Audio Engineering Society

Conference Paper

Presented at the 2022 International Conference on
Audio for Virtual and Augmented Reality
2022 August 15–17, Redmond, WA, USA

Methodology for perceptual evaluation of plausibility with self-translation of the listener

Marta Gospodarek[1,2], Olivier Warusfel[2], Pablo Ripollés[1,3,4], and Agnieszka Roginska[1]

[1]*Music and Audio Research Laboratory, New York University*
[2]*Sciences et Technologies de la Musique et du Son , IRCAM, CNRS, Sorbonne Université, Ministère de la Culture*
[3]*Department of Psychology, New York University*
[4]*Center for Language, Music and Emotion, New York University*

Correspondence should be addressed to Marta Gospodarek (gospodarek@nyu.edu)

ABSTRACT

The perceptual evaluation of the Audio Augmented Reality (AAR) experience is typically conducted using authenticity or plausibility as a measure of realism of the sounds. The previous studies usually employed methodologies where participants were comparing the exact reference with real sound. This paper proposes a novel experimental design where participants rate the plausibility of a pair of real or virtual loudspeakers playing consecutively from different positions in the room. This approach sets expectations of the user very close to real-life scenarios where similar but not identical sources exist within one environment. The objective of the study is to assess plausibility and to evaluate spatial sound attributes using two auralization methods which simulate the experimental room. Instead of using the *yes/no* paradigm, the proposed methodology employs continuous scales which allow gaining insight into the correlation of plausibility with other sound attributes. The preliminary results show that the proposed experimental design was successful in obtaining a meaningful comparison between the two auralizations and the real source. Moreover, the analysis of results suggests that plausibility is a continuous dimension.

1 Introduction

The Augmented Reality (AR) audio experience aims to create a perfect illusion of the virtual sources being part of the real environment. Quality evaluation of these experiences can be approached using two different paradigms. One is aimed to quantify the subject performance during a given task, and from it conclude the quality of the rendering system. Its advantage is to provide an objective evaluation. Another approach is to evaluate realism, which is subjective and challenging to measure. Evaluating the realism of an AR experience is a multimodal task, as many factors contribute to it such as the quality of the rendering system, task of the user, modes of interaction, or the user's personality. Thus, it is necessary to precisely define the measure of the quality being assessed.

One of the essential measures of the AR experience quality is authenticity, which is commonly defined as the perceptual identity of real and virtual events. In previous studies, the authenticity of binaural rendering was measured with a *yes/no* paradigm where partici-

pants were listening to a real loudspeaker and its virtual replication and had to respond if the sound was real or not [1, 2]. The authenticity sets a very technically challenging goal which might not be feasible to achieve in most rendering systems. Even very subtle differences, like JND of timbre between the simulation and reference sound (caused e.g. by headphones repositioning) can contribute to the identification of the simulation. Moreover, in most real life situations, the immediate exact same reference to the virtual source is not available. Plausibility, which relates to the similarity to the internal expectation and not the actual reference, seems to be a more appropriate measure of audio experience quality. Lindau suggests a definition of plausibility which applies to AR: "a simulation in agreement with the listener's expectation towards a corresponding real event" [3]. Plausibility can be evaluated without real reference [4, 5] but in this case, it is more prone to individual biases caused by experience, training, etc. [6]. In previous studies, including a real reference in the test design, resulted in a change of plausibility judgment [7]. This suggests that the type of reference presented to the listener might affect the perception of plausibility. Wirler et al. proposed another term - transfer-plausibility - in which the immediate but not exact reference is introduced for direct comparison. The experimental design proposed in this study includes several loudspeakers playing simultaneously from which participants have to pick the simulation. This approach is more similar to real applications but applies mostly to situations where multiple sources play at the same time.

1.1 Overall Approach

This paper proposes a novel experimental design where participants rate the plausibility of a pair of speakers playing consecutively from different positions in the room. This approach sets expectations of the user very close to the real-life scenario where similar but not identical sources exist within one environment. In a real scenario of the AAR system, an immediate comparison with the real source is available, whereas the reference is never identical with the virtualized sound. During the experiment, subjects walk back and forth following a predefined path which allows to evaluate the dynamic rendering of the stimuli. The use of special transparent headphones (AKG K1000) allows for direct comparison between the simulation played on headphones and the real sources played through loudspeakers positioned in the room. The questionnaire of the study not only investigates the plausibility perception but also includes several perceptual attributes taken from the Spatial Audio Quality Inventory (SAQI) [9]. The inclusion of these attributes is motivated to gain more insight into possible correlations between plausibility and other sound attributes. In order to enable such a study, the plausibility is evaluated on a continuous scale instead of the *yes/no* paradigm often proposed in the literature. Thus, here we consider plausibility as a continuous dimension as using the *yes/no* paradigm might unnecessarily limit the evaluation and richness of the data. Such as situation could occur for instance for a virtual sound source which can be distinguished from a real reference while at the same time being perceived as plausible. In addition, real sound sources could even sound not very plausible when heard from unusual situations, for instance, if the direct path is occluded. For these reasons, the proposed methodology includes not only the evaluation of pairs where a virtual source is contrasted with a real one, but also pairs of two real or two virtual sources. Analysis of the results should give insight into the influence of the comparison with a real source on the plausibility judgment.

