PERCEPTUAL EVALUATION OF BINAURAL AURALIZATION OF DATA OBTAINED FROM THE SPATIAL DECOMPOSITION METHOD

Jens Ahrens

Audio Technology Group, Division of Applied Acoustics
Chalmers University of Technology, 412 96 Gothenburg, Sweden
jens.ahrens@chalmers.se

ABSTRACT

We present a perceptual evaluation of head-tracked binaural renderings of room impulse response data that were obtained from the spatial decomposition method. These data comprise an omnidirectional impulse response as well as instantaneous propagation directions of the sound field. The subjects in our experiment compared auralizations of these data according to the originally proposed method against direct auralizations of dummy head measurements of the exact same scenarios. We tested various parameters such as size of the microphone array, number of microphones, and HRTF grid-resolution. Our study shows that most parameter sets lead to a perception that is very similar to the dummy head data particlarly with respect to spaciousness. The remaining differences that are audible are considered small and relate primarily to timbre. This suggests that the equalization procedure that is part of the approach provides potential for improvement. Our results also show that the elevation of the propagation directions may be quantized coarsely without audible impairment.

***Index Terms*—** Reverberation, spatial decomposition method, binaural rendering, head-related transfer functions

1. INTRODUCTION

The spatial decomposition method (SDM) [1] uses room impulse responses that were captured with a typically compact microphone array to obtain the omnidirectional pressure signal as well as the instantaneous arrival direction for each of the digital samples that the pressure signal is composed of. The geometry of the array is flexible as long as the required pressure signal as well as the instantaneous arrival directions can be obtained from the data. SDM has been primarily used for analysis and visualisation of the directional properties of room impulse responses and is particularly popular in concert hall acoustics [2].

SDM has also successfully been applied in auralization of acoustic spaces. Expressed slightly simplified, SDM data are auralized by distributing the digital samples of the pressure signal over the available loudspeakers such that the instantaneous arrival directions of the signal are maintained as precisely as possible. Example works are [3, 4, 5], all of which used loudspeaker systems in the auralization. The systems presented in [3, 5], for example, play the obtained signals directly from the available loudspeakers while [4] uses Ambisonics encoding of the components. Binaural auralization of SDM data is essentially similar to loudspeaker rendering as the available head-related transfer function (HRTF) measurement points are used as virtual loudspeakers [6]. The optimal data-dependent placement of a moderate number of virtual loudspeakers is investigated in [7].

An alternative approach to binauralization of omnidirectional room impulse responses is presented and evaluated in [8, 9, 10], which uses a more involved decomposition of the room impulse response into direct sound, early reflections, and late reverberation.

All works that performed SDM-based auralization report success in terms of the pleasantness of the results, but it has been largely unclear how authentic the auralization is, i.e., how similar the auralization sounds to the original room. Some amount of perceptual evaluation was presented in [6, 11] but with no comparison to a reference. We showed in [12] that the spatial data can be replaced with synthetic data with hardly causing any perceptual difference. Only [13] performed a comparison of a modified version of binaural SDM-auralization to a dummy-head-based auralization of the same scenario. The perceptual attributes that were investigated were source width, diffuseness, and source distance and were rated as almost identical. The overall perceptual distance was not investigated.

The present study aims at filling this gap. We chose binaural rendering because it constitutes the most controlled and reproducible scenario. We apply head tracking in the rendering to mitigate risks for impairment of the spatial fidelity of the rendering [14]. As in [13], we compare the binaural auralizations to dummy head auralizations of the same scenario so that their authenticity can be evaluated. This strategy has also been successfully applied with spherical microphone array data, for example, in [15, 16]. Contrary to previous studies, we incorporate a variety of microphone array geometries.

2. SPATIAL DECOMPOSITION AND AURALIZATION

SDM estimates the direction of arrival (DOA) of the sound pressure signal of a room impulse response in short time windows along the entire impulse response. These data are obtained from compact microphone arrays the geometry of which is not important as long as the desired information can be deduced. The microphone arrays do typically not comprise a scattering object.

The room impulse response is analyzed in segments. The time-difference of arrival is determined for each time window for each of the microphone pairs in the array through cross-correlation. Subsequently, a minimum mean square error problem is solved to obtain the final estimate of the nominal DOA for the time window under consideration [1]. The analysis window advances in steps of 1 sample so that one DOA estimate is obtained for each digital sample of the impulse response.

