Audio Engineering Society

Convention Paper 10325

Presented at the 148th Convention,
2020 June 2-5,Online

Optimized binaural rendering of Next Generation Audio using virtual loudspeaker setups

Felix Lau[1], Michael Meier[2]

[1,2] *Institut für Rundfunktechnik, Floriansmühlstraße 60, 80939 München*

Correspondence should be addressed to Felix Lau (felix.lau@irt.de)

ABSTRACT

A binaural rendering addon based on the EBU ADM Renderer has been developed, which renders Next Generation Audio scenes using virtual loudspeaker systems. During development, two optimization approaches emerged regarding rendering quality and efficiency. The first approach concerns the rendering of coherent signals from different emitter positions, which is particularly relevant for virtual speakers since object positioning is based on amplitude panning. The second approach concerns the reduction of computational costs when rendering the binaural room response by using different virtual speaker layouts for the rendering of the direct path and of the room response. To evaluate both approaches, a listening test was conducted. The results of this test showed that each approach positively influenced either rendering quality or performance.

1 Introduction

The current development of Next Generation Audio (NGA) offers a multitude of improvements in production, transmission and playback of audio productions by combining object-, channel- and scene-based audio in one format [1]. In particular, spatially complex productions can benefit from NGA, as it is no longer necessary to commit to a specific playback system during production. The playback of such productions, however, usually requires a quite complex multi-channel playback system, which is generally linked to higher costs and space requirements for the listener. One solution to this problem can be binaural rendering, as it only requires headphones and a renderer for playback to provide a comparable listening experience [2, 3]. For this purpose, the high availability of headphones, as well as smartphones, which can be used as renderers, offers a great advantage [4]. Since NGA-productions, compared to former channel-based workflows, are not necessarily bound to a specific playback system, it is possible to access all metadata of audio objects during the binaural rendering, even at the consumer side. With a channel-based workflow, these metadata are only accessible during the production.

Based on the EBU ADM Renderer (EAR) [5], a file-based NGA renderer for Audio Definition Model (ADM)-based formats [6], a binaural rendering addon was developed, which uses the core functionality of the EAR, which is mainly the rendering to different loudspeaker setups [7], to simulate a virtual, three-

dimensional loudspeaker setup. The complete rendering therefore consists of two parts, the loudspeaker rendering and the binaural rendering (see Fig 1).
The virtual loudspeaker approach mainly provides advantages in terms of the required computing capacity, since only the number of virtual loudspeakers needs to be rendered. This way, the processing power required for the binaural rendering is constant. For the positioning of objects between the virtual loudspeakers, amplitude panning is required. This functionality, however, is already provided by the EAR itself.

2 Binaural rendering structure

The binaural rendering is divided into three parts, the direct path, the room response and a non-binaural part, which have each assigned their own loudspeaker rendering. For the direct path, a free-field Head Related Transfer Function (HRTF) is used, which can be exchanged for the utilization of individual HRTFs. For the room response, binaural room impulse responses (BRIR) are used, which have been acquired through an acoustic simulation. This simulation is based on a virtual reference listening room and a HRTF of a KU100 dummy-head. It was carried out with the software AUVIS, which is an internal software of the Institut für Rundfunktechnik (IRT) for the simulation of studio environments [8, 9]. From the resulting BRIR, the first 60ms [10] (without the direct path) are used for the binaural rendering. Unlike the HRTF for the direct sound component, the HRTF, which is part of the BRIR, cannot be exchanged. In internal listening tests, however, this was considered reasonable, since the perception of the binaural rendering is dominated by the direct path.

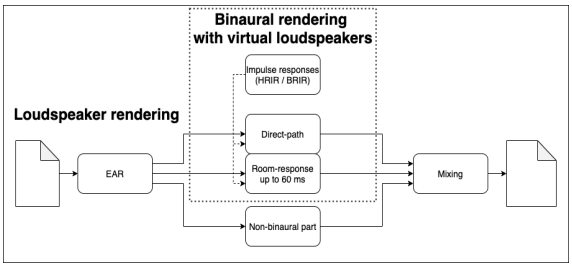

Fig 1 Binaural rendering structure

Furthermore, a non-binaural rendering path is used for enabling the usage of in-head localization, as well as distance perception by crossfading between the binaural and the non-binaural parts. For this part an adjusted stereo rendering is used to maintain correct object positioning on headphones.
After rendering, all three parts are mixed together, with time and gain relations being automatically adjusted according to the used HRTF.
The allocation into three parts provides some general advantages for the binaural rendering like the control over the direct-to-reverberant ratio by adjusting the ratio between the direct path and the room response, as well as the opportunity to process each part individually without affecting any other parts. This allocation is also the foundation for the two optimization approaches, which are described in the following.