Another aim of the study is to validate the proposed experimental design by comparing two different approaches to auralization of early reflections and late reverberation. The first method exploits a simple 3D numerical model of the room and runs a real-time beam-tracing method to calculate the time and spatial distribution of the reflections. The late reverberation is modeled with the Feedback Delay Network (FDN). The second auralization method characterizes the room with a single spatial room impulse response (SRIR), which is further manipulated in order to account for the relative listener-source distance. Both methods employ the same simulation of the direct sound propagation effect with a simple directivity model to simulate the source radiation pattern. The two methods have different limitations and were chosen to investigate the possible influence of different aspects of simulation on the plausibility judgment. With the first approach, the temporal structure of reflections should be closer to reality as it is constantly updated according to the source and listener positions in the room. In contrast, the late reverberation is only an approximation of the measured IR, as it is rendered through an FDN implementation. On the other hand, the SRIR auralization can accurately reproduce the time and frequency distribution of the

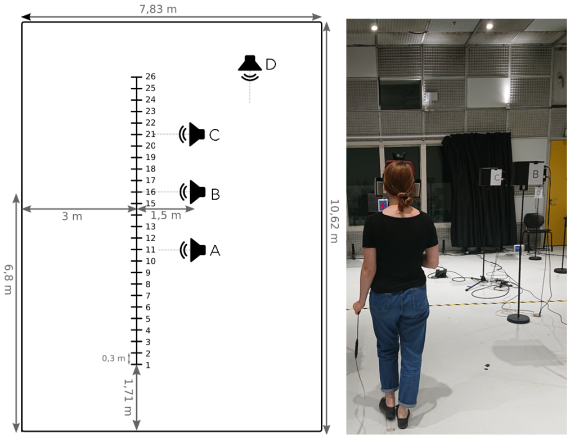

Fig. 1: Loudspeakers positions in the experimental room and participant wearing AKG K1000 headphones and HTC Vive tracker

late reverberation. However, it cannot account for the varying early reflections patterns along the walking path. The temporal structure of reflections is kept from the original SRIR and does not depend on the position of the listener and source, as it would in reality. Previous studies indicated that depending on the type of variation in the temporal structure of early reflection and position in the room, the accurate reproduction of temporal pattern might not be necessary [10].

2 Virtual acoustics environment implementation

2.1 Room and loudspeaker measurements

The experimental room is of size 10.68 m x 7.83 m x 4.17 m and cubature 348.71 m^3 as shown in Figure 1. The walls and ceiling are covered with a random arrangement of absorptive and semi-reflective panels. The floor is covered with linoleum on concrete. One wall has a big glass window covered with a thick curtain. During the experiments, the room was nearly empty. The measured reverberation time at 1 kHz was 0.28 seconds with a slight global decrease according to frequency (see Figure 3).
In order to conduct an objective comparison with the auralizations (see section 2.6), acoustic measurements were performed using sine-sweep signals for each of the four loudspeakers used in the experiment and for 26 receiver positions evenly distributed (30 cm spacing) along the walking path. The measured path is a straight line, 7.5m long and parallel to the main room axis, next to three of the speakers at a minimum distance of 1.5 m (see Figure 1). Each point was measured using the EM32 Eigenmike® (4th order 3D soundfield) spherical microphone array from *MH Acoustics* and a Neumann KU 100 dummy head. After the measurements, each of the impulse responses was denoised to allow for proper analysis and avoid any artifacts [11].

The four loudspeakers used for the experiment were dual concentric Amadeus PMX 5. One of them was measured in an anechoic chamber at IRCAM to characterize its frequency response and directivity. The microphone was set up 1m from the speaker. Measurements were taken every 15° around the speaker which resulted in a total of 24 measurement points. Thanks to the axisymmetrical design of the loudspeakers, the measurements were done only on the horizontal plane.

2.2 Direct sound and directivity modeling

The rendering of the direct sound for both auralizations was the same. At first, the propagation delay was added to the initial stimulus based on the distance between the listener and the source. After that, the signal was filtered with the on-axis spectral response of the experimental loudspeaker, then the source directivity model described below was applied. Next, the intensity attenuation according to distance was applied following the conventional inverse square law. Finally, the signal was convolved with a proper Head-Related Transfer Function (HRTF) controlled by the rotation of the listener's head and the relative position of the source.