Fig. 1 exemplary depicts the first 20 ms of the data of the room Hall. The data where obtained from a star-shaped array of radius 50 mm comprising six omnidirectional microphones (cf. Fig. 2).

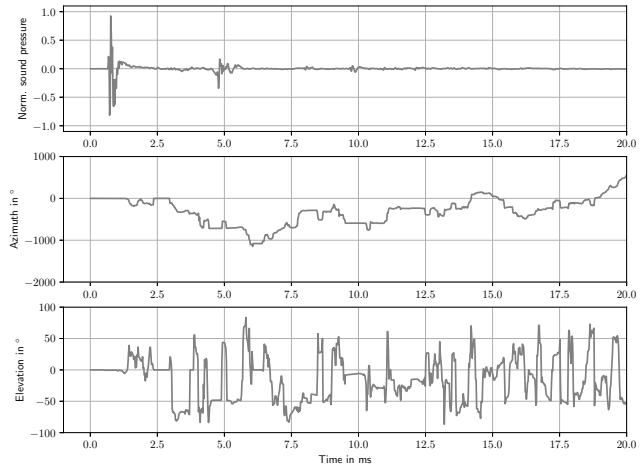

Figure 1: Sound pressure signal (top) of room Hall as well as unwrapped azimuth (middle) and elevation (bottom) as determined by SDM

The direct sound occurring between approx. 0.5 and 1.5 ms is apparent, and a few reflections might be discernible at approx. 5 ms and 10 ms. The direct sound and the mentioned reflections are also apparent in the azimuth: The determined azimuth appears to be fairly stable for the duration of the related event. The direct sound is also apparent in the computed elevation (approx. 0°), but the arrival elevations of the reflections are not always as obvious.

It has been unclear which array geometry and which number of microphones are most favorable particularly when it comes to binaural rendering. The authors of the original method [1] achieved good results with a small 6-element star-shaped array similar to ours but smaller.

The pressure signal can be obtained in different ways depending on the microphone array. If the array comprises omnidirectional microphones and the array is compact, the signal of any of the microphones may be employed directly. Arrays that do not employ omnidirectional microphones such as tetradedral Ambisonics microphones [17] require different dedicated solutions.

Auralization of the obtained pressure and DOA signals is performed by either distributing the signal samples over the available loudspeakers via nearest neighbor interpolation [3], vector-base amplitude panning [1], Ambisonics encoding [4], and the like, with compensation for potentially varying loudspeaker distance from the listening location applied. The loudspeaker signals finally need to be equalized in frequency bands and time windows to achieve the correct signal spectrum at the listening location [3]. The resulting impulse responses for each of the loudspeakers are then convolved with the source signal.

Head-related transfer functions (HRTFs) may be interpreted as virtual loudspeakers so that the same procedure can be applied to auralize the data binaurally [6].

3. DATA

We acquired impulse responses from the rooms Lab (reverb decay time: 0.14 s) and Hall (reverb decay time: 0.9 s) to the ears of a KEMAR dummy head (DH) as well as to different microphone arrays. The arrays as well as as the DH were located at the exact same positions as verified with a laser positioning device. The DH

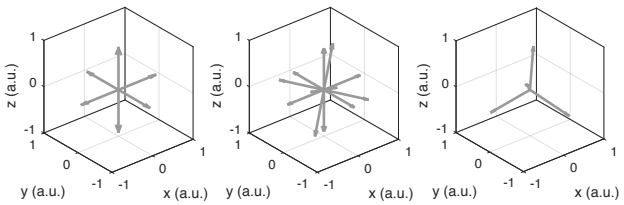

Figure 2: Geometry of the employed arrays: 6 elements – star-shaped (left), 12 elements – star-shaped (middle), and tetrahedron (right); the 12-element array was obtained by using the 6-element array and adding a rotated copy of it

Array type	Radius (mm)	Identifier
6-element star-shaped	20	SS_6^{20}
6-element star-shaped	50	SS_6^{50}
6-element star-shaped	100	SS_6^{100}
12-element star-shaped	50	SS_{12}^{50}
4-element B-format	24	BF_4^{24}
4-element tetrahedron	50	TH_5^{50}

Table 1: List of microphone arrays employed; the B-format microphone was a Tetramic by Core Audio (the B-format signals were extracted using the standard method [17]); all other arrays employed omnidirectional microphones; refer to Fig. 2 for an illustration of the geometries

data were acquired for different head orientations in steps of 1° to enable head-tracked auralization. The distances between the sound sources and the receivers were approx. 3 m and 5 m, respectively, so that the response was dominated by the direct sound.