3 Alignment of HRTFs

During the development of the binaural rendering addon, two optimization approaches emerged that aim to improve the rendering quality and efficiency. The first approach concerns the alignment of impulse responses from different emitter positions in a HRTF dataset to improve the summing of coherent signals after rendering with these impulse responses. The necessity for this alignment is the varying temporal relation between different emitter positions in relation to the ears. The origin of this relations can be found in the measurement method of HRTFs, in which

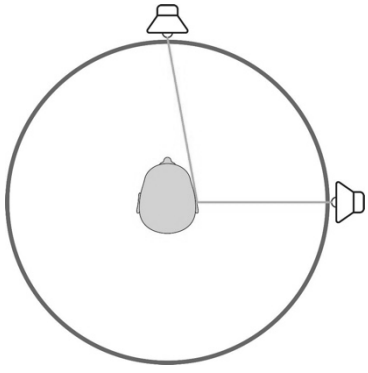

Fig 2 Illustration of different distances from 2 loudspeakers to the listeners ear

emitters are usually positioned on a sphere whose center point equals the center of the head. However, since the ears are located on the side of the head, the distance between emitter and the ears differs from position to position, which causes different time of arrivals (TOAs) for each emitter position (see Fig 2). The problem of different TOAs from different emitter positions in a set of HRTFs has already been discussed in connection with the interpolation of HRTFs. Matsumoto et al., in this context, describes an improvement of the interpolation by prior temporal alignment [11]. Since amplitude-panning of signals can be described as a linear interpolation of HRIRs that take place after convolution, it can be assumed that the same approach may be applicable. This assumption is based on the distributivity of the convolution, which is described by

$$x(n) * [f(n) + f(n)] = [x(n) * f(n)] + [x(n) * f(n)] \quad [1]$$

where x(n) is the signal and f (n) and f (n) are the (aligned) impulse responses of two emitters. The left side of the equation corresponds to the convolution with the interpolated impulse response where both impulse responses are first added together, whereas the right side of the equation corresponds to the summation of signals after the convolution with the aligned responses.

With the rendering being based on virtual loudspeakers, where object positioning is done via amplitude panning, all phantom sources are affected by the alignment.

To improve the summation of coherent signals that have been rendered with impulse responses of different emitter positions, the impulse responses of the emitter positions were time-aligned according to the difference of their TOAs. To do this it is first necessary to determine the TOAs of the emitter positions. For this, an approach has been used that detects the first significant peak in an impulse response and uses this peak for the alignment. Since the interaural time difference (ITD) of the impulse responses should not be modified, the alignment was carried out using the ipsilateral impulse responses, since this part of the HRTF contains the most energy. The contralateral side is time-shifted according to the ipsilateral side.

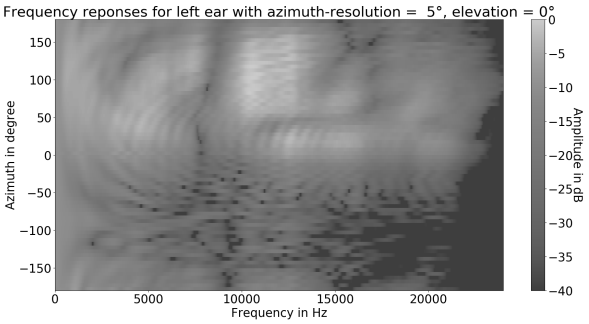

Fig 3 Frequency responses for the left ear in 5° azimuth-steps with virtual loudspeakers at every object position (no amplitude panning)

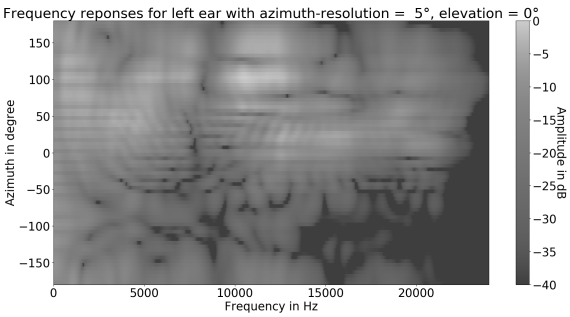

Fig 4 Frequency responses for the left ear in 5° azimuth-steps with 16 virtual loudspeakers (intermediate positions established with amplitude panning) – without alignment

