Improving Externalization in Ambisonic Binaural Decoding

Daniel Rudrich and Matthias Frank
Institute of Electronic Music and Acoustics
University of Music and Performing Arts Graz, Austria
Email: {rudrich, frank}@iem.at

Introduction

In the recent years, Ambisonics has become the standard format for virtual reality and 360° video. One reason is its ability to follow head-rotations during playback via headphones, without the need of potentially artifact-afflicted filter switching. These static filters are based on head-related impulse responses (HRIRs), which describe the direction-dependent path from a sound source to the listener's ears under free-field conditions. Binaural Ambisonics often lacks externalization, the impression of a sound source positioned at a certain distance outside the head. However, the externalization of virtual sources is essential for a plausible and realistic experience in a virtual or augmented reality.

The phase information of HRIRs at higher frequencies require a very high spatial resolution. In terms of spherical harmonics, only Ambisonic signals with orders over 30 would correctly represent this information. However, this information is perceptually not necessary, as human hearing does not rely on phase relations in the higher frequency range [1]. Recent developments in binaural Ambisonic rendering for headphones facilitate this property and manipulate the high frequency phase in order to lower the necessary Ambisonic order. This so-called magnitude least-squares (MagLS) approach proposed in [2] also renders personalized HRIR data correctly, without any perceptual differences to direct HRIR rendering. Consequently, with state-of-the-art methods, Ambisonic binaural rendering can be ruled out as a reason for externalization problems.

Perception of distance primarily relies on the direct-to-reverberant energy ratio [3]. A source located directly at the ear canal creates a high ratio, as the reverberant part is neglectably low in comparison to the direct energy. This also happens in an anechoic room without any wall reflections contributing to the reverberant part. In both situations and without a given visual cue, the impression of distances breaks down. It is only through reflections, created by a source interacting with its surrounding room, that enables externalization [4–6].

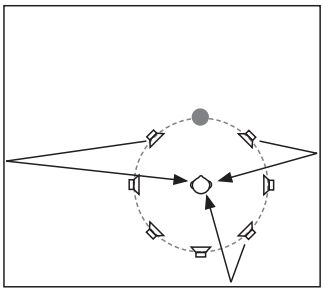
In general, synthetically created audio scenes often include dry sounds without room information, and also HRIR sets for rendering are measured under anechoic conditions. We pin down externalization problems to the lack of reflections, and propose a method to generically create appropriate room reflections in order to improve externalization of dry audio scenes. Our aim is low computational complexity, e.g. for application on mobile devices, and little coloration of the original audio material.

Room Rendering Methods

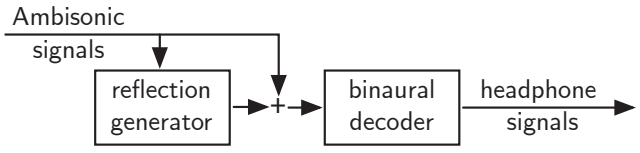
Level, direction, and timbre of room reflections depend on the acoustical properties of the surrounding walls, and the positions of the sound source and receiver in the room. Hence, simulated room reflections should also be influenced by the source's directions within the Ambisonic mix. As Ambisonics is a scene-based method, there's no access to the individual sources within, so a generic solution is necessary.

The proposed strategy to improve externalization in binaural playback of Ambisonics is depicted in Figure 1a. The Ambisonic signals are decoded to virtual loudspeakers which are positioned in a virtual room. The room reflections excited by these loudspeakers are then simulated and encoded back into Ambisonics. The direct path of the room simulation is not rendered, but the original signal is added before the binaural decoder with the appropriate delay, cf. Figure 1b. This is done to preserve the favorably properties of the MagLS binaural decoder that might be otherwise compromised by rendering to the virtual loudspeakers.

In case the Ambisonic orders of direct path and reflections differ, a separate binaural decoder for each path can be used. Here, third-order Ambisonics is used for both paths. Head-tracking-controlled scene rotation is applied within the *binaural decoder* block to also rotate the listener inside the simulated room.