SDM may not be considered a physically accurate decomposition of the room response as only one propagation direction is obtained for each time window in which one has to assume that many wave fronts are generally contained. SDM is rather an approximation. As to our awareness, the effect of geometry of the arrays not been investigated in detail. We therefore employed microphone arrays with 4, 6, and 12 microphones as depicted in Fig. 2. The 6-element array was available with different radii as listed in Tab. 1. The arrays were 3D-printed, and the impulse responses where measured sequentially (cf. Fig. 3). The sizes of the arrays range from 20 mm radius (measured from the center of the array to the center of the microphone diphragm) to 100 mm. This corresponds to 2 sensors per wavelength in radial direction at approx. 4.2 kHz and 0.86 kHz, respectively.

All microphone arrays other than the B-format one (BF_4^{24}) yield consistent and plausible spatial data for at least the direct sound and the strong reflection around 5 ms as evident from Fig. 4. The spatial data for the direct sound and the strong reflection are marked by black arrows. The data from BF_4^{24} are more erratic and are plotted with transparency in order not to obstruct the view. The data that we yielded for all other time instants were very different for the different microphone array sizes and geometries. One cause for the differences can be the circumstance that we used longer analysis windows to perform the required cross-correlation for the larger arrays as the microphones were farther apart. This circumstance needs further investigation, and we are not able to interpret the differences that are apparent.

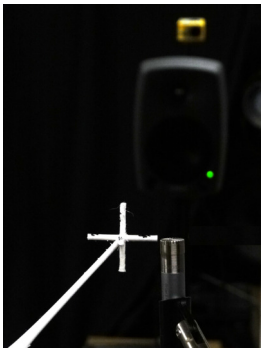

Figure 3: Photograph of the measurement procedure; the array SS_6^{50} and the B&K measurement microphone are shown in the foreground and the loudspeaker and the laser device in the background

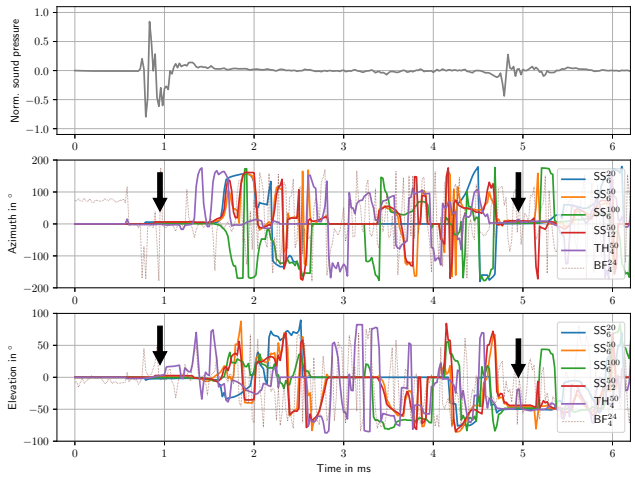

Figure 4: SDM data for the room Hall (Fig. 1) obtained from the different microphone arrays from Tab. 1

For the binaural SDM-auralization, we have available a dense anechoic HRTF grid of the DH with which above described data have been acquired. The set covers the whole sphere and both azimuth and elevation are sampled with steps of 1°. This results in more than 65.000 measurement points. It is intuitive that a subset of the available measurement points exists that leads to an SDM-auralization that is indistinguishable from the complete set. To obtain information on this, we employed HRTF subsets in the auralization that contain only certain elevation angles. We used sets with the elevation quantized in steps of 10°, 45°, 90°, as well as one set with only measurement locations inside the horizontal plane. We denote the sets H_{10}, H_{45}, H_{90}, and H_{hor}, respectively. All subsets included the horizontal plane. Note that the 90°-set contains only the horizontal plane as well as the two poles.