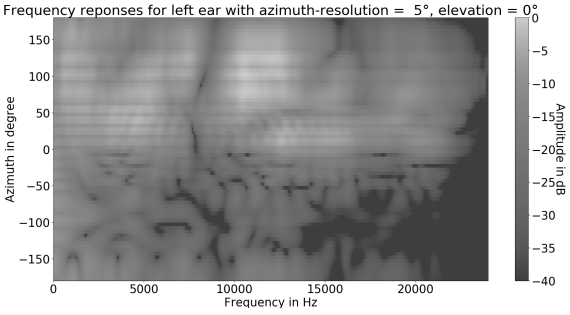

Fig 5 Frequency responses for the left ear in 5° azimuth-steps with 16 virtual loudspeakers (intermediate positions established with amplitude panning) – with alignment

Figure 3 shows the direct frequency response of HRTFs for the left ear in 5° azimuth steps. Each position corresponds to a virtual loudspeaker and therefore no amplitude panning was applied. Figure 4

in comparison shows the frequency response when using 16 virtual speakers, where all intermediate positions were realized using amplitude panning, but no alignment was used. Figure 5 in comparison to Figure 4 shows the frequency responses using a previous time alignment. It can be seen that Figure 4 shows significant drops in the frequency responses of the side facing the speakers (0 to 180°), whereas Figure 5 is much closer to the original frequency responses. The side facing away from the loudspeakers (0 to -180°), however, shows similar changes in both Figure 4 and Figure 5. This is because the ipsilateral sides of the HRTFs are used for the alignment.

This first-peak approach was as well compared to a cross correlation approach during development, whereby the frequency responses of the first-peak method showed fewer discontinuities than the cross-correlation approach though. However, since there is a wide range of possibilities to determine TOAs, it should be pointed out that this is a critical part of the alignment, which requires further research.

3.1 Evaluation

To evaluate the influence of the alignment process, a listening test has been conducted. 23 subjects participated in this listening test. Different binaural test signals were played to the subjects via headphones, which were rendered once with prior alignment and once without alignment.

All signals were rendered with a HRTF of a KU100 dummy-head from Neumann, measured at the "Technische Hochschule Köln" [12]. This measurement of the KU100 was considered to be very neutral and a dummy-head HRTF as most appropriate for multiple listeners.

The aim of the listening test was to find out whether the changes caused by the alignment had a positive effect on the overall rendering quality or not. For this purpose, the participants were asked to evaluate which of the two variants they preferred and to what extent.

The participants were advised to pay special attention to the effects on the frequency response and the localization acuity. These two parameters were considered particularly important in previous informal expert listening tests. To increase the participants' awareness of these parameters, a training session was held before the listening test.

The parameter localization acuity was deliberately used in preference to an exact position specification, since generic HRTFs were used for the experiment. Since these already have a decisive influence on the localization, a deviation from this position would likely not have resulted in a reliable statement.

Nr.	Description	Source format	Azimuth	Elevation	Median	Mean	StdDev
1	Electronic beat	Mono	158	0	1	0.476	1.680
2	Electronic beat	Mono	5	15	2	1.238	1.109
3	Electronic beat	Mono	-75	60	2	1.19	1.531
4	Audio drama	Multi-mono	-	-	0	0	0.926
5	Guitar, solo	Mono	158	0	1	1.095	1.231
6	Guitar, solo	Mono	5	15	1	0.381	1.290
7	Guitar, solo	Mono	-75	60	2	1.286	1.750
8	Jazz-ensemble	Surround (5.0)	-	-	0	0.19	0.906
9	Romantic orchestra	Stereo	-15	0	0	-0.143	0.467
10	Speech, dry, male	Mono	158	0	2	1.286	1.278
11	Speech, dry, male	Mono	5	15	1	1.333	1.168
12	Speech, dry, male	Mono	-75	60	2	1.571	1.591
13	Speech, dry, male (2)	Mono	360° movement	0	0	0.381	1.045
14	Violin, solo	Mono	158	0	-1	-0.143	1.807
15	Violin, solo	Mono	5	15	1	1.381	1.045
16	Violin, solo	Mono	-75	60	1	0	1.877