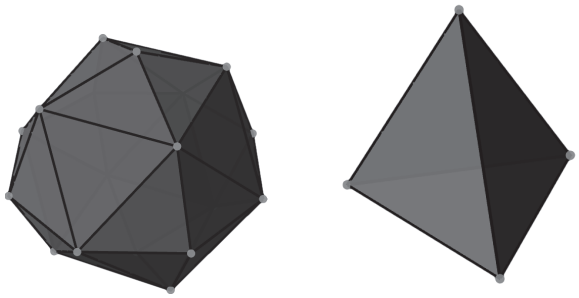


(a) Qualitative illustration of the virtual room layout. Reference signals (grey dot) are located in the room with the same distance as the virtual loudspeaker layouts (radius 1.75 m). Layouts are located off-center, with listener in the middle of the array.



(b) Signal flow of used rendering method

Figure 1: Basic strategy for improving externalization.



(a) third-order layout with two imaginary loudspeakers

(b) first-order layout (tetrahedral arrangement)

Figure 2: Virtual loudspeaker layouts for third-order and frist-order Ambisonic decoding.

Reflections are both simulated and encoded into Ambisonics using the IEM RoomEncoder[1] plug-in, with three different virtual loudspeaker decoding strategies described below. The virtual room has a size of 7.4 m (depth), 8.3 m (width), and 3.0 m (height). Its reflection attenuation was set to −4.0 dB with an additional attenuation for frequencies above 5 kHz by −3.7 dB, resulting in a RT_{60} reverberation time of 230 ms below 5 kHz, and 166 ms above. Direct path *zero-delay* and *unity-gain* options were enabled in the plug-in, so that the system didn't introduce any latency or attenuations. The direct path itself was excluded from rendering in the virtual room, as it was handled separately.

The virtual loudspeaker decoding was realized in three strategies, which differ in number of loudspeakers, and consequently in the used Ambisonic order of the decoder. The first one decodes the full third-order Ambisonic signal to 16 virtual loudspeakers, which are located on three equiangular rings: eight loudspeakers at 0° elevation, four at -45°, and four at +45° elevation; with two additional imaginary loudspeakers at zenith and nadir for use with the IEM AllRADecoder plug-in. This loudspeaker layout is shown in Figure 2a and is labeled **o3** in the following.

The second variant decodes only the first-order part of the Ambisonic signals to four loudspeakers with a tetrahedral arrangement, shown in Figure 2b. With only first order, the spatial resolution of the excitation signal is lowered, and with only four loudspeakers exciting the room, this variant has a lower effective number of reflections. This method is labeled **o1**. Both layouts have a radius of 1.75 m and are positioned off-center in the virtual room, with the listener being positioned in the center of the layout as depicted in Figure 1a.

The third variant requires the lowest computational cost: it uses only the first channel of the Ambisonic signal, the omni component, placing it directly at the listener position in the virtual room. With the direct path disabled, it can be seen as the room response when the listener is omni-directionally radiating sound. As it uses mono, or zeroth-order Ambisonics, it is labeled **o0**.

[1]The RoomEncoder plug-in is part of the free and open-source IEM Plug-in Suite: `https://plugins.iem.at/`

Listening Experiment I

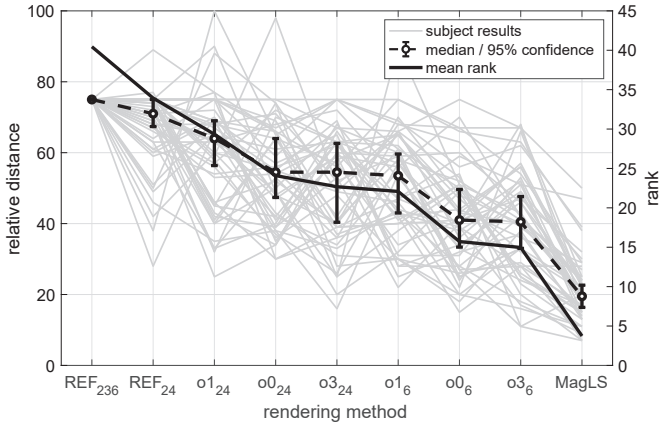
The first listening experiment investigates the effect of the different rendering methods described above on distance perception, timbre, and localization. A MUSHRA-like paradigm was chosen to compare the following ten stimuli:

- $\mathbf{REF_{236}}$ and $\mathbf{REF_{24}}$: direct auralization at a distance of 1.75 m (loudspeaker layout radius) with 236 and 24 reflections, which equals a 7th- and 2nd-order image source model, respectively
- $\mathbf{o3_{24}}$ and $\mathbf{o3_6}$: third-order decoding strategy with 24 and 6 reflections (2nd and 1st order image source model, respectively)
- $\mathbf{o1_{24}}$ and $\mathbf{o1_6}$: first-oder decoding strategy with 24 and 6 reflections
- $\mathbf{o0_{24}}$ and $\mathbf{o0_6}$: zeroth-oder decoding strategy with 24 and 6 reflections
- **magLS**: dry binaural decoding without any reflections, using the MagLS approach
- **stereo**: mid/side decoding of the first two Ambisonic channels (stereo, anchor)

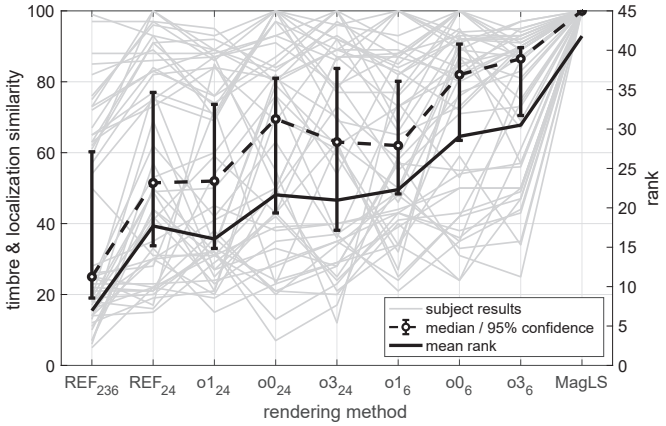
The experiment was divided into two parts, each consisting of five trials with different source directions (azimuth: 90°, 45°, 0°, -45°, -90°; elevation 0°). For the source signal, male speech was used [7]. In the first part, the subjects were asked to rate the perceived distance of the stimuli in comparison to the reference REF_{236} (no hidden reference). The rating scale consisted of five labels: *very close*, *closer*, *little closer*, *same distance*, and *further away*; with a quantization of 100 steps. In the second part, the subjects rated both the effects on timbre and localization in comparison to the reference magLS (also no hidden reference). The rating scale of the second part consisted of three labels: *very different*, *different*, and *identical*; again with 100 steps.

Ten experienced listeners (staff and students of the IEM) participated in the experiment, listening to the stimuli via headphones. Head-tracking was enabled and the participants were asked to restrict their head-movements to only small rotations.

The results are depicted in Figure 3. The individual ratings presented in light grey show a large deviation for the attribute *timbre/localization*, which are caused by inter-subject differences. Hence, the non-parametric Friedman test was carried out to test for a significant main effect for the factor *rendering method*. The mean ranks are presented as solid lines. Data shows an inverse relationship between distance perception and signal coloration: the more distant a rendering method was rated, the stronger the coloration of the audio signal. In general, a low number of reflections introduces only little coloration, however with a smaller perceived distance.



(a) distance; 0 *very close*, 100 *further away*, 75 *same as reference*



(b) timbre/localization; 0: *very different*, 100 *identical* (reference)

Figure 3: Results of the first listening experiment. Light grey lines show the individual ratings of the ten subjects, median with 95% confidence intervals are connected with dashed lines, mean ranks of the Friedman test are presented with solid lines.

The Friedman test showed a highly significant main effect *rendering method* for both ratings *distance* and *timbre/localization* ($p \ll 0.001$). Bonferroni-Holm corrected p-values of pairwise comparisons using the Wilcoxon signed-rank test are given in Table 1. These results show that the virtual loudspeaker approach can achieve the same performance as a direct auralization ($o3_{24}$ and REF_{24}). The number of simulated reflections adjusts the trade-off between distance and coloration. With a fixed number of reflections, first-order decoding (o1) performs better than third and mono (o3 and o0).