The standard SDM according to [1, 6] with the following modifications was employed for the auralization: We discarded any special treatment of the low-frequency region, and we modified the equalization procedure to reduce the amount of time aliasing that it produces. The spatial data, i.e., the instantaneous azimuth and elevation of the arrival direction of each sample of the pressure impulse response, was computed for all combinations of rooms and array type at a sampling frequency of 48 kHz.

4. EXPERIMENT

It was shown in [13] without head tracking that spaciousness as represented by perceived source width, perceived distance, and perceived diffusity is very well preserved when performing SDM-auralization. We can confirm this through informal listening based on our data. A pre-study showed that audible differences between the SDM-auralization and the DH auralization with respect to both spaciousness and timbre can occur, but it is difficult to verbalize them because they are small and do not precisely relate to any single one of the commonly considered attributes. We therefore conducted an experiment in which the subjects rated the overall perceptual difference between a reference – for example the DH auralization – and a set of stimuli – for example binaural SDM-auralization with a range of parameters.

All stimuli were presented with head tracking. We used rock drums and male speech as signals. The signals were looped continuously, and switching between the stimuli and/or the reference occurred without interruption or other artifacts. The presentation of the stimuli was performed using the software SoundScape Renderer (SSR)[1] [18] running in binaural room synthesis mode. SSR convolves a given input signal with that pair of impulse responses that corresponds to the instantaneous head orientation. We employed a Polhemus Patriot head tracker and AKG K702 open-design headphones. The experiment was conducted in an acoustically treated laboratory room. We refer the reader for further details in the setup to [16] where we used an identical one.

The setup is inspired by MUSHRA [19] in that sets of stimuli were compared against a reference, and a hidden reference as well as an anchor were always present. The anchor was identical to the reference but lowpassed with a 4-th order butterworth filter with a critical frequency of 3000 Hz.

Two different sets of stimuli were presented:

1) A stimulus set with the DH as reference: The motivation was to obtain information on which of the parameter sets related to the capture side produces the result that is least different from the DH, which is considered the baseline. See Fig. 5 for the complete set of stimuli.
2) A stimulus set with primarily data from one microphone array but with different quantizations of the HRTFs. The densest HRTF set was the reference, and the motivation was to obtain information on how much the rendering side can be simplified without loosing authenticity. See Fig. 6 for the complete set of stimuli.

Each stimulus set occurred twice in the experiment. This resulted in 16 sets of stimuli to be rated (2 sets x 2 rooms x 2 audio signals x 2). The order of the sets and the order of the stimuli inside a set were randomized whereby all sets for one audio signal were completed before the audio signal was changed. The experiment was preceded by written instructions and 2 manually picked sets of stimuli as training.

The subjects reported the perceived different by means of continuous sliders with a scale ranging from “No difference” via “Small difference”, “Moderate difference”, and “Significant difference” to “Huge difference” (cf. the ordinate in Fig. 5 and 6).

[1] http://spatialaudio.net/ssr/

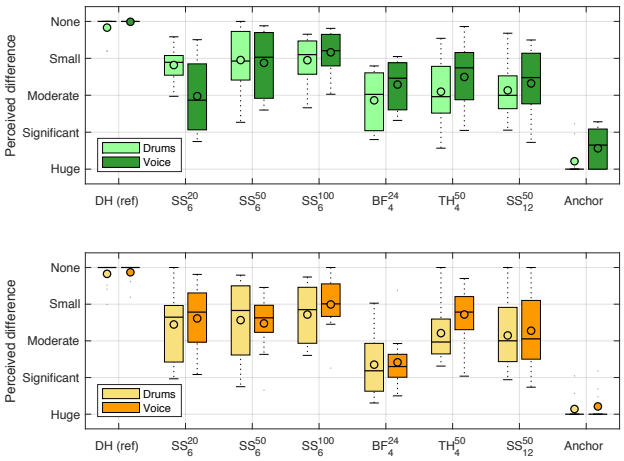

Figure 5: Results with the DH as reference for Lab (green, top) and Hall (orange, bottom); the rendering of the array data always used the HRTF set H_{10}