The directivity of the source is synthesized through the implementation of directional filters. The directivity modeling is based on beamforming up to 4th order HOA which allows to approach a given radiation pattern [12]. For simple radiation patterns such as the one of concentric loudspeakers, the simulation is approximated with a single spatial dirac distribution bandlimited with the order of the HOA decomposition: the higher the order the narrower the beam. To ensure the best possible match of the model for each frequency band, the simulation was verified by comparing the fit of the curves from the measurements and simulation focusing on the range of angles from which the listener would hear the speakers (-70° 70°) along the path. The operation was repeated in eight frequency bands [13]. The resulting directivity index for simulation and measurements are compared on Figure 2.

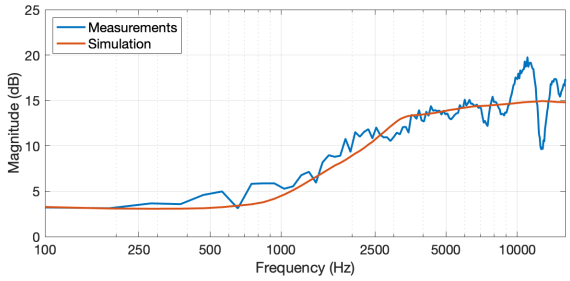

Fig. 2: Directivity index of loudspeaker PMX5

2.3 Auralization based on GA simulation

The first auralization method, labeled GA in the following, is based on geometrical acoustic modeling of the room. The method combines a real-time beam-tracing algorithm for the simulation of early reflections, and FDN for the rendering of late reverberation. The expected advantage of the method is that it calculates the early reflections segment of the RIR based on the actual geometry of the room and according to the instantaneous position and orientation of the source and listener in the room - thus potentially giving more accurate space-time distribution of early reflections. On the other hand, the late reverberation is an approximation of the actual RIR decay based on the reverberation time estimated in a limited number of frequency bands.

A simplified 3D model of the experimental room was designed and input to the modeler. The model was then calibrated to obtain the same frequency dependent reverberation time as BRIRs measured with the KU 100 and averaged over six positions (1 to 6, see Figure 1). The floor material (linoleum on concrete) was assigned an absorption value based on the literature [14]. As the absorption coefficients for wall and ceiling materials were not known, they were estimated from the measured reverberation time, using Eyring's formula.

The rendering system was implemented using EVERTims module of Spat5 library running in the Max/MSP software [15]. EVERTims is an open-source framework for 3D models auralization [16]. The modeler unit constructs a beam tree for the current scene geometry as well as the positions of the listener and the source. The beam tree is a base to generate a list of image sources which are then sent to the auralization object. The modeler characterizes each reflection path by its direction of arrival, propagation delay, filtering due to the source directivity and frequency-dependent material properties, and air absorption. Directivity of the source in the modeler is applied to both the direct sound and image-sources according to the model described in section 2.2. Image sources up to the 3rd order, and limited to reflections earlier than 100ms were implemented in the system. Each of them was encoded into a 4th order HOA soundfield according to its direction of incidence. All together, they form a single ambisonic stream representing the early reflections segment of the RIR.

Synthesizing the late reverberation through image sources modeling would however not be efficient since the computation cost increases exponentially with the reflection order [17]. The late reverberation was simulated with an 8-channel FDN, which parameters (decay rate and modal density) were set to match the BRIRs measured in the room (see Figure 3). The incoming signal feeding the FDN was equalized according to the power spectrum radiated by the loudspeaker. FDN are characterized by a slow building up of first reflections before reaching a high density reverberation. Hence, an anti-phase filter was used to cancel out this building up process until 80ms (i.e., with a small overlap with the latest image source reflections). This guarantees that only the first reflections provided by the image source model are delivered to the listener [18]. The transition time of 80ms between the image source reflections and the FDN late reverberation was chosen to match the mixing time observed on the measured SRIRs (i.e. the time when a sufficiently high echo density is achieved). The mixing time was estimated from the analysis of the spatial coherence of the measured SRIRs [11]. The eight FDN output channels were encoded to 4th order HOA with diffuse panning. Both the first reflection ambisonic stream and the late reverberation stream were mixed before being sent to the binaural decoder.

2.4 Auralization based on SRIR synthesis

The second auralization method, labeled SRIR in the following, is based on a convolution approach [19] using a single reference SRIR among the measurements described in section 2.1. The reference SRIR corresponds to loudspeaker B measured with the microphone set at position M1 (see Figure 1) which represents the maximum distance between the listener and that source for the considered path. The time and frequency envelope of the SRIR is then modified dynamically to emulate the relative source-listener distance along the

walking path. The real-time modifications include time delay, level and spectral changes, applied to different segments of the impulse response. Thanks to the SRIR encoding into the HOA domain, the rotation of the listener may be easily compensated for in real-time before being decoded in binaural mode. The expected advantage of this method is that it exploits the SRIR measured in the room thus reflecting the actual characteristics of the room acoustics. However, in contrast with the GA method, the space-time distribution of early reflections is not updated according to the position of the listener and source in the room, which limits the auralization accuracy.