The reduced distance of the third-order decoding is assumed to be caused by an increase of low frequencies due to similar impulse responses of neighboring virtual loudspeakers. Informal listening by the authors using music with prominent low-frequencies content confirmed the coloration due to the additive reflections at lower frequencies. An analysis of impulse responses revealed a gain of up to 5 dB below 200 Hz. Former studies [6, 8] have shown that correlated bass in binaural playback can reduce externalization. It is also shown that near-field HRIRs have a pronounced low frequency gain in comparison to free-field HRIRs [9].

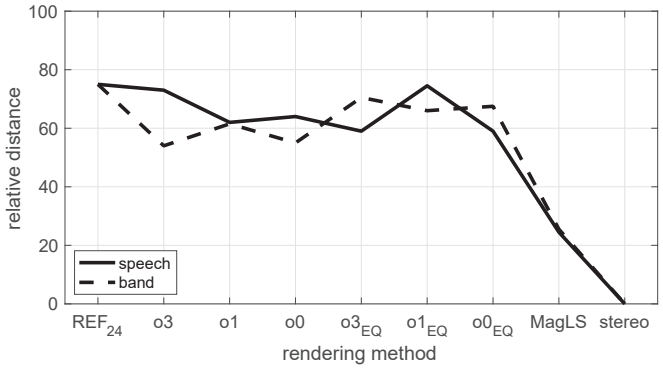
Listening Experiment II

The second listening experiment investigates the effect of high-pass filtering the additive reflections on distance perception, timbre, and localization. Here, two different Ambisonic scenes were played back: male speech, as before, encoded at 45° azimuth, and a band setting (drums, bass, guitar, and brass at 45°, -45°, -135°, and 135° azimuth, respectively; audio signals from *Spicy Funk Cake* by *Rhythmussportgruppe* from [10]). The band setting was chosen, as it contains more lower frequency content as the speech signal, and also to evaluate the rendering methods for a more complex scene.

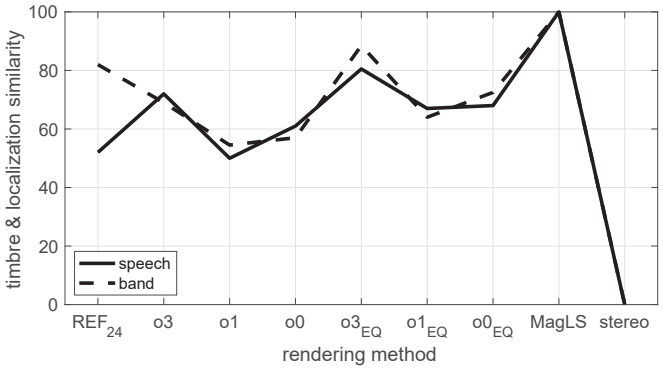
Again a MUSHRA-like paradigm was chosen to compare the following nine rendering methods:

- **REF_{24}**, **$o3_{24}$**, **$o1_{24}$**, **$o0_{24}$**, and **magLS**: same as in experiment I
- **$o3_{24}$EQ**, **$o1_{24}$EQ**, and **$o0_{24}$EQ**: high-pass @ 200 Hz
- **stereo**: M/S decoded, with low-pass filter @ 2.8 kHz

The results of the second experiment are shown in Figure 4. A Friedman test yielded a significant effect of *rendering method*, although Bonferroni-Holm corrected pairwise comparisons with Wilcoxon signed-rank test showed no interesting relationships.



(a) distance ratings



(b) timbre/localization ratings

Figure 4: Ratings of the second listening experiment. Solid and dashed lines show median ratings for speech and band signals, respectively.

Table 1: Bonferroni-Holm corrected p-values of pairwise comparisons of the first listening test using Wilcoxon signed-rank test; significant differences are marked bold; lower triangle shows results for distance, upper one for timbre.