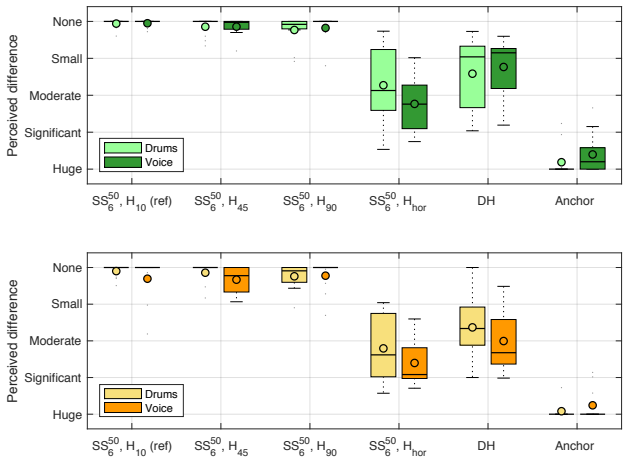

Figure 6: Results with array SS_6^{50} with HRTFs H_{10} as reference for Lab (green, top) and Hall (orange, bottom)

5. RESULTS

11 grown-up subjects of different genders and with self-reported normal hearing participated. The mean duration of the experiment was 26 min. The results are depicted as boxplots in Fig. 5 and 6. The median values of the data are shown via the horizontal line, the mean via the black circle, the 25th and 75th percentiles via the box, the whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually as dots.

We performed an n-way ANOVA [20] with the subject index as additional random factor to identify conditions with statistically different means of the subjects' ratings. Only the statistical analysis for the data from Fig. 5 is reported in detail as the situation is obvious for the data from Fig. 6. We also excluded the ratings for the anchor and the hidden reference to mitigate some of the concerns raised in [21]. The main observations are:

- The hidden reference was rated consistently.
- The audio signal did not have a significant influence ($F(1, 510) = 2.3, p > 0.13$).
- The perceived difference to the reference was slightly smaller for the room Lab compared to the more reverberant room Hall (Fig. 5 top vs. bottom; $F(1, 510) = 4.9, p \leq 0.027$).
- The quantization of the elevation has no audible effect as long as some information is presented outside of the horizontal plane (Fig. 6). Note that the HRTF set H_{90} comprises only HRTFs in the horizontal plane and at the two poles. Note that Lab exhibited a very absorptive ceiling, and Hall is 10 m high.
- If only horizontal HRTFs are used, then a clear difference to the conditions that include non-horizontal signals is perceived (Fig. 6).
- The array geometry is significant in Fig. 5 ($F(5, 510) = 20.9$, $p < 0.001$). The 6-element star-shaped array particularly for the radii of 50 mm and 100 mm tends to produce the smallest perceptual differences to the DH auralization (Fig. 5), which are on average in the order of "small".
- The B-format signals tend to cause a larger perceived difference than most of the other array geometries (Fig. 5).

The circumstance that the differences that are perceived for different array geometries are somewhat comparable suggests that these are all perceived as rather similar. We can confirm this through informal listening. This also confirms the results from [12], where we showed that SDM with synthetic spatial data can sound very similar to the original data. Recall from Sec. 3 that the different microphone arrays can yield very different spatial data for the same scenario. It seems that all geometries that we employed other than the B-format-based array yield spatial data that are plausible when auralized binaurally. However, this needs to be confirmed formally.

We interviewed all subjects after the experiment. The main statements that we distilled manually from the responses are the following:

- Many of the stimuli were hardly distinguishable from the reference.
- The remaining differences were primarily related to the timbre.
- In those cases where considerable differences were perceived, differences with respect to both timbre and spaciousness occurred.

6. CONCLUSIONS

The basic SDM-based binaural auralization produces pleasant results for a variety of microphone array geometries, whereby a 6-element star-shaped array tends to produce data that sounds most similar to a dummy head auralization of the same scenario. The authenticity is slightly higher for rooms with short reverberation compared to rooms with long reverberation and is only slightly lower than for binaural rendering of the data from large spherical microphone arrays [16]. The involved equalization procedure seems to exhibit potential for improvement as the remaining differences are reported to be primarily timbre related.

7. ACKNOWLEDGMENTS

We thank Jan-Gerrit Richter, Hark Braren, and Janina Fels of RWTH Aachen University for providing us with the KEMAR HRTFs.