2.4.1 SRIR manipulations

The propagation delay was applied in real-time to the direct sound, early reflections, and late reverberation sections, according to the relative distance between the source and the listener. Filtering was applied to the early reflections and late reverberation segments based on Barron's revised theory [20, 21]. The revised theory takes into account the fact that whereas the reverberation level is assumed to be constant in the room, it exhibits a spatial dependency when counted from the arrival time of the direct sound. This property is expressed through the following formula, that links the frequency-dependent reverberated energy $E(f,r_n)$ observed at distance r_n with respect to the energy $E(f,r_{ref})$ measured at distance r_{ref} :

$$\frac{E(f,r_n)}{E(f,r_{ref})} = exp(\frac{-(r_n - r_{ref}) * 0.04}{RT_{60}(f)})$$

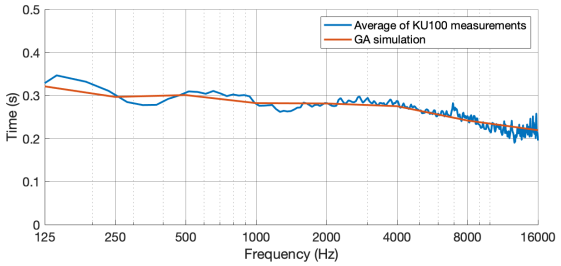

Fig. 3: Measured RT_{60} and its FDN synthesis

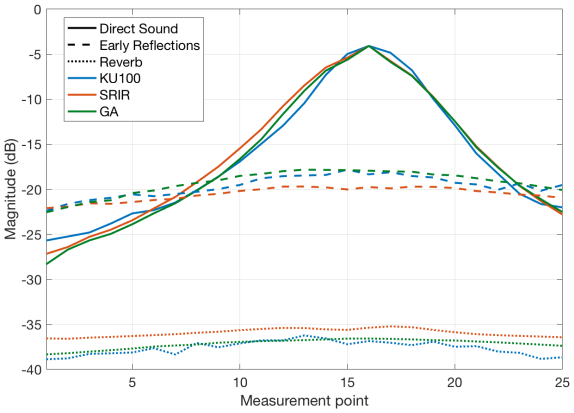

Fig. 4: Energy of time segments along the path

2.5 HOA to binaural decoding and equalization

Both methods are delivering their output under a 4th order HOA format which needs to be decoded into binaural signals. The HOA soundfield was first decoded on a set of 24 virtual loudspeakers evenly distributed on the sphere (slightly sub-optimal compared to the theoretical 25 loudspeakers required for 4th order HOA streams). Then, each loudspeaker channel was filtered with the corresponding HRTF of the KU 100 dummy head available from the HRTF Bili database [22].
Several compensation filters were applied to each segment of the synthesized RIR to compensate for measurement and reproduction chain components. The GA auralization included a filter compensating for the encoding process of FDN into 4th order HOA and then decoding virtual speakers and binaural output. For the SRIR auralization, it included compensation for the diffuse field response of EM32 microphone and a compensation filter for the decoding process of SRIR into virtual speakers and binaural output.
During the subjective listening tests, participants were wearing special open headphones (AKG K1000) with transducers distant from the ear pinnae, in order to guarantee high acoustic transparency of real sound sources. A filter was applied to compensate for the headphone related transfer function (HpTF). As individualized measurements of participants were not possible to be recorded, this HpTF was averaged from measurements conducted on the KU 100 dummy head.

2.6 Calibration and objective comparison

For both methods, the different time sections are simulated separately. Hence, it was possible to calibrate their energy level with regard to a reference point. This was done considering the KU 100 measurement for point M16, which corresponds to the shortest listener to loudspeaker distance (see Figure 1). Thanks to this, it was then possible to check the evolution of the energy levels along the walking path for each time section and to compare them with the measurements (see Figure 4 and section 2.1). For the GA auralization, the evolution is very close to the measurements except for the direct sound on points M1 to M6, which show a slight underestimation (up to -2.6 dB). For the SRIR auralization the late reverberation is slightly overestimated (+1.7 dB), while the early reflections sections is underestimated for short distances (up to -2.2 dB).

3 Subjective listening tests

The main goal of the listening test was to evaluate audio clips played either through real speakers positioned in the room or through virtualized ones on headphones. During each trial, participants walked forward and back following a line drawn on the floor. While they were walking, an audio clip was played once during the way forward from a given real or virtualized loudspeaker and repeated during the way back but from a different real or virtualized loudspeaker. After each trial, participants answered a short questionnaire to rate the two audio clips in terms of their respective plausibility and other audio quality attributes.
The speakers positioned in the room were associated by pairs. For each trial, the audio clips were played consecutively on the two loudspeakers of a given pair (A-C or B-D, see Figure 1). Pair A-C represents speakers which provide similar listening perspective in relation to the room and to the walking path. Thus, the influence of the speaker position on the plausibility and other attributes rating of the two audio clips is expected to be minimal. Pair B-D represents a very different loudspeakers perspective (in terms of distance as well as orientation). Thus the influence of the speaker position on the plausibility and other sound attribute ratings of the two audio clips may be higher.