Table 1 Detailed description and results of the test signals

For the assessment, a 7-point Likert scale has been used to rate both variants in comparison.
A 49-channel system was used as a virtual loudspeaker setup, with 16 loudspeakers each, placed circularly 30° below, 30° above and at ear level, as well as one additional loudspeaker directly above the listener. This layout is based on nearly all defined loudspeakers of the ITU-R BS 2051-2 recommendation [7] and aimed to achieve a high level of compatibility, especially for all channel-based input signals.
To be able to conduct the test simultaneously for two subjects, two cabins were set up in the listening-room, separated by mobile walls. Both cabins were equipped with a laptop, an RME "Babyface Pro" audio interface and Sennheiser "HD 800" headphones. The test was carried out with the software STEP by Audio Research Labs, which is specially developed for listening tests.
To minimize disturbance, both subjects started each test at the same time after a collaborative introduction and training. In this way, the subjects would not be distracted by someone talking in the room.

3.2 Results

The average result of the test was 0.72, with a total number of 336 ratings, and a standard deviation of 1.477. Figure 6 shows a non-normal distribution for all ratings, which was further verified by a Shapiro-Wilk test [13]. This test yielded a p-value smaller than 0.05 which led to the rejection of the null hypothesis.
To investigate whether the aligned versions of the signal were rated better than the non-aligned version, a Mann-Whitney U test was conducted. This test was chosen because ratings are non-normal distributed, and the rating-scale is ordinal. For both these restrictions, the Mann-Whitney U test is applicable [14].
The result was therefore compared to a hypothetical result with the same number of ratings and a normal distribution with a mean value of 0. The null hypothesis would therefore be that the sample distributions are equal. The test yielded a p-value lower than the significance level of 0.05, and this led to a rejection of the null hypothesis.
Since the mean value of all ratings is above 0, it was assumed that overall, the aligned versions are rated significantly better than the non-aligned versions. Consequently, the aligned versions were preferred over the non-aligned versions.
When looking at the rating of the different test signals (see table 1), clear differences between the signal types were noticeable. A closer look revealed a correlation between signal complexity and rating. Especially with more complex signals or signals with several components, no or hardly any differences were noticed between the aligned and the not aligned version (see table 1, signal 4, 8, 9 and 13).
To compare the distribution of the mono-signals to the more complex signals with a higher number of channels/objects, a Mann-Whitney U test was conducted. Signals were therefore divided into two groups. The test yielded a p-value smaller than the significance level of 0.05. The null hypothesis can therefore be rejected, meaning that the sample distributions are not equal. This shows that mono-signals were rated differently than signals with more channels or objects.

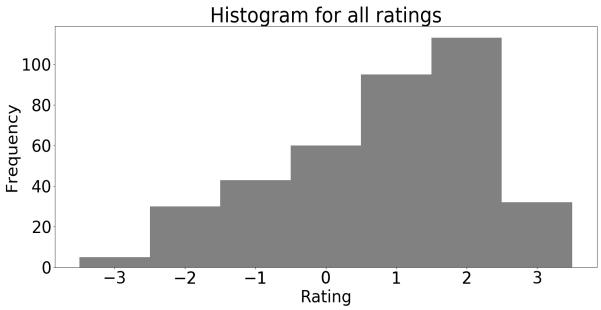

Fig 6 Distribution of all ratings

-3: The non-aligned version is much better
-2: The non-aligned version is better
-1: The non-aligned version is slightly better
0: both versions are equally
+1: The aligned version is slightly better
+2: The aligned version is better
+3: The aligned version is much better

4 Rendering of the room response

The second approach aims to reduce the computational costs for the rendering of the binaural room response. For this purpose, a smaller virtual speaker setup is used to render the room response than for rendering of the direct path. By doing so, the

number of convolutions that are required to render the room response can be reduced. As the impulse responses, that are used for rendering the room response, are significantly longer, a lot of computing capacity can be saved thereby, without reducing the spatial resolution of the direct path. Since both loudspeaker renderings reproduce the same content, but on different virtual speaker setups, only the room excitation changes. This leads to a change in the direction of the reflections. The localization/perception of these reflections can be described by the precedence-effect though [15, 16], whereby this shift is expected to be almost inaudible. A similar approach has also been described and evaluated by Picinali et al., but with Ambisonics-based rendering instead of virtual speakers [17].