	REF_{236}	REF_{24}	$o3_{24}$	$o3_6$	$o1_{24}$	$o1_6$	$o0_{24}$	$o0_6$	MagLS
REF_{236}		**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
REF_{24}	**<0.001**		1.000	0.555	0.400	0.253	**<0.001**	**<0.001**	**<0.001**
$o3_{24}$	**<0.001**	0.0967		0.253	0.250	**0.006**	**<0.001**	**<0.001**	**<0.001**
$o3_6$	**<0.001**	**<0.001**	0.085		1.000	1.000	**0.002**	**<0.001**	**<0.001**
$o1_{24}$	**<0.001**	**<0.001**	0.050	1.000		1.000	**0.002**	**<0.001**	**<0.001**
$o1_6$	**<0.001**	**<0.001**	**0.008**	0.730	1.000		**0.015**	**<0.001**	**<0.001**
$o0_{24}$	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.024**	**0.005**		0.870	**<0.001**
$o0_6$	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.006**	1.000		**<0.001**
MagLS	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	

In order to study the effect of high-pass filtering on distance and timbre, Friedman tests were carried out for each combination of attribute (*distance*, *timbre/localization*) and source signal (*speech*, *band*). The test showed significant differences between filtered and non-filtered versions of $\mathbf{o3_{24}}$, $\mathbf{o1_{24}}$, and $\mathbf{o0_{24}}$ for the combinations *distance-band* ($p = 0.0011$), *timbre-speech* ($p = 0.0057$), and *timbre-band* ($p = 0.0011$). Only combination *distance-speech* showed no significant differences. As the band setting contains more lower frequency contain, these differences are more distinctive than with the speech signal.

Results show that high-pass filtering of the additive reflections not only improves coloration of signals with lower frequency content, but also increases the perceived distance of sources to the listener; both desired effects without a trade-off as with the results of the first listening experiment.

Summary & Outlook

This study has proposed and evaluated methods for improving externalization in binaural playback of dry Ambisonic signals. Listening tests rating perceived distance and timbral similarity showed an inverse relationship of the favorable attributes distance and audio quality: an increase in perceived distance is accompanied by a stronger coloration of the signal. With the number of generated reflections, the trade-off between these attributes can be controlled.

Coloration due to the added room was more noticeable in the lower frequency range. A second listening test has shown the effect of high-pass filtering of the additive reflections: a reduction of coloration with an increase in perceived distance.

In general, only a small amount of reflections, and also a low Ambisonic order when decoding to the virtual loudspeakers is sufficient for satisfying results. Even reflections excited by a mono signal (zero order; **o0**) create the desired effect, which makes the method even more efficient.

We plan to implement these methods as an optional feature in a future release of the IEM BinauralDecoder, with the option to choose between the different rendering methods and number of reflections.

References

[1] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3616–3627, 2018.

[2] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Fortschritte der Akustik - DAGA*, 2018.

[3] A. Kolarik, S. Cirstea, and S. Pardhan, "Discrimination of virtual auditory distance using level and direct-to-reverberant ratio cues," *J. Acoust. Soc. Am.*, vol. 134, no. 5, pp. 3395–3398, 2013.

[4] R. Crawford-Emery and H. Lee, "The Subjective Effect of BRIR Length on Perceived Headphone Sound Externalization and Tonal Coloration," in *Audio Engineering Society Convention 136*, 2014.

[5] F. Völk, "Externalization in data-based binaural synthesis: effects of impulse response length," in *Fortschritte der Akustik - DAGA*, pp. 1075–1078, 2009.

[6] J. Catic, S. Santurette, and T. Dau, "The role of reverberation-related binaural cues in the externalization of speech," *J. Acoust. Soc. Am.*, vol. 138, no. 2, pp. 1154–1167, 2015.

[7] EBU, "EBU SQAM CD: Sound Quality Assessment Material recordings for subjective tests," 2008.

[8] F. Zotter and M. Frank, "Low-frequency trick to improve externalization with non-individual HRIRs," *Fortschritte der Akustik - DAGA*, 2018.

[9] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources. head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1465–1479, 1999.

[10] D. Leckschat, C. Epe, and N. Dahlheimer, "Komposition und Studioproduktion von Musikstüken des Jazz/Funk-Genre zur Verwendung als Stimuli in virtuellen Umgebungen ," in *Fortschritte der Akustik - DAGA*, 2018.