3.1 Experimental setup

During the experiment, participants were wearing AKG K1000 open headphones with a small tracking device attached (see Figure 1). The tracking system was implemented using the HTC Vive Pro system. A small sensor - Vive Tracker attached to the top of the headphones - allowed to track participants' rotation as well as absolute position. The system employed four infrared cameras mounted in the corners of the room to track the position of the sensor.
In order to help participants adjust their walking speed to the stimuli duration, two iPads were set on each end of the path, which displayed simple visual signs indicating the time to start walking, rotate, or stop.
In order to limit the test duration and the fatigue of participants, only one audio clip was used. The sound stimulus was a 10 second long excerpt from the anechoic recording of a male voice reading short sentences in English. The stimulus was processed in real-time using the above described auralization methods for the playback on headphones, or played back directly from one of the four real loudspeakers standing in the experimental room.

3.2 Conditions

There were 28 combinations of the stimuli pairs (see Table 1). The conditions included 2 pairs of source positions (pair A-C or B-D), 2 orders of playback within a given speaker pair, and 7 combinations of methods: auralization GA or SRIR vs real source (in both orders), two real sources or two auralizations of the same method. From these 28 combinations, 20 were presented twice. The eight conditions with two auralizations (21-28) were not repeated to limit the test duration. All of the trials were randomized for each participant. The average length of the experiment was 75 minutes.

Nr	Methods	Ldspkrs	Nr	Methods	Ldspkrs
1	SRIR R	A-C	15	R GA	B-D
2	SRIR R	C-A	16	R GA	D-B
3	SRIR R	B-D	17	R R	A-C
4	SRIR R	D-B	18	R R	C-A
5	R SRIR	A-C	19	R R	B-D
6	R SRIR	C-A	20	R R	D-B
7	R SRIR	B-D	21	SRIR SRIR	A-C
8	R SRIR	D-B	22	SRIR SRIR	C-A
9	GA R	A-C	23	SRIR SRIR	B-D
10	GA R	C-A	24	SRIR SRIR	D-B
11	GA R	B-D	25	GA GA	A-C
12	GA R	D-B	26	GA GA	C-A
13	R GA	A-C	27	GA GA	B-D
14	R GA	C-A	28	GA GA	D-B

Table 1: Conditions used in the listening test (Methods and Loudspeaker pairs).

3.3 Collection method

Thirty three participants with self-reported normal hearing, with a median age of 29 (min 18, max 47, 23 men, 10 women), participated in the experiment. All participants were expert listeners or students of sound engineering programs. After each trial, participants answered a short questionnaire about the perceived plausibility, localization accuracy, externalization, reverberation, and timbre differences between the two stimuli. The attributes were chosen based on previous studies on the topic [23, 5, 8] and the Spatial Audio Quality Inventory [9]. Participants rated the attributes of different stimuli on a visual scale with the exception of localization, for which they used a simple graph to indicate the perceived position of the sound clips. Participants responded to the questionnaire on a laptop and were guided with a short explanation of the different questions:

- *Plausibility* - For each audio clip, rate the plausibility that it was actually played by one of the loudspeakers (0 - not at all plausible, 6 - very plausible)
- *Localization* - Drag two circles to indicate the localization of audio clip 1 and 2. In case the sound was coming not exactly from the loudspeaker but very close to it - you can put it in the area around the speaker. If the sound was localized even further from the speaker, you can put it anywhere in the picture
- *Blur* - Rate how precise was the localization of the 1st and 2nd audio clips (0 - focused, 6 - blurred)
- *Externalization* - Choose the area that matches the externalization of the 1st and 2nd audio clips (outside of the head, close to the head, inside the head)
- *Timbre* - Rate the difference of timbre between the two audio clips (0 - not different, 6 - very different)
- *Reverberation* - Rate the difference of room reverberation between the two audio clips (0 - not different, 6 - very different)

4 Results

The plausibility, blur, and timbre difference evaluation collected from participants was analyzed to investigate the influence of several factors: acoustic rendering method, localization of virtual and real loudspeakers, as well as the inter-subject and intra-subject variability. Given the repeated measures design of the experiment (two repetitions of the majority of the trials), simply averaging values was avoided by using linear mixed modeling (LMM). This also allowed to include random intercepts to account for individual differences in internal scales and other subjective scores and to add control parameters that might account for variance in the data. Generalized linear mixed modeling was performed in *R* (version 4.0.2) and *RStudio* (version 1.3.959) using the *lmer4* package. Post-hoc comparisons were computed using the package *emmeans*. For each of the preliminary analyses of plausibility, blur, timbre, and plausibility difference ratings two models were generated. The first included speaker position and method of playback as fixed factors (anticipated as the most important factors affecting ratings). The second model comprised an interaction between the two factors. The models also included a random intercept for participant. The best model for each analysis was chosen based on the Akaike information criterion (AIC) to determine the best candidate to explain the variance.