4.1 Evaluation

Nr.	Description	Source format	Azimuth	Elevation
1	Guitar, solo	Mono	5	0
2	Guitar, solo	Mono	-75	15
3	Jazz-ensemble	Surround (5.0)	-	-
4	Speech, dry, female	Mono	0	0
5	Speech, dry, female	Mono	45	30
6	Percussion, dry	Mono	-47	84
7	Audio-drama	Multi-mono	-	-
8	Speech, dry, male	Mono	5	0
9	Speech, dry, male	Mono	-75	15

Table 2 Detailed description of the test signals

To evaluate the influence of using different loudspeaker renderings for the rendering of the room response, a listening test has been conducted, in which 23 subjects participated.

In preliminary expert listening panels, different speaker layouts have been chosen for the rendering of the room response. This resulted in five different speaker layouts. These layouts consist of 49, 13, 7, 5, and 2 speakers, whereby the last two systems are not three-dimensional (5.0-Surround and Stereo). All other systems are three-dimensional layouts. The comparatively large gap between the layout with 49 speakers and the layout with 13 speakers arose because the differences between larger layouts were considered to be almost inaudible.

The test was carried out as a full paired AB comparative test using a 7-point Likert-scale. Therefore, to compare five different renderings of one signal, 10 trials were required. With a total of nine signals, this resulted in 90 trials. The differences between the variants were partly considered as relatively subtle. Therefore, in order to simplify the evaluation, the AB test design was used, where in contrast to a multi-stimulus test only 2 versions have to be compared per trial. This increases the number of trials but was considered to make the rating easier.

All signals were rendered with a HRTF of a KU100 dummy-head from Neumann, measured at the "Technische Hochschule Köln" [12].

To be able to conduct the test simultaneously for two subjects, two cabins were set up in the room, separated by mobile walls. Both cabins were equipped with a laptop, an RME "Babyface Pro" audio interface and Sennheiser "HD 800" headphones. The test was carried out with the software STEP by Audio Research Labs, which is specially developed for listening tests.

To minimize disturbance, both subjects started each test at the same time after a collaborative introduction and training. In this way, the subjects would not be distracted by someone talking in the room.

4.2 Results

In this test, the subjects were asked to rate which of the both presented versions they prefer according to the parameters of spaciousness and localizability. They were given advice to pay special attention to artefacts such as comb filtering or artificial delays.

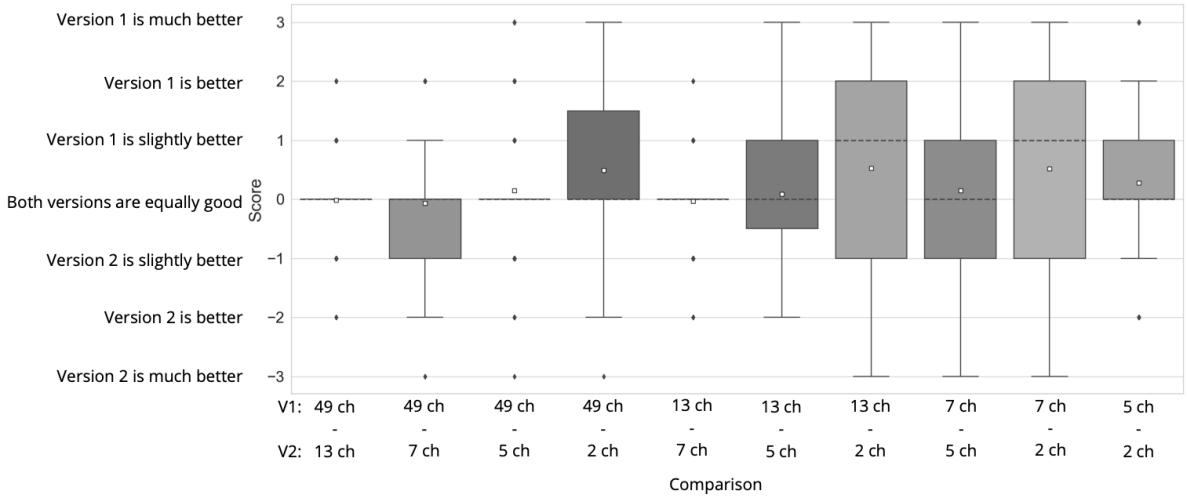

Fig 7 Comparisons of different speaker layouts for the rendering of the room response, the white square shows the mean value, whereas the dashed line shows the median

To evaluate the results, first the average results of each of the comparisons over all test signals were examined. It was found that the systems with 49, 13, 7 and even 5 loudspeakers were evaluated very similarly (see Fig 7). Only the systems with 2 loudspeakers showed clearer differences.