8. REFERENCES

[1] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *J. Audio Eng. Soc*, vol. 61, no. 1/2, pp. 17–28, 2013.

[2] J. Pätynen, S. Tervo, and T. Lokki, "Analysis of concert hall acoustics via visualizations of time-frequency and spatiotemporal responses," *The Journal of the Acoustical Society of America*, vol. 133, no. 2, pp. 842–857, 2013.

[3] S. Tervo, J. Pätynen, N. Kaplanis, M. Lydolf, S. Bech, and T. Lokki, "Spatial Analysis and Synthesis of Car Audio System and Car-Cabin Acoustics with a Compact Microphone Array," *Journal of the Audio Engineering Society*, vol. 63, no. 11, pp. 914–925, 2015.

[4] M. Frank and F. Zotter, "Spatial impression and directional resolution in the reproduction of reverberation," in *Proc. of DAGA*, Aachen, Germany, 2016, pp. 1304–1307.

[5] S. V. Amengual Garí, W. Lachenmayr, and E. Mommertz, "Spatial analysis and auralization of room acoustics using a tetrahedral microphone," *The Journal of the Acoustical Society of America*, vol. 141, no. 4, pp. EL369–EL374, 2017.

[6] J. Pätynen, S. Tervo, and T. Lokki, "Amplitude panning decreases spectral brightness with concert hall auralizations," in *55th International Conference of the AES*, Helsinki, Finland, Aug. 2014.

[7] O. Puomio, J. Pätynen, and T. Lokki, "Optimization of virtual loudspeakers for spatial room acoustics reproduction with headphones," *Appl. Sci.*, vol. 7, no. 12, 2017.

[8] C. Pörschmann, P. Stade, and J. Arend, "Binauralization of Omnidirectional Room Impulse Responses - Algorithm and Technical Evaluation," *Proceedings of the DAFx 2017*, pp. 345–352, 2017.

[9] P. Stade, J. Arend, and C. Pörschman, "A parametric model for the synthesis of binaural room impulse responses," in *Proc. Mtgs. Acoust. 30*, Boston, MA, USA, June 2017, p. 015006.

[10] P. Stade, "Perzeptiv motivierte, parametrische Synthese binauraler Raumimpulsantworten [text in German]," Dissertation, Technische Universität Berlin / TH Köln, 2018.

[11] N. Kaplanis, S. Bech, S. Tervo, J. Pätynen, T. Lokki, T. van Waterschoot, and S. H. Jensen, "A method for perceptual assessment of automotive audio systems and cabin acoustics," in *60th International Conference of the AES*, Leuven, Belgium, Feb. 2016.

[12] J. Ahrens, "Auralization of omnidirectional room impulse responses based on the spatial decomposition method and synthetic spatial data," in *IEEE ICASSP*, Brighton, UK, May 2019, pp. 146–150.

[13] M. Zaunschirm, M. Frank, and F. Zotter, "BRIR synthesis using first-order microphone arrays," in *144th Convention of the AES*, 2018.

[14] D. R. Begault, A. S. Lee, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," in *108th Convention of the AES*, May 2000.

[15] B. Bernschütz, "Microphone arrays and sound field decomposition for dynamic binaural recording," PhD thesis, Technische Universität Berlin, 2016.

[16] J. Ahrens and C. Andersson, "Perceptual Evaluation of Headphone Auralization of Rooms Captured with Spherical Microphone Arrays with Respect to Spaciousness and Timbre," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2783–2794, Apr. 2019.

[17] J.-M. Batke, "The B-Format Microphone Revisited," in *Proceedings of the Ambisonics Symposium*, Graz, Austria, June 2009.

[18] M. Geier, S. Spors, and J. Ahrens, "The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods," in *124th Convention of the AES*, May 2008.

[19] ITU-R, "BS.1534: Method for the subjective assessment of intermediate quality levels of coding systems," Recommendation, Internation Telecmmunications Union, 2015.

[20] J. Bortz, *Statistik* [*text in German*], 6th ed. Berlin/Heidelberg: Springer, 2006.

[21] C. Mendonça and S. Delikaris-Manias, "Statistical tests with MUSHRA data," in *144th Convention of the AES*, Milan, Italy, May 2018, p. 10006.