Analysis of the results showed that the best model for predicting plausibility ratings (scale 0 - not at all plausible, 6 - very plausible) was: loudspeaker pair * playback method. There was a significant interaction between speaker and method ($x^2(2) = 114.78$, $P<0.001$), where participants rated both auralization methods as slightly less plausible in comparison to the real speaker for speakers A, C, and D (see Figure 5). In particular, for speaker A there was only a 0.33 point difference for SRIR and R method and 0.36 for GA and R method, and for speaker C, there was a 0.56 point difference for SRIR and R method and 0.77 for GA and R method. The lowest ratings of plausibility were obtained for speaker D: the SRIR synthesis was rated as 3.32, and GA synthesis even lower - 2.79. Importantly, for speaker B there were no differences in plausibility between auralization methods and real loudspeaker.

Similarly to plausibility evaluation, results indicated that the best model for predicting blur ratings (scale: 0 - very focused, 6 - very blurry) was: loudspeaker pair * playback method. There was a significant interaction between speaker and method ($x^2(2) = 79.99$, $P<0.001$) where ratings were higher for the simulated methods than for the real sound for speakers A, B, and C (see Figure 5). Speaker A was rated as slightly more blurry than the real speaker for both rendering methods, but the difference was minimal: 0.29 between SRIR synthesis and the real speaker and 0.33 between GA synthesis

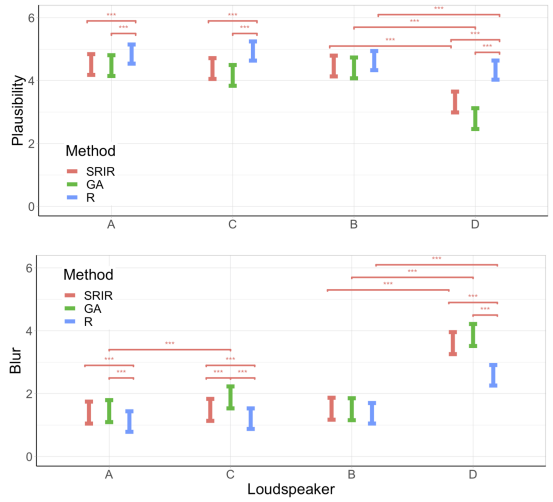

Fig. 5: Predicted values and 95% confidence intervals for plausibility (scale: 0 - not at all plausible, 6 - very plausible) and blur ratings (scale: 0 - very focused, 6 - very blurry) (***P < 0.05)

and the real speaker. The difference was a bit more pronounced between GA and R for speaker position C: 0.62 points. Speaker D had the highest ratings of blur for all 3 methods of rendering. However, both auralizations were rated as significantly more blurry than the real speaker: the difference between SRIR and R was 1.03, and between GA and R 1.29 points. Importantly, there was no significant difference between blur ratings for auralizations and the real speaker for speaker B.

During the experiment, participants also rated the timbre difference within each pair of speakers A-C or B-D (scale: 0 - not different, 6 - very different). Analysis of the results showed that the best model for predicting timbre difference ratings was: loudspeaker pair + playback method. There was a significant main effect of loudspeaker pair (($x^2(1) = 159.0$, $P<0.001$). The timbre difference for pair A-C (1.64) was rated statistically significantly lower than for pair B-D (2.58). During the test, participants also listened to pairs of stimuli with different combinations of playback methods. The possible pairs of methods were: R-R, SRIR-R, GA-R, SRIR-SRIR, GA-GA (see Figure 6). There was a significant influence of playback method pair on the ratings of timbre difference ($x^2(4) = 60.3$, $P<0.001$). The lowest rating was obtained by a pair of two real speakers (R-R, 2.01) and two SRIR auralizations (SRIR-SRIR, 2.11). The highest rating of timbre difference was observed on pair GA-R (2.79). Pairs SRIR-R and GA-GA were rated similarly with the mean estimate slightly higher than for pair R-R (2.58, 2.46 respectively).
In order to compare the results of timbre difference evaluation with plausibility ratings, for each trial, plausibility difference was calculated. The plausibility difference was obtained by taking the absolute value of difference between the plausibility ratings for each stimuli pair. After that, a separate analysis was performed to allow for direct comparison of plausibility difference with timbre difference results (see Figure 6). The best model for predicting plausibility difference was: loudspeaker pair + playback method. There was a significant main effect of loudspeaker pair ($x^2(1) = 90.1$, $P<0.001$). The results for plausibility difference were similar to the timbre difference evaluation. The ratings for pair A-C and B-D were significantly different, with pair A-C obtaining a lower difference (1.09) than pair B-D (1.78). There was also a significant influence of playback method pair on the plausibility difference ($x^2(4) = 30.2$, $P<0.001$). The influence of method was also similar to timbre difference ratings. The lowest difference was obtained by a pair of two real speakers (R-R, 1.40). The largest difference of plausibility rating was observed on pair GA-R (1.99). Pairs SRIR-SRIR, SRIR-R, and GA-GA obtained similar values with the mean estimate slightly higher than for pair R-R (1.75, 1.78, 1.85 respectively).