It was to be expected that this system would show the most obvious differences. But that the differences are so small in all other systems clearly confirms the assumption that using a smaller loudspeaker system to render the room response causes hardly any noticeable differences but allows enormous savings in computing capacity.

For further evaluation, the influence of the different test signals was examined. It was found that clearer differences were detected in relatively dry test signals than in test signals where a certain amount of reverberation was already present in the test signal. This indicates that the room component of the rendering mixes with the reverberation component of the test signal and thus becomes less differentiable. A Mann-Whitney-U test was performed to compare the distribution of both groups. The test yielded a p-value smaller than 0.05 which led to a rejection of the null hypothesis, meaning that the distribution of both groups is different.

However, in order to be able to make a more precise statement, it would make sense to carry out more specific tests, where the direct to reverberation ratio is controlled on purpose.

5 Conclusions

The intention of this research was the development and the optimization of a virtual loudspeaker based binaural rendering. This resulted in a binaural rendering addon for the EBU ADM Renderer, as well as two optimization approaches, which were generally evaluated with a listening test. On the one hand it was shown that the alignment of HRTFs can improve the rendering quality for coherent signals that are rendered with impulse responses from different emitter positions. This improvement, however, seems to be dependent on the signal-type, whereas a more complex signal reduces this improvement. On the other hand, it was found that a separate loudspeaker rendering for the rendering of the room response can reduce the number of required convolutions without creating a perceivable quality loss.

However, further investigations are needed, especially concerning the HRTF alignment, to determine whether and how the alignment affects different test signals and rendering scenarios. In particular, the exact effect on localization should be examined, for which individualized HRTFs should be used though.

Even though the optimization approaches are somehow related to the virtual loudspeaker approach, they should be applicable for other binaural rendering approaches as well.
The binaural rendering addon, including the implementation of the prior described optimization approaches can be found at https://github.com/IRT-Open-Source/binaural_nga_renderer.

References

[1] European Broadcast Union, *Next Generation Audio. Optimized audio regardless of where and how it is consumed,* https://tech.ebu.ch/nga.

[2] P. Zahorik, F. Wightman, and D. Kistler, "On the discriminability of virtual and real sound sources," in *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Acoustics:* IEEE, Oct, 1995, pp. 76–79.

[3] E. H. A. Langendijk and A. W. Bronkhorst, "Fidelity of three-dimensional-sound reproduction using a virtual auditory display," *The Journal of the Acoustical Society of America*, vol. 107, pp. 528–537, 2000.

[4] C. Cieciura, R. Mason, P. Coleman, and M. Paradis, Eds., *Survey of Media Device Ownership, Media Service Usage, and Group Media Consumption in UK Households,* 2018.

[5] European Broadcast Union, "TECH 3388. ADM Renderer For Use In Next Generation Audio Broadcasting,", 2018.

[6] International Telecommunication Union, "Recommendation ITU-R BS.2076-1. Audio Definition Model,", pp. 2–3, 2017.

[7] International Telecommunication Union, "Recommendation ITU-R BS.2051-2. Advanced sound system for programme production,", 2018.

[8] S. Goossens and R. Stumpner, "Auralisation of room acoustics - A tool for planning broadcast production rooms?,", 2004.

[9] S. Goossens, "Simulation und Auralisierung kleiner Räume,", 2010.

[10] F. Völk, *Externalization in data-based binaural synthesis: effects of impulse response length,* 2009.

[11] M. Matsumoto, S. Yamanaka, M. Tohyama, and H. Nomura, "Effect of arrival time correction on the accuracy of binaural impulse response - Interpolation interpolation methods of binaural response," *Journal of the Audio Engineering Society. Audio Engineering Society*, vol. 52, 2004.

[12] B. Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU100,", 2013.

[13] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, pp. 591–611, 1965.

[14] Jason Brownlee, *How to Calculate Nonparametric Statistical Hypothesis Tests in Python,* https://machinelearningmastery.com/nonparametric-statistical-significance-tests-in-python/, 2018.

[15] H. Wallach, E. B. Newman, and M. R. Rosenzweig, "The Precedence Effect in Sound Localization," *The American Journal of Psychology*, vol. 62, p. 315, 1949.

[16] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *The Journal of the Acoustical Society of America*, vol. 106, pp. 1633–1654, 1999.

[17] L. Picinali, A. Wallin, Y. Levtov, and D. Poirier-Quinot, *Comparative perceptual evaluation between different methods for implementing reverberation in a binaural context,* 2017.