Majdak *et al.*, "Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions," in *Proc. of the 134th Conv. of the Audio Eng. Soc.* Audio Eng. Soc., 2013.