5 Discussion

We presented a novel experimental design for plausibility evaluation with self-translation of the user. The design of the study focused on resembling a real-life scenario where real sounds are present but do not allow for a direct comparison, as they may be originating from different types of sources (e.g. different voices) and might not be heard from the same perspective. In the experimental design, stimuli were played in pairs by real and virtual loudspeakers with the same acoustical characteristics, but with varying positions.
Importantly, plausibility evaluation seems to be effective using a continuous scale. The presence of results where participants rated the real speaker lower than the maximum of the scale (e.g. loudspeaker D), shows that the plausibility judgment is not binary, as also the real sources might not always be perceived as fully plausible. This poses an interesting area of investigation to answer what factors cause real sounds to be perceived

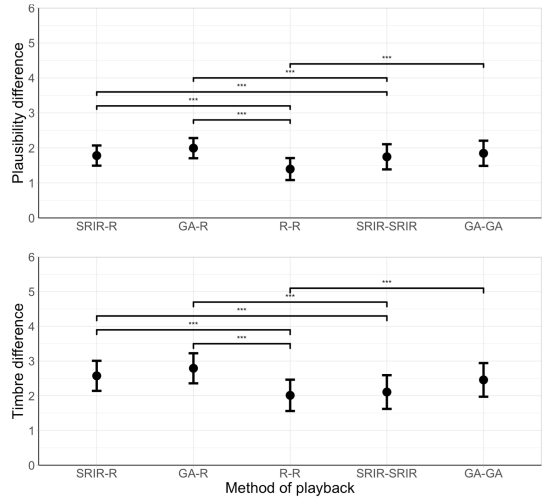

Fig. 6: Predicted means and 95% confidence intervals for plausibility and timbre difference ratings (***P < 0.05)

as less plausible.

The ratings were strongly affected by the position of the loudspeaker in the room. The largest difference between rendering methods was seen for speaker D which might be explained by the inaccuracy of the simulation at this loudspeaker position. As the levels for RIR segments were calibrated to the point in the center of the room, the points furthest away from the center might have the biggest inaccuracies in reproduced levels of RIR. Also, the further away the source from the listener, the bigger role is played by reverberation, which might have been reproduced less accurately. Perception of close sources depended mostly on direct sound.

The design of the study - which included self-translation of the listener - allowed to perceive the three loudspeakers A, B and C under a very similar hearing perspective although they were located at different positions. Only loudspeaker D located further away from the walking path imposed a different hearing perspective. Thus, the two pairs of loudspeakers represented two types of comparisons between real references and auralizations. Pair A-C provided a situation where the hearing perspectives are very similar considering the forward and backward phases of walking, while pair B-D exemplifies a situation where the comparison with the real source is more difficult. This may explain why the plausibility judgment for the virtualized speaker B was not significantly different from the real speaker. In contrast, despite high plausibility scores, the simulation of speakers A and C could be still discriminated from their respective real references. Analysis of timbre and plausibility difference evaluation for different methods suggests that the ratings were correlated. The relative difference between different methods of playback is very similar for both attributes. This may indicate that the plausibility is affected by timbre of the source but further studies are needed to explore this area.

The preliminary results showed that the experimental design was successful in obtaining meaningful comparison of auralizations with a real reference. Both rendering methods obtained similar high plausibility scores, with method SRIR performing slightly better than GA. The limitation of this study is that both rendering methods differ in several aspects, so it is hard to assess which aspect of rendering was crucial for the plausibility evaluation. However, further analysis of the reverberation, timbre and externalization ratings as well as a more detailed objective evaluation of auralizations might reveal some dependencies between rendering and perceptual judgment and specific features of auralization methods.

Future work should include further investigation of the correlation between plausibility and other sound attributes. Both auralizations should be further investigated in perceptual studies to assess how simplification of rendering parameters might affect the plausibility evaluation for expert and naive listeners.

6 Acknowledgments

This material is based upon research supported by the Chateaubriand Fellowship of the Office for Science Technology of the Embassy of France in the United States. The research protocol was approved by the Research Ethical Committee (CER) of Sorbonne University (CER-2021-083) and by the University Committee on Activities Involving Human Subjects (UCAIHS) of New York University (IRB-FY2021-5112).

References

[1] Brinkmann, F., Lindau, A., and Weinzierl, S., "On the authenticity of individual dynamic binaural synthesis," *The Journal of the Acoustical Society of America*, 142(4), 2017.

[2] Bailey, W. and Fazenda, B. M., "The effect of visual cues and binaural rendering method on plausibility in virtual environments," *Proceedings of the 144th AES Convention*, 2018.

[3] Lindau, A. and Weinzierl, S., "Assesing the Plausibility of Virtual Acoustic Environments," *Acta Acustica united with Acustica*, 98(5), 2012.

[4] Werner, S., Klein, F., Neidhardt, A., Sloma, U., Schneiderwind, C., and Brandenburg, K., "Creation of auditory augmented reality using a position-dynamic binaural synthesis system—technical components, psychoacoustic needs, and perceptual evaluation," *Applied Sciences*, 11(3), 2021.

[5] Neidhardt, A., Tommy, A. I., and Pereppadan, A. D., "Plausibility of an interactive approaching motion towards a virtual sound source based on simplified BRIR sets," in *Proceedings of the 144th AES Convention*, Milan, Italy, 2018.

[6] Kuhn-Rahloff, C., *Prozesse der Plausibilitätsbeurteilung am Beispiel ausgewählter elektroakustischer Wiedergabesituationen*, Phd thesis, Technischen Universität Berlin, 2011.

[7] Neidhardt, A. and Zerlik, A. M., "The Availability of a Hidden Real Reference Affects the Plausibility of Position-Dynamic Auditory AR," *Frontiers in Virtual Reality*, 2(September), pp. 1–17, 2021.

[8] Wirler, S. A., Meyer-Kahlen, N., and Schlecht, S. J., "Towards Transfer-Plausibility for Evaluating Mixed Reality Audio in Complex Scenes," in *AES Conference on Audio for Virtual and Augmented Reality*, Virtual, 2020.

[9] Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkmann, F., and Weinzierl, S., "A Spatial Audio Quality Inventory (SAQI)," *Acta Acustica united with Acustica*, 100(5), 2014.

[10] Shinn-Cunningham, B. and Ram, S., "Identifying where you are in a room: Sensitivity to room acoustics," in *Proc. of the 9th Int. Conference on Auditory Display*, Boston, MA, USA, 2003.

[11] Massé, P., Carpentier, T., Warusfel, O., and Noisternig, M., "Denoising directional room impulse responses with spatially anisotropic late reverberation tails," *Applied Sciences*, 10(3), 2020.

[12] Carpentier, T. and Einbond, A., "Spherical correlation as a similarity measure for 3D radiation patterns of musical instruments," in *16ème Congrès Français d'Acoustique*, Marseille (FR), 2022.

[13] Kronlachner, M. and Zotter, F., "Spatial transformations for the enhancement of Ambisonic recordings," *2nd International Conference on Spatial Audio*, (2), 2014.

[14] Fediuk, R., Amran, M., Vatin, N., Vasilev, Y., Lesovik, V., and Ozbakkaloglu, T., "Acoustic Properties of Innovative Concretes: A Review," *Materials*, 14(2), 2021.

[15] Carpentier, T., "Spat: a comprehensive toolbox for sound spatialization in Max," *Ideas Sonicas*, 13(24), 2021.

[16] Poirier-Quinot, D., Katz, B., and Noisternig, M., "EVERTims: Open source framework for real-time auralization in VR," *ACM International Conference Proceeding Series*, Part F1319, 2017.

[17] Vorländer, M., *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, Springer, 2008.

[18] Greenblatt, A., Abel, J., and Berners, D., "A Hybrid Reverberation Crossfading Technique," in *ICASSP*, 2010.

[19] Nowak, J. and Klockgether, S., "Perception and prediction of apparent source width and listener envelopment in binaural spherical microphone array auralizations," *The Journal of the Acoustical Society of America*, 142(3), 2017.

[20] Barron, M. and Lee, L. J., "Energy relations in concert auditoriums. I," *Journal of the Acoustical Society of America*, 84(2), 1988.

[21] Jot, J.-M., Audfray, R., Hertensteiner, M., and Schmidt, B., "Rendering Spatial Sound for Interoperable Experiences in the Audio Metaverse," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–15, 2021.

[22] Carpentier, T., Bahu, H., Noisternig, M., and Warusfel, O., "Measurement of a head-related transfer function database with high spatial resolution," in *7th Forum Acusticum (EAA)*, 2014.

[23] Olko, M., Dembeck, D., Wu, Y.-H., Genovese, A. F., and Roginska, A., "Identification of perceived sound quality attributes of 360 audiovisual recordings in VR using a Free Verbalization Method," in *Proceedings of the 143rd AES Convention*, New York, USA, 2017.